Measuring Income Inequality Across Countries and Over Time: The Standardized World Income Inequality Database [DRAFT]*

Frederick Solt *University of Iowa*

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

Keywords: income inequality, measurement

Introduction

Collecting the SWIID Source Data

The SWIID source data consists of observations of the Gini coefficient in various countries and years.[1] The Gini is most intuitively defined as the average difference in income between all pairs in a population, divided by twice the average income in the population. It is by far the most commonly encountered summary statistic for measuring income inequality. The Gini has drawbacks—it is most sensitive to changes in the middle of the income distribution, rather than among those with the highest or lowest incomes, and it is not easily decomposed—that other metrics, such as the Atkinson index and the Theil indices, overcome. However, in light of the SWIID's goal of providing data for the broadest possible sample of countries and years, the ubiquity of the Gini makes it the only plausible choice for this purpose.

To be included in the SWIID source data, Gini observations need to encompass the entire population of a country without regard to age, location, 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, as shown in Figure~1. Although

*The paper's revision history and the materials needed to reproduce its analyses can be found on Github here. Corresponding author: frederick-solt@uiowa.edu. Current version: 08 March 2019.

[1]Because the Gini index is simply the Gini coefficient multiplied by 100, the two are equivalent, and both can be referred to as 'the Gini' without much cause for confusion.

[2]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. This requirement that the observation must not be restricted to only the employed is new; it means nearly 600 observations on the distribution of wages across employed individuals that were included in the source data of earlier 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 were not 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.

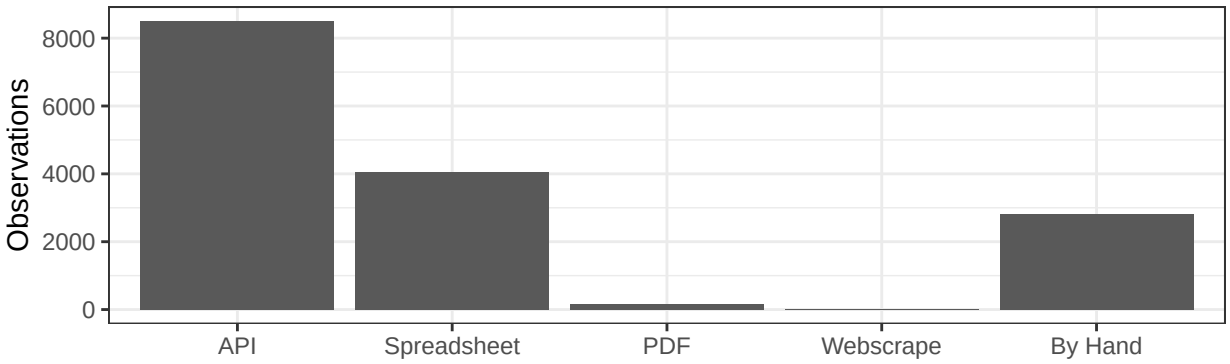


Figure 1: Income Inequality Observations by Method of Collection

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 downloading and reading 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 rare absence of any 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. Other sources, such as many books, cannot be read into R. Finally, one source contains crucial information encoded in the typeface of its tables, information lost when the tables are read directly into R (see Mitra and Yemtsiv 2006, 6). All of the entries in this spreadsheet were checked repeatedly for errors, and I excluded repeated reports of the exact same observation from different sources.[6]

In the end, I was able to automate the collection of 82% of the source data and an even higher percentage of the observations that will be updated or are subject to revision, greatly facilitating incorporating these changes in future versions.

The resulting dataset comprises 15,549 Gini coefficients from 2,951 country-years in 196 countries or territories; as shown in Figure 2, 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 other secondary collections—the World Income Inequality Database (UNU-WIDER 2018), which contains no original data and so is not drawn on at all, and the *All the Ginis* database (Milanovic 2019), which contains very little original data and so is not drawn on much—the SWIID source data incorporates all of the data in these other datasets.

Turning from how the source data were collected to how they are composed reveals that there is much more data available about the income distribution in some countries than in others. Which countries are most data-rich? Figure 3 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 736 observations, many more than any other country. The United Kingdom, Germany, and the United States are next, followed by a group dominated by European

[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/fs_added_data.csv.

[6]Which, of course, is not to say that these entries are error-free. If you spot any problems or know of sources I might have missed, *please* let me know at https://github.com/fsolt/swiid/issues/6.

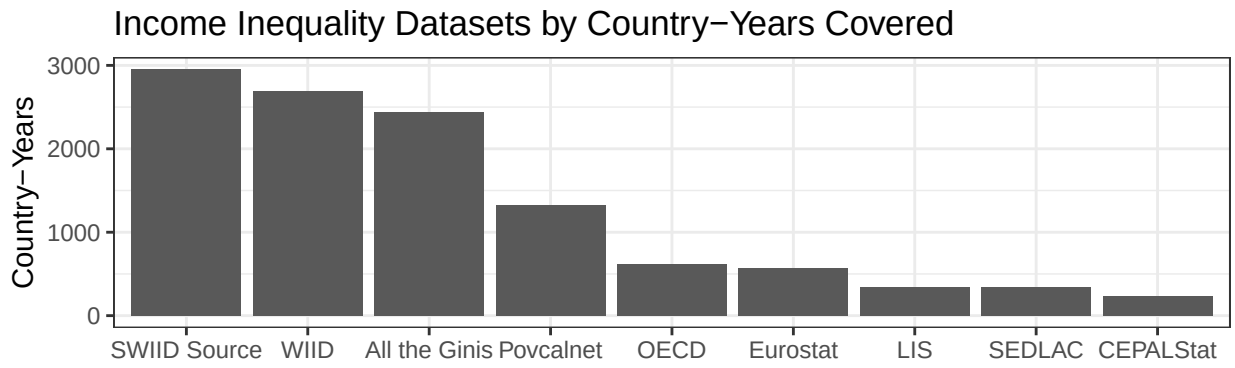


Figure 2: Income Inequality Datasets by Country-Years Covered

countries but including Mexico, Taiwan, and Brazil. All of these data-rich countries are members of the LIS. On the other hand, the eleven most data-poor countries have only a single observation each in the SWIID source data.

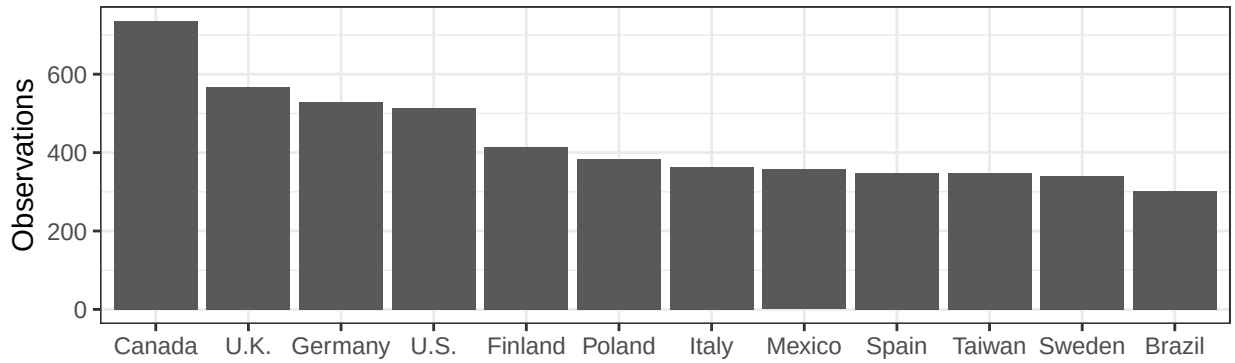


Figure 3: Countries with the Most Observations in the SWIID Source Data

As discussed in the next section, 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 inequality estimates for countries with fewer country-year observations will include more interpolated values, which in turn will have more uncertainty, and—because estimates are not extrapolated beyond the years observed in the source data—more years without any estimates at all. The countries with the most observed years are shown in the left panel of Figure 4. The source data includes observations in every covered year for Sweden and the United Kingdom. There are observations in all but four years for the United States, and in all but eight years for Japan and Taiwan. Iran and Argentina—countries that are not members of the LIS—also make the top ten, with observations in 43 and 42 country-years, respectively. The median country, though, has observations in just nine different country-years.

We can also get a sense of the breadth of the available income inequality data by turning the question around and asking about the number of countries covered across time. The right panel of Figure 4 shows, for each year, the number of countries for which the SWIID source data includes at least one observation. There are observations for 123 countries in 2005, the year with

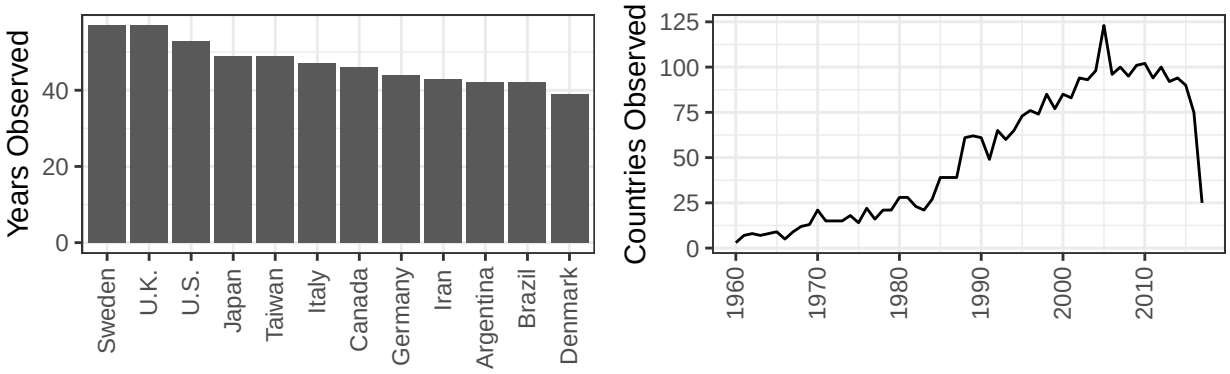


Figure 4: Country-Year Coverage in the SWIID Source Data

the broadest coverage. Coverage is relatively good in all of the years from 2000 to 2015, with at least 80 countries observed per year, before dropping to 75 countries for 2016 and just 25 for 2017. Before then, country coverage is pretty thin each year through the 1960s and 1970s and still is not great until the late 1980s.[7]

Above I mentioned that to be included in the SWIID source data observations need to have an identifiable welfare definition and equivalence scale; let us now consider these two aspects of the data. A welfare definition is an answer to the question, this Gini measures 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 the 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] As can be seen in the left panel of Figure 5, in the SWIID source data, Ginis of disposable income are much more common than those

[7]This is partly a result of the 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 it is hard to have much confidence in them.

[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). Moreover, 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 or value-added taxes, 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 (cf. 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.

using other welfare definitions.

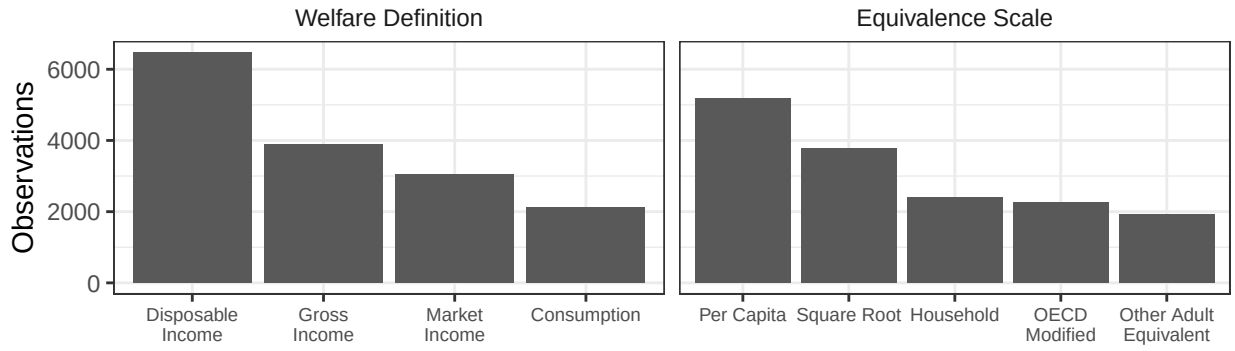


Figure 5: Welfare Definitions and Equivalence Scales in the SWIID Source Data

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. Although undoubtedly superior to simply looking at households, the per capita scale is not ideal: 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 is the most important 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 itself actually does not use, preferring the square-root scale) counts the first adult as 1, all other adults as .5, and each child as .3. And there are many 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 in fact 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. The right panel of Figure 5 shows how many observations in the SWIID source data use each equivalence scale.

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.

Measuring Income Inequality Comparably

Evaluating the Comparability of the SWIID's Estimates With Cross-Validation

How can we know if the approach just described actually works? In previous work, I provided the most stringent test I could come up with: I examined LIS data on country-years that had been

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

included in previously-released versions of the SWIID (Solt 2016, 1277-1278).[12] The results were reassuring in some ways—only seven percent of the differences between new LIS observations and old SWIID estimates were statistically significant and larger than two Gini points, a far better record than that achieved by data carefully selected from the UNU-WIDER database or the *All the Ginis* dataset adjusted in accordance with its instructions–but less so in others. Most disappointingly, only 72% of the differences had 95% confidence intervals that included zero, suggesting that the SWIID's standard errors were often too small. I've been working hard on the SWIID's estimation routine to fix these issues since I conducted that test back in 2014, but the LIS doesn't release new data frequently enough to allow for continuous testing of these revisions. So, instead, I've drawn on a technique developed in data science and machine learning, k-fold cross-validation, to assess the SWIID's progress.

To get how k-fold cross-validation works, it helps to first understand the simpler form of cross-validation in which the available data are first divided into two groups of observations: the *training* set and the *testing* set. The model parameters are then estimated on only the training set. Finally, these results are used to predict the values of the testing set (that is, again, observations that were not used to estimate the model's parameters). By comparing the model's predictions against the test set, we avoid overfitting and get a good sense of how well the model performs in predicting other, as yet unknown, data.

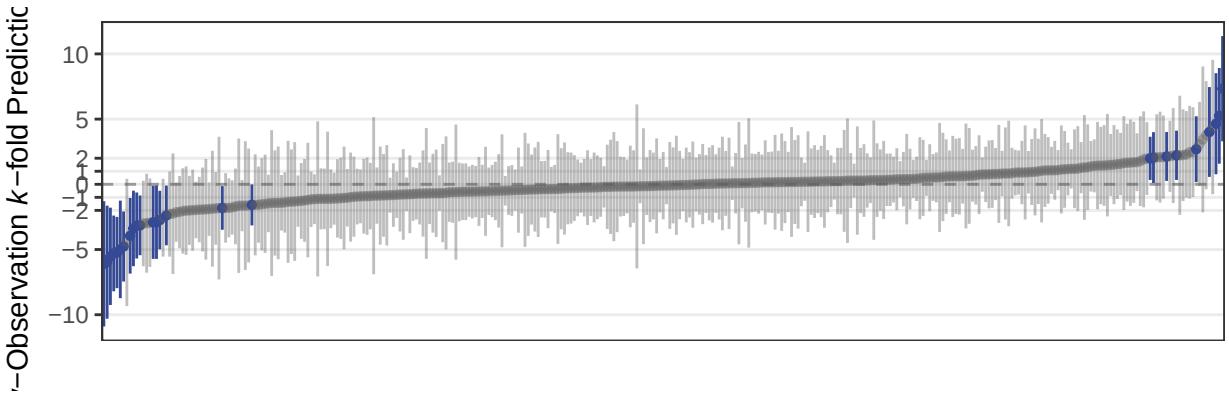
Still, that sense may be biased by the exact observations that happened to be assigned to each set. We can reduce this bias by performing the process repeatedly: this is k-fold cross-validation. The available data are divided into some number k groups. One at a time, each of the k groups is treated as the testing data, with all other groups forming the training data for estimating the model. The model's performance is then evaluated by considering how well it predicts *all* of the groups, and because every observation is included in the testing data at some point, the process allows us to check whether and for which observations the model is doing particularly poorly.

To provide a first assessment of the SWIID's ability to predict the LIS, I randomly assigned the available LIS observations into groups of three, with an added check to ensure that no group included two observations from the same country.[13] (Because the SWIID routine relies only on relationships observed within-country for the countries included in the LIS, the check that only a single observation from a country be assigned to the test data at a time means that the exact size of the group doesn't really matter.)[14] The figure below plots the difference between the SWIID prediction generated from this k-fold cross-validation and the LIS data for each country-year included in the LIS. Observations for which the 95% credible interval for this difference includes zero are gray; those for which it doesn't are highlighted in blue.

[12]For the initial kernel of this idea, I remain grateful to participants in the Expert Group Meeting on Reducing Inequalities in the Context of Sustainable Development, Department of Economic and Social Affairs, United Nations, New York, October 24–25, 2013.

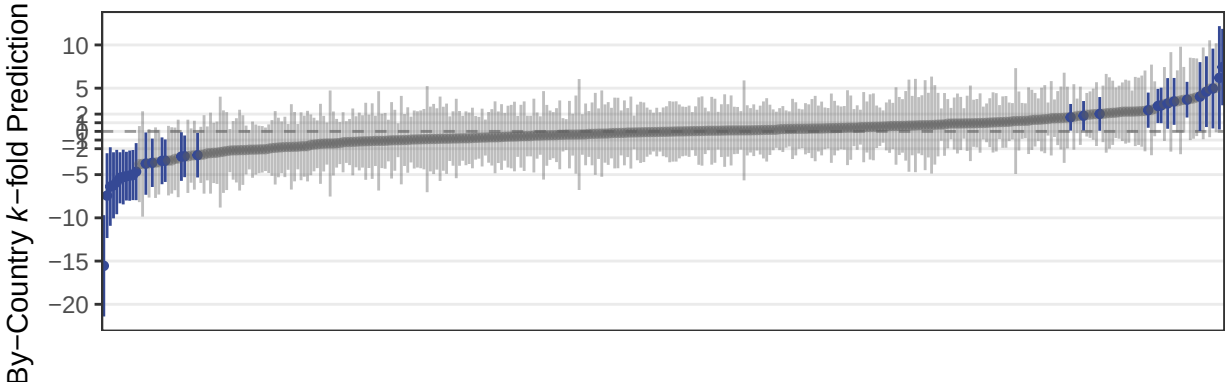
[13]The goal of this exercise is really to assess how well the SWIID works within the LIS countries, so Egypt 2012, the only LIS observation for that country, is excluded from the analysis. This is because holding out that observation makes Egypt a *non*-LIS country.
What happens when the SWIID is used to predict all of a country's LIS observations at once is discussed below.

[14]My experiments with groups sized from just one observation each up to six observations each confirmed this. Three observations per group struck a nice balance between the time it takes to randomly generate the groups (which increases with group size because it becomes more likely for a group to be rejected for containing two observations from a single country) and the demand the work puts on UI's high performance computing cluster (which increases with the number of groups–which is also the number of times the SWIID routine is re-run). I probably don't really need to worry about the latter, but whatever—like I said, it doesn't actually matter.



The results show that the SWIID does a very good job of predicting the LIS: the 95% credible interval for the difference between the two includes zero for 92% of these observations. The point estimates for these differences are generally small, with 86% less than 2 Gini points and 62% less than a single Gini point. It's true that there are a few observations for which the estimated difference is quite large–on the far left of the plot, the SWIID routine underestimated the LIS Gini for Hungary in 1991 by 6 ± 4 points, and on the extreme right the SWIID routine overestimated that for Guatemala in 2014 by 7 ± 4 points–but there doesn't really seem to me to be much pattern in which countries and years are estimated poorly.

This test, though, really only assesses how well the SWIID predicts LIS-comparable inequality figures in years without LIS data in the (now fifty) countries that are *included in the LIS*. We can get a better sense of how well the SWIID does predicting countries not covered by the LIS with another cross-validation that, one country at a time, excludes *all* of the LIS observations for that country. The results of just such a test are plotted below.



Overall, the plot looks very similar to the one above. With each country's entire run of LIS data taking a turn being excluded, the 95% credible interval for the difference between the resulting SWIID estimate and the excluded LIS data contains zero 91% of the time. And here, too, most of the point estimates for these differences are small: 74% are less than 2 Gini points, and 52% are less than one Gini point.

This analysis, though, does point to a few rough spots in need of future attention. The first appears on the far left of the plot above. There we find that the largest difference is for the sole country-year for Egypt in the LIS–for 2012–which the SWIID routine underestimates by 16(!) ± 6 Gini points. Egypt is currently the only country in the LIS with just a single country-year observation; given that excluding the one observation is equivalent to excluding all of the country's observations, I skipped omitting it in the first cross-validation. LIS researchers Checchi et al. report in a footnote that Egyptian income surveys before 2012 did not include any questions to capture self-employment income, and it's also true that most of the available Ginis for Egypt are based

on the distribution of consumption expenditure, which sometimes only loosely track those for the distribution of income (see, e.g., India). These factors, however, are present in many non-LIS countries as well, so I'll continue working to come up with ways to improve the SWIID routine for such cases.

The second is that there are two other countries for which the 95% credible interval for the differences between the LIS data and the SWIID routine's estimates for those countries when all of their LIS data are excluded does not contain zero in *any* of the country's observations: Brazil and Peru. For Brazil, the cross-validation's estimates of the country's four LIS observations are all too high—by 2.5 ± 2.0 Gini points to 3.7 ± 2.1 Gini points. The cross-validation's estimates for Peru's four LIS observations, on the other hand, are all too low—by between 5.0 ± 2.9 and 5.4 ± 3.0 Gini points. So there is some room for improvement here too, and I'll keep working on it also.

All in all, though, these k-fold cross-validation exercises show that the SWIID does a very good job of predicting the LIS, which inspires confidence that the SWIID is indeed maximizing the comparability of income inequality data across countries and over time.

References

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.

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. 2019. "Description of *All The Ginis* Dataset." Graduate Center, City University of New York and Stone Center on Socio-economic Inequality.

Mitra, Pradeep and Ruslan Yemtsiv. 2006. "Increasing Inequality in Transition Economies: Is There More to Come?" World Bank Policy Research Working Paper 4007, September 2006.

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. 2018. "World Income Inequality Database, Version 4, December 2018." https://www.wider.unu.edu/database/world-income-inequality-database-wiid4.

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.