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

From its origins now over a decade ago, the goal of the Standardized World Income Inequality Database has been to provide estimates of income inequality for as many countries and years as possible while ensuring that these estimates are as comparable as the available data allow (see **?**). That is to say, the SWIID's first prority is breadth of coverage, and its second is comparability. The starting point for the SWIID estimates is a dataset with the complementary priorities: the Luxembourg Income Study, which aims to maximize comparability and, given that primary concern, to include as many countries and years as possible.[1] Then, the SWIID routine estimates the relationships between Gini indices based on the LIS and all of the other Ginis available for the same country-years, and it uses these relationships to estimate what the LIS Gini *would be* in country-years not included in the LIS but available from other sources. This approach has made the SWIID a preferred source of income inequality data for researchers pursuing broadly cross-national work.

The SWIID has recently been completely revised. This article explains, first, how the source data for the SWIID is now collected. It turns next to how these data are used to generate comparable estimates of income inequality across countries and over time. To demonstrate that this procedure succeeds in producing estimates are comparable to the LIS data, it then reviews evi-

*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: 12 June 2019.

[1]Still, even for the LIS prioritizing comparability has given way, to some degree, to the desire to cover more countries. Checchi et al. (2018) have recently written about the difficulties the LIS team has encountered including middle-income countries due to the greater importance of non-monetary and self-employment income as well as the differences in direct taxation and social security contributions in these countries in comparison to high-income countries. Despite these issues, the LIS remains the most comparable income inequality data available.

dence from two k-fold cross-validations. Finally, it concludes by explaining how researchers may best make use of the SWIID in their own research.

Collecting the SWIID Source Data

The SWIID source data consists of observations of the Gini coefficient in various countries and years.[2] 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 is more sensitive to changes in the middle of the income distribution than some other metrics, such as the Atkinson index and the Theil indices, and unlike those measures it cannot be analytically decomposed. But in light of the SWIID's goal of providing information on income inequality 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.[3] 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.[4]

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,

[2]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' with little cause for confusion.

[3]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.

[4]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. The source data can be accessed and explored graphically on the web at https://fsolt.org/swiid/swiid_source.html.

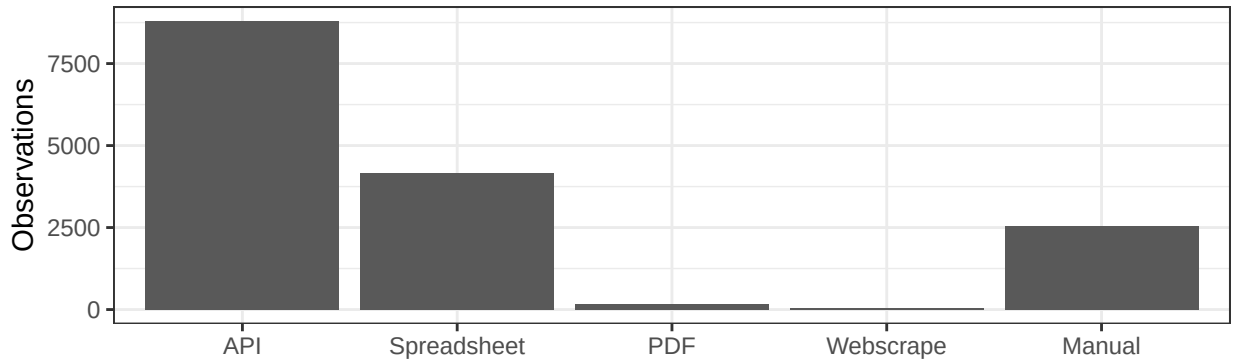

Figure 1: Income Inequality Observations by Method of Collection

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 sources with APIs were relatively few, they contained the most data: 56% of the observations were collected this way. In the absence of an API, I scripted downloading and reading any available spreadsheets (see Wickham 2016). When 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.[5]

Still, for a variety of reasons, a source's data may have been consigned to being entered manually in a separate spreadsheet.[6] 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 directly into R. Finally, one source contains crucial information encoded in the typeface of its tables (see Mitra and Yemtsiv 2006, 6), information lost when the tables are read directly into R. 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.[7]

In the end, I was able to automate the collection of 84% of the source data and an even higher percentage of the observations that will be updated or are subject to revision, greatly facilitating

[5]Code for the entire process can be viewed here: https://github.com/fsolt/swiid/blob/master/R/data_setup.R.
[6]See https://github.com/fsolt/swiid/blob/master/data-raw/fs_added_data.csv.
[7]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.

incorporating these changes in future versions.

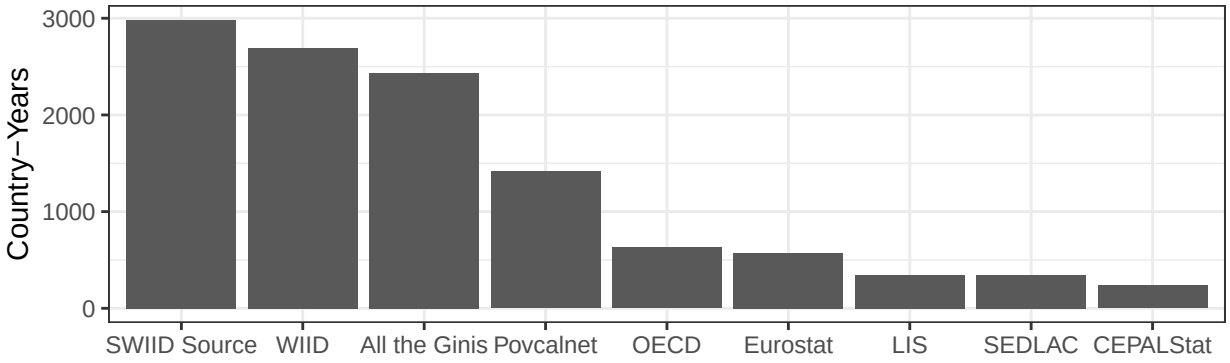

Figure 2: Income Inequality Datasets by Country-Years Covered

Data were collected from a total of 384 sources. But the fact that observations are drawn from a single source does not necessarily make them comparable: as Atkinson and Brandolini (2009, 392) observed, inequality statistics can suffer from "a break in continuity" if "the underlying source is the same, but the methods changed." For example, Eurostat (2019) indicates breaks in the time series of its data for France occurred in 2000, 2003, and 2007. Rather than a single consistent series, then, these three breaks mean the Eurostat data for France actually consist of four separate series (1994-1999, 2000-2002, 2003-2006, and 2007-2016) that are not comparable with each other. That these breaks result in year-to-year jumps ranging in size from an entire Gini index point (in 2000) to more than three points (in 2007) underscore the importance of taking such breaks in continuity into account. Defining a series, then, as a group of one or more observations from the same country calculated using the same methodology, the resulting dataset comprises 2,616 series.

All told, the SWIID source data include 15,730 Gini coefficients from 2,984 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 775 observations, many more than any other country. The United Kingdom, Germany, and the United States are next, followed by a group dominated by European 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.

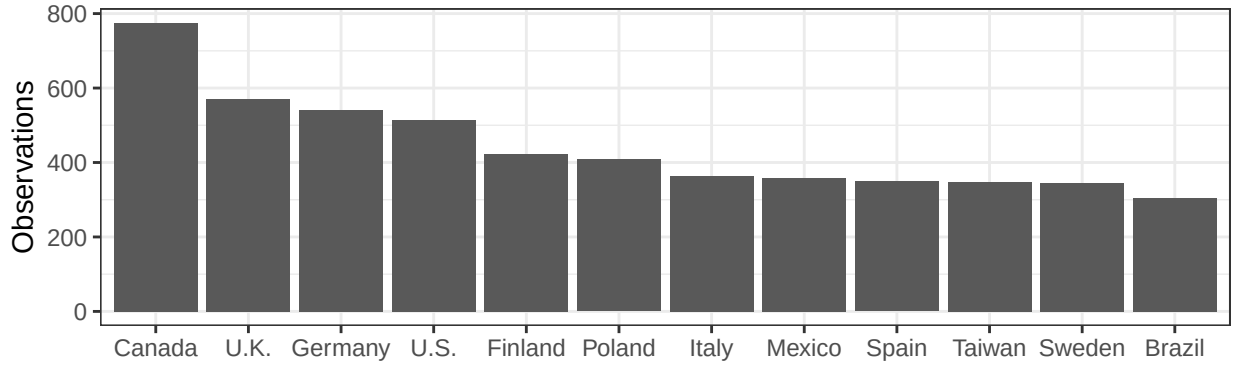

Figure 3: Countries with the Most Observations in the SWIID Source Data

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 59-year timeframe, from 1960 to 2018, 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 all but one covered year for Sweden and the United Kingdom. There are observations in all but six years for the United States, and in all but ten years for Japan and Taiwan. Iran and Argentina—countries that are not members of the LIS—also make the top ten, with observations in 45 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

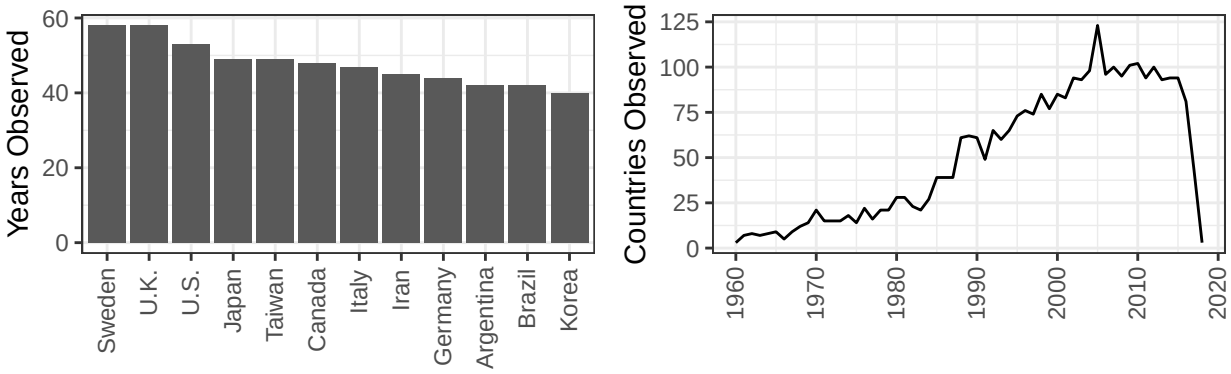

Figure 4: Country-Year Coverage in the SWIID Source Data

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 the broadest coverage. Coverage is relatively good in all of the years from 2000 to 2016, with at least 80 countries observed per year, before dropping to 43 countries for 2017 and only 3 for 2018.[8] Before then, country coverage is pretty thin each year through the 1960s and 1970s and still is not great until the late 1980s.[9]

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.[10] Gross income is the sum of market income and government transfer payments; it is 'pre-tax, post-transfer' income.[11] Disposable income, in turn, is gross income minus

[8]Data collection for version 8.1 was completed on May 20, 2019.

[9]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.

[10]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).

[11]In previous versions of the SWIID, market and gross income were treated as a single welfare definition, and I am

direct taxes: 'post-tax, post-transfer' income.[12] Consumption does not refer to the money coming into the household at all but rather to the money going out. 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 using other welfare definitions.

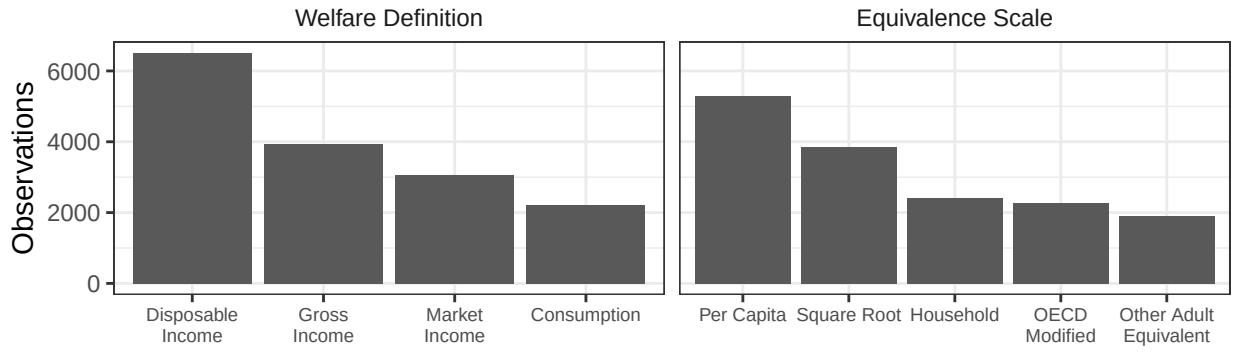

Figure 5: Welfare Definitions and Equivalence Scales in the SWIID Source Data

Equivalence scales are the ways in which the size and composition of a household are 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'[13] 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 among the most important reasons 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)

glad to finally be able to split them apart (cf. Solt 2016, 1272).

[12]Note that disposable income still does not take into account, on the one hand, indirect taxes such as 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'—a notable exception being that generated by the impressive work of the Commitment to Equity Institute (see Lustig 2018)—so I exclude it from the SWIID source data at least for the time being.

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

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.

All twenty combinations of welfare definition and equivalence scale are present in the source data. These broad differences between observations, along with less readily evident distinctions such as in the comprehensiveness of the income or consumption definitions employed or in the reporting period,[14] mean that they are far from comparable. So while the SWIID source data constitute the most comprehensive collection of the available observations of the distribution of income across the populations of the world's countries, they are not comparable and so not suitable for cross-national research. And if were we willing to put aside the less obvious incomparabilities across its observations and use the only data with most common combination of welfare definition and equivalence scale, disposable income per capita, we would be left with only 1766 observations, or just 11% of the total information available.[15] How to make use of all of these available data to estimate levels of income inequality that can be compared across countries and over time is the subject of the next section.

Generating Comparable Income Inequality Estimates

The starting point for the SWIID's estimates are two sets of Gini indices from the LIS, one of market income inequality and one of disposable income inequality, each using the square-root

[14]One source of differences within a single welfare definition is the manner and extent to which nonmonetary income or expenditure—such as the value of food grown for the household's own consumption or of housing that the owner occupies—is included. Consumption definitions can vary considerably in whether and how observations treat expenditures on durable goods. Income definitions sometimes exclude such major categories as interest and dividends (see Atkinson and Brandolini 2001, 785) or self-employment income (see Checchi et al. 2018, 6).

[15]This seemingly sensible approach, in addition to discarding the vast majority of the available information, also does a surprisingly poor job at yielding comparable income inequality figures (see Solt 2016, 1278) and can raise issues of problematic researcher degrees of freedom (see Solt 2015, 686). Moreover, if the welfare concept relevant to our theory were in fact market income, these data would not be fit for our purposes regardless (see Atkinson and Brandolini 2009, 389, 393).

equivalence scale (LIS 2018). The LIS has meticulously harmonized its survey microdata to be as comparable cross-nationally as possible; inequality observations calculated from these harmonized microdata therefore serve as the baselines for the SWIID.[16] These two LIS baselines anchor the SWIID's estimates of market and disposable income inequality.[17] The SWIID routine estimates the relationships between these LIS baseline Gini indices and all of the other Ginis available for the same country-years and uses these relationships to estimate what the LIS Gini would be in country-years not included in the LIS but available from other sources.

These relationships fall into four categories, varying in accordance with how much information is available: two for observations of countries included in the LIS and two more for countries that, as of yet, are not included. First, for each series s—again defining a series as a group of observations from the same country calculated using the same methodology—that includes more than two country-years that also appear in the LIS baseline,[18] the ratio $\hat{\rho}_s$ such that

$$G_{bkt} \sim \mathcal{N}(\hat{\rho}_s \times G_{skt},\, \sigma_s^2)$$

where G_{bkt} is the LIS Gini for baseline welfare definition b in country k at time t, G_{skt} is the Gini from series s in that same country k and time t, and σ_s^2 is the series-specific variance of the error term. That is, $\hat{\rho}_s$ is the estimated ratio of the LIS baseline to the series s in all years that the two overlap. This relationship is not only country-specific and includes the welfare definition and equivalence scale, but because each series uses a consistent methodology, such factors as the comprehensiveness of the income definition and the reporting period are also taken into account. These $\hat{\rho}_s$ can be estimated for series encompassing nearly half of the observations and over 70% of the observed country-years in the SWIID source data for the countries included in the LIS. As

[16]These Ginis are calculated directly from the LIS microdata on the same basis as that used for the LIS Key Indicators. The code employed for this purpose can be found here: https://github.com/fsolt/swiid/blob/master/R/lissy.R. Most controversially, this procedure truncates high incomes at ten times the median income; this is done in an effort to maximize comparability across different countries given that the raw data may have been subject to different degrees of such top coding before being submitted to the LIS. The tradeoff, however, is that Ginis calculated with top coding underestimate the true extent of income inequality. I am grateful to Zsolt Darvas for drawing my attention to this issue.

[17]Although New Zealand does not participate in the LIS, I treat the four observations of disposable income inequality (1982, 1986, 1991, 1996) specially prepared by Statistics New Zealand Statistics New Zealand (1999, 73) to be comparable with the LIS to be part of the collection of disposable income inequality statistics from the LIS.

[18]Because the two LIS baselines do not contain the identical country-years (the disposable income inequality baseline has 351 observations, while that for market income inequality has only 342), these series vary somewhat when estimates for each of the two welfare definitions are being generated.

shown in the column labeled "Total" at the right of Figure 6, these country-years constitute just over a third of all observed country-years in the SWIID source data. This figure also reveals that the extent to which series-specific relationships can be estimated vary greatly by region: $\hat{\rho}_s$ can be estimated for over 70% of the observed country-years in the advanced English-speaking countries and nearly 60% of those in Western Europe, but only about 7% of the country-years in the countries of Africa and developing Asia.

The SWIID routine next expands the observations used to estimate the relationship between the source data and the LIS baseline to include all of the observations in the country with the same welfare definition and equivalence scale. The relationship $\hat{\rho}_{kwe}$ is the estimated ratio of the LIS baseline to Ginis with the same country-year for each country k, welfare definition w, and equivalence scale e. Because it requires only that an observation's combination of welfare definition and equivalence scale—rather than the observation's series—share country-years with the LIS baseline data, $\hat{\rho}_{kwe}$ gives up the benefits of harmonization and consequently yields more uncertain estimates of these ratios. On the other hand, the advantage of $\hat{\rho}_{kwe}$ is that it can be estimated for all of the remaining observations in the source data for countries included in the LIS, some 7% of all the observed country-years.[19] The share of observed country-years for which $\hat{\rho}_{kwe}$ provides the best estimate ranges from 35% for Japan and the Asian Tigers to just over 3% among the developing ex-Communist countries.

However, $\hat{\rho}_{kwe}$ cannot be estimated for any of the observations in the countries that are not included in the LIS: no matter how rich the country's data in any combination of welfare definition and equivalence scale, those data will have no country-years in common with the LIS baseline. For those countries, when possible, the relationship between source data observations and the LIS baseline data is estimated as the product of two factors. The first factor seeks to account for differences in welfare definition—the policy-sensitive effects of taxes and transfers—by reference only to data from the country in question. For each country k and welfare definition w, $\hat{\rho}_{kw}$ is the estimate of the ratio of Ginis with the baseline welfare definition to those with the welfare definition w in the same country-year. The second factor, to account for the much smaller differences across

[19] Previous versions of the SWIID allowed $\hat{\rho}_{kwe}$ to vary over time, but much of what those parameters was capturing was the breaks in continuity across different series within each combination of welfare definition and equivalence scale; this variation, of course, is now parsed directly by $\hat{\rho}_s$. Allowing $\hat{\rho}_s$ to vary by time for series of sufficient length is an enhancement planned for future versions of the SWIID.

equivalence scales, is $\hat{\rho}_{re}$. The ratio $\hat{\rho}_{re}$ is the estimate of the ratio of LIS baseline observations to Ginis of the same country-year with the baseline welfare definition and equivalence scale e across the region r in which country k is located.[20] The product of $\hat{\rho}_{kw}$ and $\hat{\rho}_{re}$, then, estimates the relationship between a Gini in country k in region r with welfare definition w and equivalence scale e and the LIS baseline. This product can be estimated for more than two-thirds of the source data observations for countries outside the LIS, though these constitute only 57% of the observed country-years for those countries. Across all countries, as can be seen in the right-most column of Figure 6, these are 27% of the observed country-years in the SWIID source data. Less than 10% of the observed country-years for the countries of Western Europe—but nearly 60% of those for the developing ex-Communist countries—are best estimated in this fashion.

Lastly, where the source data for the country of an observation does not allow the relationship $\hat{\rho}_{kw}$ across welfare definitions to be estimated from in-country data as described in the preceding paragraph, the estimated relationship $\hat{\rho}_{rwe}$ is used: the estimated ratio of the LIS baseline to Ginis with the same country-year for welfare definition w and equivalence scale e across the same region r as the country of the observation. Across all regions, data scarcity imposes this unfortunate reliance upon information only from other countries in the region to estimate the relationship between observed source data and the LIS baseline in 20% of the observed country-years in the SWIID source data. None of the relationships for the country-years of the advanced English-speaking countries rely on $\hat{\rho}_{rwe}$ (or indeed any information beyond in-country data); however, the relationships to the LIS baseline of nearly 60% of the observed country-years for the countries of Africa and developing Asia are based on such regionwide data.

Although often neglected, like any other statistic calculated from a survey sample rather than the entire population of interest, the Gini has an associated standard error. In the source data, 44% of the observations include standard errors. The 99th percentile of the relative standard errors of these

[20]The regions used are defined as follows: (1) the advanced English-speaking countries, (2) Western Europe, (3) Japan and the Asian Tigers, (4) the advanced ex-Communist countries, (5) Latin America, (6) the developing ex-Communist countries, and (7) Africa and developing Asia. This follows the practice of previous versions of the SWIID with the exception that Africa and developing Asia, once considered separately, are now considered a single region. The combination of the relative paucity of data and the heterogeneity of relationships observed—particularly between consumption and other welfare definitions—among the countries of these two continents counseled in favor of considering them together. Cross-validation tests confirmed that doing so best preserves the uncertainty implied by the scarcity and heterogeneity of the data available; as more LIS data becomes available, the question will be re-examined. The next section discusses cross-validation in detail. The complete list of countries and regions may be consulted at https://github.com/fsolt/swiid/blob/master/data/reg.csv.

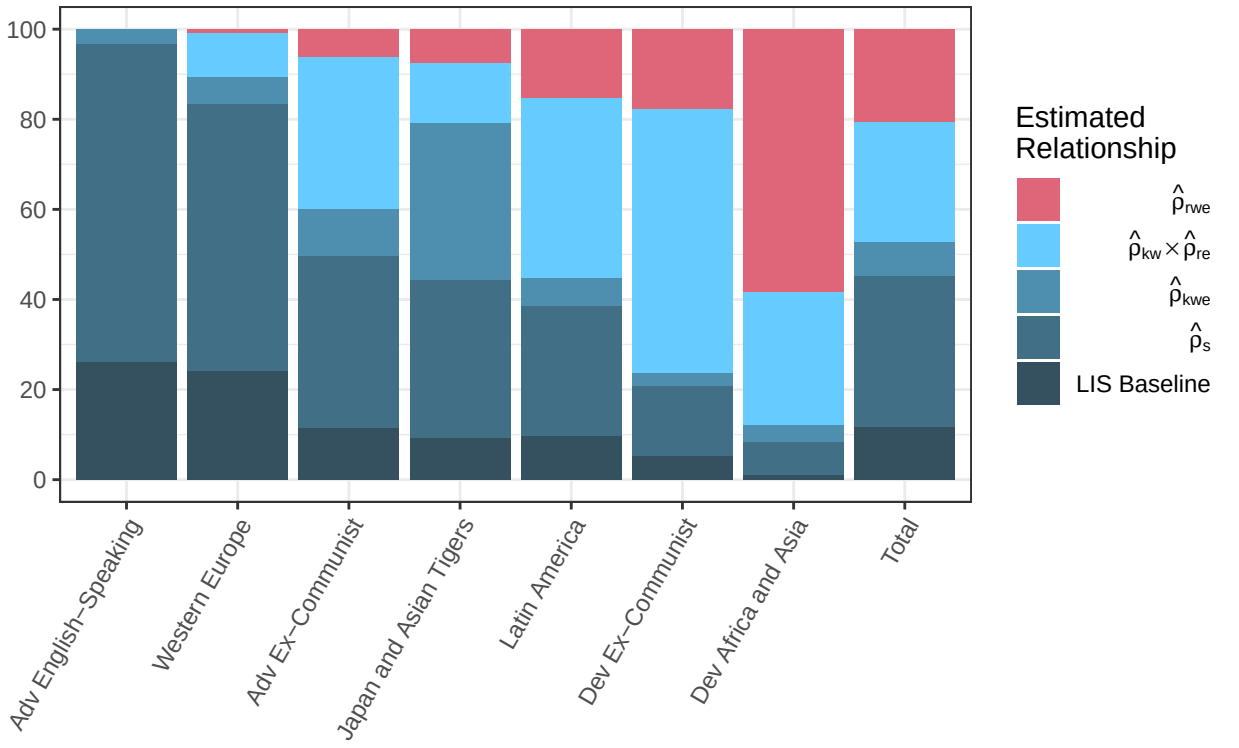

Figure 6: Estimated Relationships by Region, Observed Country-Years

observations was used to provide a (very conservative) imputed value for the remaining observations. This measurement error was taken into account before estimating any of the relationships described above by modelling the true quantity of each observation in the source data as normally distributed around the observed Gini with deviation equal to the associated standard error.

To generate $\hat{G}_{bkt}$, the SWIID estimate of the LIS baseline based on the available source data for countries and years in which the LIS baseline is not available, I employ a fully Bayesian approach. To take into account the fact that Ginis generally change only gradually from one year to the next (see, e.g., Atkinson and Brandolini 2009, 392), I specify a random walk prior process: within each country k, the SWIID estimate in year t has a normal prior distribution that is centered on its value in the year $t-1$:[21]

$$\hat{G}_{bkt} \sim \mathcal{N}(\hat{G}_{bk(t-1)}, \sigma_G^2)$$

Estimates for country-years with LIS baseline data are updated from a normal distribution centered

[21]If a LIS baseline observation is available in the first observed year for country k, this data and its associated standard error are used as a normally distributed informative prior; otherwise, a weakly informative, lognormally distributed prior with $\mu = -1$ and $\sigma = .25$ is employed.

on the observed Gini with deviation equal to its associated standard error. Estimates for country-years without LIS baseline data but with other source-data observations are updated using $\hat{\rho}_s$, $\hat{\rho}_{kwe}$, the product of $\hat{\rho}_{kw}$ and $\hat{\rho}_{re}$, and $\hat{\rho}_{rwe}$. For example, if G_{skt} is present in the source data for a country-year kt without LIS baseline data, then $\hat{G}_{bkt}$ is updated using $\hat{\rho}_s$ and σ_s^2 (recall that these quantities are estimated from those country-years in which both the LIS baseline data and series s are observed):

$$\hat{G}_{bkt} \sim \mathcal{N}(\hat{\rho}_s \times G_{skt},\, \sigma_s^2)$$

Country-years with more observations in the source data have more information to update the estimate. The random-walk specification smoothly manages those country-years without any Ginis in the source data between each country's first and last observed years: if no observation is available, the random walk process spans the gap between the last year to appear in the source data and the next year to appear.

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 put the SWIID to the most stringent test that occurred to me: I compared LIS data on country-years that had been included in previously-released versions of the SWIID (Solt 2016, 1277-1278).[22] 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 (2014) database or by the *All the Ginis* dataset adjusted in accordance with its instructions (Milanovic 2013, 8)—but less so in others. Most disappointingly, only 72% of those differences had 95% confidence intervals that included zero, suggesting that the SWIID's standard errors were often too small. The new SWIID estimation routine described in the preceding section was written and revised to address these issues, but the difficulty of the work done by the LIS team to add new observations means that those additions do not come quickly enough to allow for continuous testing of the SWIID's revisions. So, instead, I have drawn on a technique developed in data science and machine learning,

[22]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.

k-fold cross-validation, to assess the SWIID's progress.

To understand how k-fold cross-validation works, it helps to first consider 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.[23] (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 does not really matter, a point I confirmed in testing.) 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 does not are highlighted in blue.

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 85% less than 2 Gini points and 62% less

[23]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.

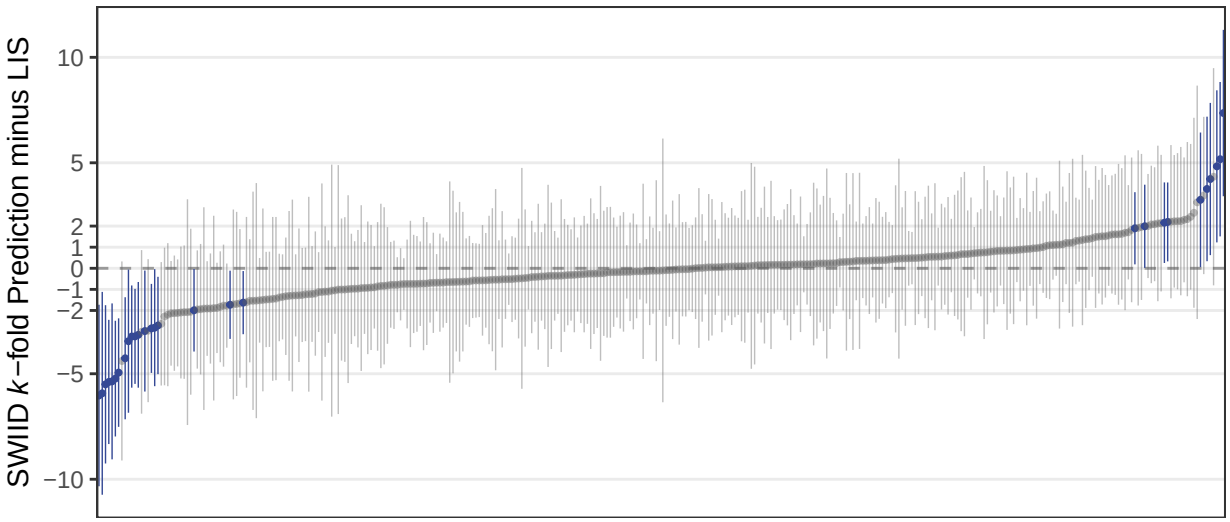

Figure 7: k-fold By-Observation Cross-Validation Results

than a single Gini point. It is 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 does not seem 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 this test are plotted below in Figure 8.

Overall, the plot looks very similar to the one above in Figure 7. With each country's entire run of LIS data being excluded in turn, 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: 76% are less than 2 Gini points, and 51% are less than one Gini point.

This analysis, though, does point to two areas that are 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

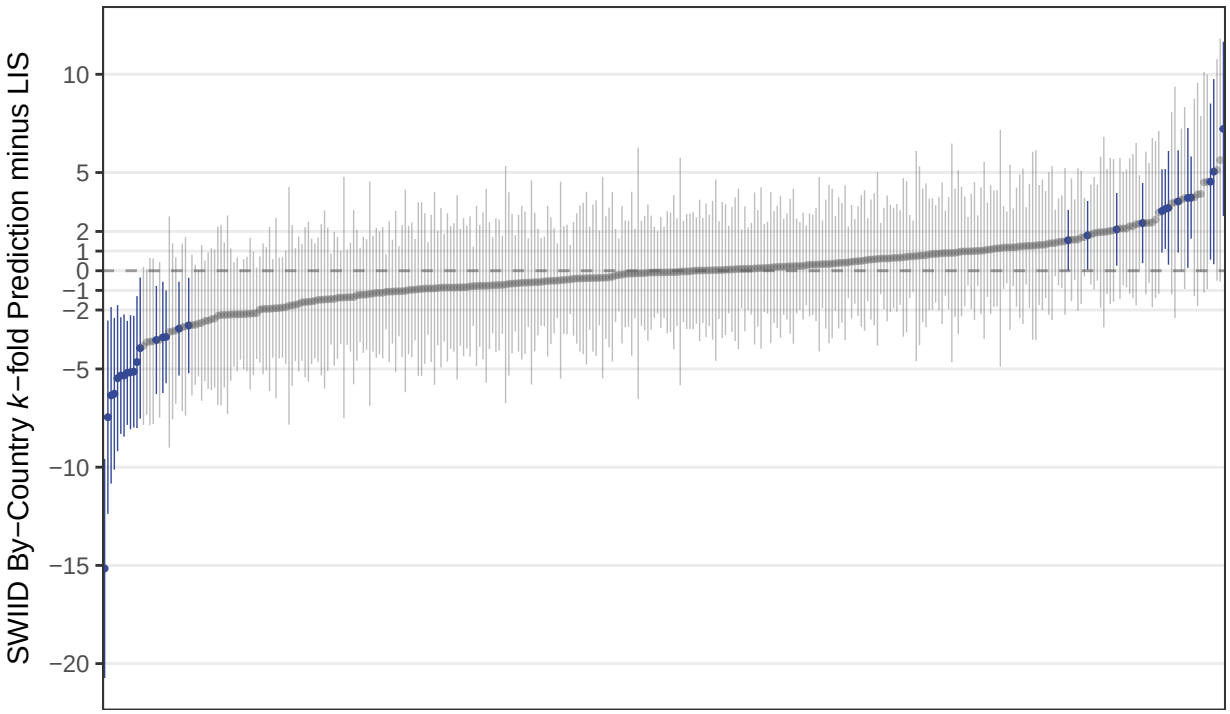

Figure 8: k-fold By-Country Cross-Validation Results

Gini points. Egypt is currently the only country in the LIS with only a single observation; given that excluding this one observation is equivalent to excluding all of the country's observations, it was not considered in the first cross-validation. LIS researchers Checchi et al. (2018, 6) report that Egyptian income surveys before 2012 did not include any questions to capture self-employment income, and it is 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 (as can be seen in the SWIID source data for, e.g., India). These factors, however, are present in many non-LIS countries as well. Finding ways to improve the SWIID routine to address them will be a priority.

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 2.4 ± 2.1 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 5.2 ± 3.0 and 5.3 ± 3.0 Gini points. There is room for improvement here too, and this too will receive continued efforts.

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.

Conclusion

References

Atkinson, Anthony B. and Andrea Brandolini. 2001. "Promise and Pitfalls in the Use of 'Secondary' Data-Sets: Income Inequality in OECD Countries as a Case Study." *Journal of Economic Literature* 39(3):771–799.

Atkinson, Anthony B. and Andrea Brandolini. 2009. "On Data: A Case Study of the Evolution of Income Inequality Across Time and Across Countries." *Cambridge Journal of Economics* 33(3):381–404.

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

Checchi, Daniele, Andrej Cupak, Teresa Munzi and Janet Gornick. 2018. "Empirical Challenges Comparing Inequality Across Countries: The Case of Middle-Income Countries from the LIS Database." *WIDER Working Paper* 2018/149. Helsinki: UNU-WIDER. https://www.wider.unu.edu/sites/default/files/Publications/Working-paper/PDF/wp2018-149.pdf.

Eurostat. 2019. "Gini Coefficient of Equivalised Disposable Income, ilc_di12." `http://appsso.eurostat.ec.europa.eu/nui/show.do?dataset=ilc_di12&lang=en`. February 19, 2019.

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.

LIS. 2018. "Luxembourg Income Study Database." `http://www.lisdatacenter.org`. Multiple countries; December 2018. Luxembourg: LIS.

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

Lustig, Nora, ed. 2018. *Commitment to Equity Handbook: Estimating the Impact of Fiscal Policy on Inequality and Poverty*. New Orleans and Washington: CEQ Institute at Tulane University and Brookings Institution Press.

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

Milanovic, Branko. 2013. "Description of *All The Ginis* Dataset." World Bank, Research Department.

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. 2015. "On the Assessment and Use of Cross-National Income Inequality Datasets." *Journal of Economic Inequality* .

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

Statistics New Zealand. 1999. *New Zealand Now: Incomes*. Wellington: Statistics New Zealand.

UNU-WIDER. 2014. "World Income Inequality Database, Version 3b, September 2014." `http://www.wider.unu.edu/research/WIID3-0B/en_GB/database`.

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`.