Updating the Standardized World Income Inequality Database

Objective: *Methods*: *Results*: *Conclusion*:

The SWIID Source Data

To be included in the SWIID's source data, observations need to encompass the entire population of a country without regard to age, location,[1] or employment status.[2] They need to have an identifiable welfare definition and equivalence scale (more on these below). Finally, to ensure that these original sources are easily available to SWIID users, observations need to be available online, although not necessarily without paywalls.[3]

Hand-entering data is tedious and error-prone work, so I automated as much of the process of data collection as practicable. Most international organizations and a few national statistical offices use application programming interfaces (APIs) that facilitate downloading their data, and often the R community has built packages using these APIs to make the task even easier (see Magnusson, Lahti and Hansson 2014; Lahti et al. 2017; Lugo 2017; Blondel 2018; Wickham, Hester and Ooms 2018). I took as much advantage of these resources as possible. Although the sources with APIs were relatively few, they contained the most data: 55% of the observations were collected this way. In the absence of an API, I scripted downloads of any available spreadsheets (see Wickham 2016). If there was no spreadsheet, but data were available in pdf files, I automated downloading these files and then used the `tabulizer` package (Leeper 2016) to read the tables into R. In the absence of a file to download, I scripted the process of scraping the data from the web.[4]

Still, for a variety of reasons, a source's data may have been consigned to being entered in a separate spreadsheet.[5] Many sources contain just a handful or fewer observations, making the payoff to the often laborious process of data cleaning too small to justify the effort. Some sources–including most academic articles–are behind paywalls, making reproducibility particularly challenging in any event. Some sources, like books, cannot be read straight into R. All of the entries in this spreadsheet were checked repeatedly

---

[1]The requirement for complete territorial coverage was relaxed for minor deviations such as data on Portugal that excludes Madeira and the Azores. It was relaxed somewhat further for early series that covered only the urban population of three highly urbanized countries: Uruguay, Argentina, and South Korea. The general rule, however, is that data is excluded if it measures the income distribution of only urban or rural populations, or of only selected cities, or some other such incomplete territory.

[2]This last requirement is new; it means nearly 600 observations on the distribution of wages across employed individuals that were included in the source data of previous versions of the SWIID are now excluded. Between the lack of information on those out of the workforce and on how workers formed households, these data weren't very strongly related to LIS data on income inequality in the entire population anyway.

[3]For scholarly articles, DOIs or JSTOR stable URLs were the preferred web addresses, but if those were unavailable the publisher's website or another repository was used. For books, the link is to the relevant page in Google Books.

[4]Code for the entire process can be viewed here: https://github.com/fsolt/swiid/blob/master/R/data_setup.R.

[5]See https://github.com/fsolt/swiid/blob/master/data-raw/article_data/fs_added_data.csv.

for errors,[6] and I excluded repeated reports of the exact same observation from different sources. In the end, I was able to automate the collection of approximately three quarters of the source data and a much higher percentage of the series that will be updated or are subject to revision, facilitating incorporating these changes in future versions.

The resulting dataset comprises 14487 Gini coefficients from 2833 country-years in 192 countries or territories; as shown in Figure 1, this makes the coverage of the SWIID source data broader than that of any other income inequality dataset. This is not surprising given that, with the exceptions of the two secondary-source collections–the World Income Inequality Database (UNU-WIDER 2017), which is not drawn on at all anymore, and the *All the Ginis* database (Milanovic 2014), which is not drawn on much–the SWIID source data incorporates all of the data in these other datasets.
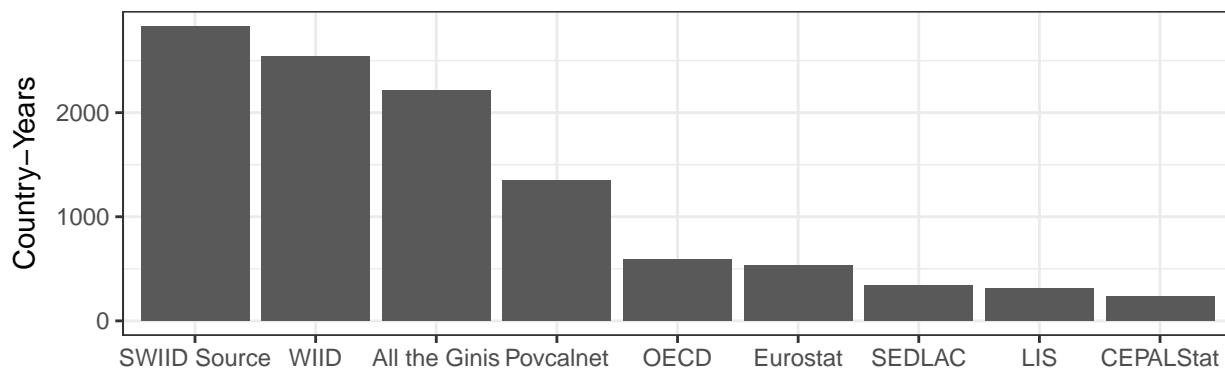

Figure 1: Income Inequality Datasets by Country-Years Covered

Turning from how the source data were collected to how they are composed, we find, not unexpectedly, that there is much more data available about the income distribution in some countries than in others. Which countries are most data-rich? The plot below shows the top dozen countries by the count of observations. Canada, by virtue of the excellent Statistics Canada as well as longstanding membership in the OECD and LIS, has 733 observations, many more than any other country. The United Kingdom and United States are next, followed by an interesting mix of countries from around the world with not surprisingly a sizable European representation. All are members of the LIS. On the other hand, ten countries have only a single observation.

As we will see below, observations for the same country in the same year, but with different welfare definitions and equivalence scales or from different sources, are important to generating the SWIID's cross-nationally comparable estimates. Still, we might be interested to know which countries have the most coverage of the years in the SWIID's current 57-year timeframe, from 1960 to 2017, because the SWIID's

---

[6]Which, of course, is not to say that they are error-free. If you spot any problems or, better still, know of sources I might have missed, *please* let me know at https://github.com/fsolt/swiid/issues/6.
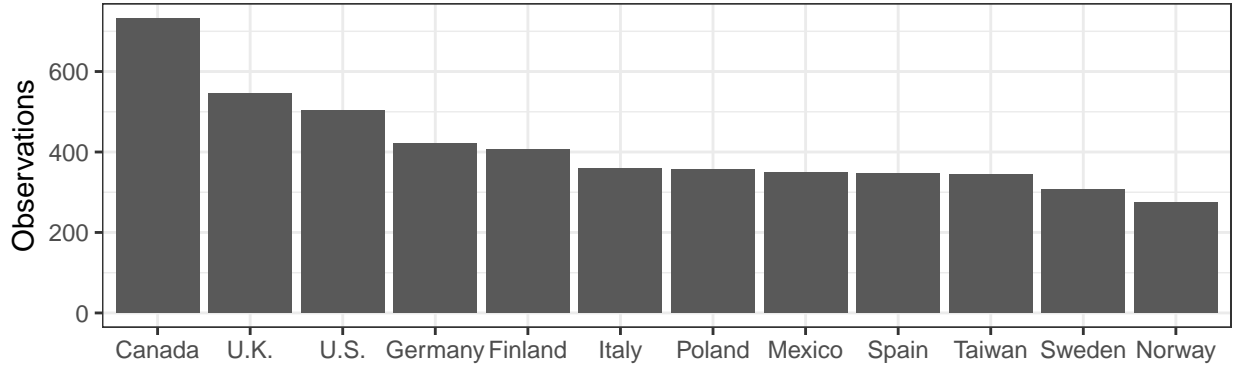
Figure 2: Countries with the Most Observations in the SWIID Source Data

inequality estimates for countries with fewer country-year observations will include more interpolated values, which in turn will have more uncertainty.
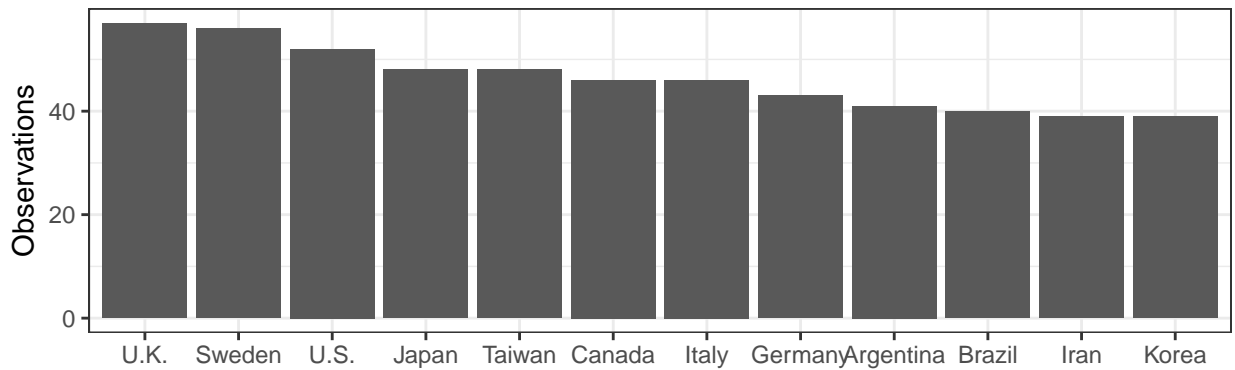

Figure 3: Countries with the Most Years Covered in the SWIID Source Data

The source data includes observations in every covered year for the United Kingdom, in all but one of these years for Sweden, and in all but five for the United States. Argentina—a country not included in the LIS—makes the top ten, with 41 country-year observations. The median country has observations in just nine different country-years.

We can also get a sense of the available inequality data by turning the question around and asking about coverage across countries over time. There are observations for 122 countries in 2005. Coverage is relatively good in the years from 2000 to 2015, at least 80 countries per year, before dropping to 46 countries for 2016 and just 6 for last year. Country coverage is pretty thin each year through the 1960s and 1970s and still is not great until the late 1980s.[7]

Earlier I mentioned that to be included in the SWIID source data observations need to have an identifiable welfare definition and equivalence scale. A welfare definition is an answer to the question, this Gini measures

---

[7]This is partly a result of my decision to insist on sources that are available online, but it's just as well: so little information is available about many of the so-excluded observations on that era that I find it hard to have much confidence in them.
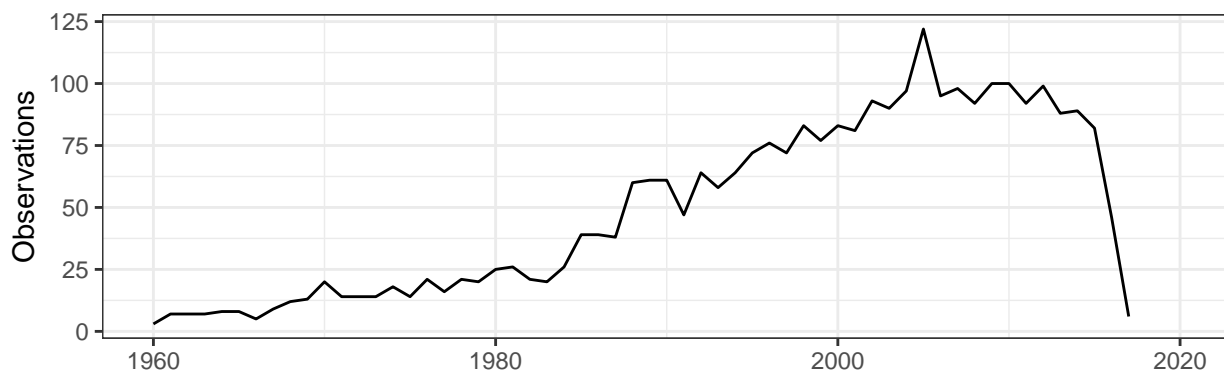
Figure 4: Countries Covered by Year in the SWIID Source Data Data

the distribution of what? The four welfare definitions employed in the SWIID source data are market income, gross income, disposable income, and consumption. Market income is defined as amount of money coming into the household, excluding any government cash or near-cash benefits, the so-called 'pre-tax, pre-transfer income.'[8] Gross income is the sum of market income and government transfer payments; it is 'pre-tax, post-transfer income.' Disposable income, in turn, is gross income minus direct taxes: 'post-tax, post-transfer income.'[9] Consumption does not refer to the money coming into the household at all but rather to the money going out.[10] In the source data, Ginis of disposable income are much more common than those using other welfare definitions.

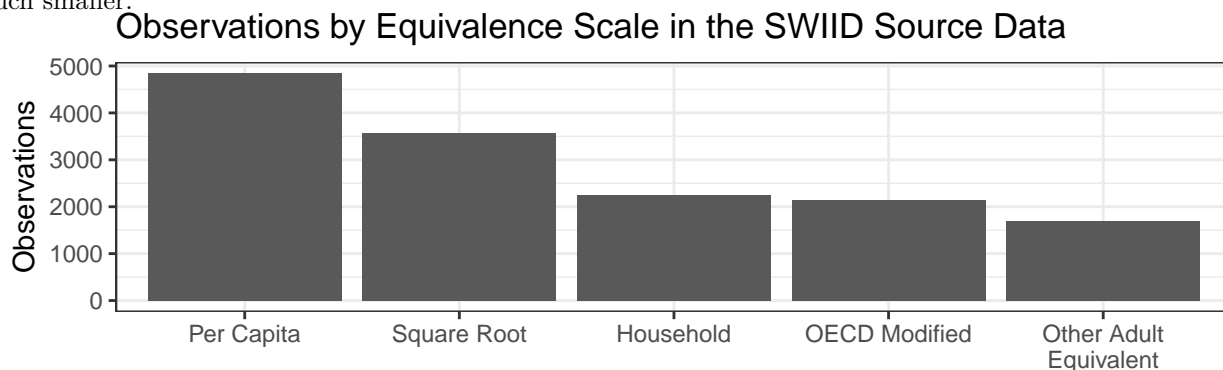Observations by Welfare Definition in the SWIID Source Data

---
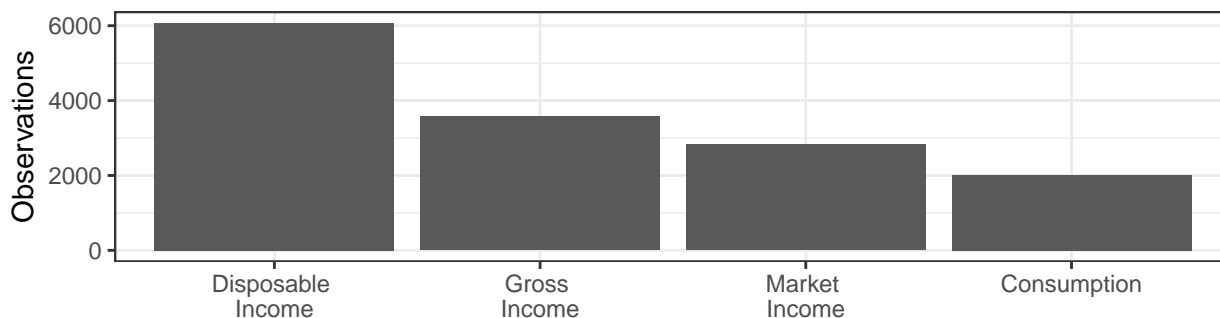
[8]It's important, though, to not think of the distribution of market income as 'pre-government.' Beyond taxes and transfers, governments seeking to shape the distribution of income have a wide array of 'market-conditioning' or 'predistribution' policy options, with minimum wage regulation and labor policy two obvious examples (see, e.g., Morgan and Kelly 2013). Further, even taxes and transfers can profoundly shape the distribution of *market* income through 'second-order effects.' Where robust public pension programs exist, for example, people save less for retirement, leaving many of the elderly without market income in old age and so raising the level of market-income inequality (see, e.g., Jesuit and Mahler 2010).

[9]Note that disposable income still does not take into account, on the one hand, indirect taxes like sales taxes and VAT, or, on the other, public services and indirect government transfers such as price subsidies. There is very little information available about the distribution of such 'final income,' pretty much only that generated by the Commitment to Equity Institute, so I exclude it from the SWIID source data at least for the time being.

[10]In previous versions of the SWIID, market and gross income were treated as a single welfare definition, and I am glad to finally be able to split them apart (c.f., Solt 2016, 1272). The consumption welfare definition *might* now be the most heterogeneous within the SWIID source data, varying considerably in whether and how observations treat expenditures on durable goods. Another source of differences within a single welfare definition is the extent to which nonmonetary income— such as the value of food grown for the household's own consumption or of housing that the owner occupies—is included. The SWIID source data include the variable `monetary` that indicates whether any nonmonetary income is taken into account, but at present this information is not incorporated into the classification of welfare definitions.

Equivalence scales are the ways in which the size and composition of a household is incorporated into the calculation of its members' welfare. On the one hand, these factors can simply be ignored, with all households with the same amount of income or consumption treated as if they enjoy the same level of welfare, regardless of their size. One can improve on this household 'scale'[11] by dividing the household's income by its number of members, that is, by using a per capita scale. But a household of two members and an income of $100,000 is better off than one with a single member and $50,000 due to economies of scale—that's the big reason why people look for roommates. There are a variety of ways to try to account for these economies by calculating the number of "equivalent adults" in the household. Of the most commonly used adult-equivalent scales, the square-root scale is the most straightforward: one simply divides the household income by the square root of the number of members. The "OECD-modified" scale for the number of adult equivalents (which the OECD actually no longer uses) counts the first adult as 1, all other adults as .5, and each child as .3. And there are plenty of other adult-equivalent scales, from the "old OECD" scale (1 for the first adult, 0.7 for each additional adult, and 0.5 for each child) to caloric-requirement-based scales (which are actually very nearly per capita, as it turns out) to a number of country-specific scales. In previous versions of the SWIID, all adult-equivalent scales were considered a single category. Now, the square-root scale and the OECD-modified scale have both been split out, leaving the remaining catch-all adult-equivalent category much smaller.

Observations by Equivalence Scale in the SWIID Source Data

Differences in the welfare definition and the equivalence scale employed constitute the biggest source of incomparability across observations in the source data, and all twenty of the possible combinations are represented. I'll take up how we get from these incomparable observations to the SWIID estimates in the next post. In the meantime, if you'd like to see the source data, you can download it from here.

---

[11]Scare quotes because, strictly speaking, nothing is being scaled at all; it's simply treating the household as the unit of analysis.

Observations by Welfare Definition and Equivalence Scale in the SWIID S
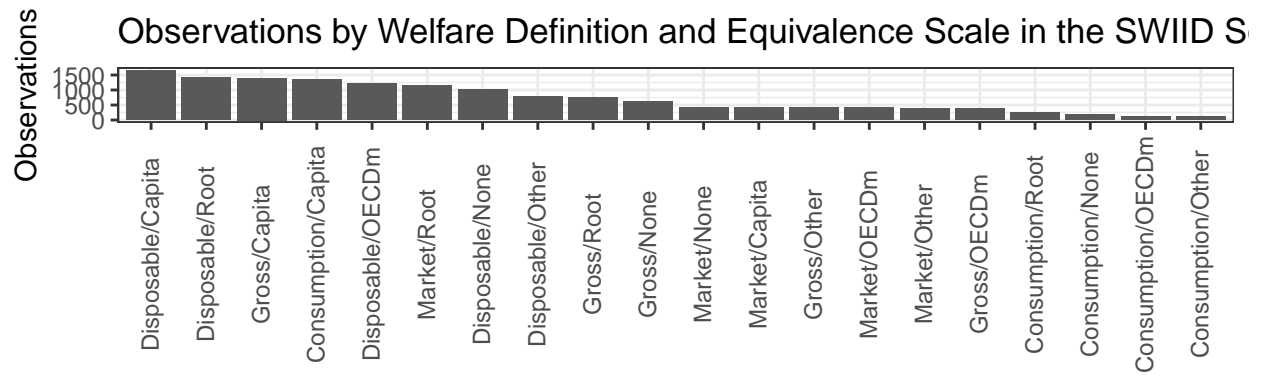
Standardizing the Data

References

Blondel, Emmanuel. 2018. "rsdmx: Tools for Reading SDMX Data and Metadata." R package version 0.5-11. `https://CRAN.R-project.org/package=rsdmx`.

Jesuit, David K. and Vincent A. Mahler. 2010. "Comparing Government Redistribution Across Countries: The Problem of Second-Order Effects." *Social Science Quarterly* 91(5):1390–1404.

Lahti, Leo, Janne Huovari, Markus Kainu and Przemysław Biecek. 2017. "Retrieval and Analysis of Eurostat Open Data with the eurostat Package." *The R Journal* 9(1):385–392.
URL: *https://journal.r-project.org/archive/2017/RJ-2017-019/index.html*

Leeper, Thomas J. 2016. *tabulizer: Bindings for Tabula PDF Table Extractor Library.* R package version 0.1.24.

Lugo, Marco. 2017. "CANSIM2R: Directly Extracts Complete CANSIM Data Tables." R package version 0.12. `https://CRAN.R-project.org/package=CANSIM2R`.

Magnusson, Mans, Leo Lahti and Love Hansson. 2014. "pxweb: R tools for PX-WEB API." `http://github.com/ropengov/pxweb`.

Milanovic, Branko. 2014. "Description of *All The Ginis* Dataset." World Bank, Research Department.

Morgan, Jana and Nathan J. Kelly. 2013. "Market Inequality and Redistribution in Latin America and the Caribbean." *Journal of Politics* 75(3):672–685.

Solt, Frederick. 2016. "The Standardized World Income Inequality Database." *Social Science Quarterly* 97(5):1267–1281.

UNU-WIDER. 2017. "World Income Inequality Database, Version 3.4, January 2017." `https://www.wider.unu.edu/database/world-income-inequality-database-wiid34`.

Wickham, Hadley. 2016. "rvest: Easily Harvest (Scrape) Web Pages." R package version 0.3.2. `https://CRAN.R-project.org/package=rvest`.

Wickham, Hadley, James Hester and Jeroen Ooms. 2018. "xml2: Parse XML." R package version 1.2.0. `https://CRAN.R-project.org/package=xml2`.