World Income Inequality Databases: an assessment of WIID and SWIID

Stephen P. Jenkins

Email: s.jenkins@lse.ac.uk

Note. The full paper (long) will be available shortly as a discussion paper in the ISER (U of Essex) and IZA series. This talk (short) covers only some of the material in the paper

LSE

World Income Inequality Databases

Secondary data compilations of inequality statistics, especially Ginis, and widely-used (these versions and earlier ones)

WIID2c (2008)

[UNU-WIDER]

- 161 countries
- 1867–2006
- Quality ratings (4 ratings)
- Ginis based on different definitions and sources
- Missing country-year obs

SWIID4.0 (2013)

[Frederick Solt]

- Based on WIID, plus extra
- 173 countries
- 1980–2010
- Quality ratings not used
- Standardized ‘net income Gini’ definition
- No missing country-year obs
  - Multiple imputation model used to ‘fill in the gaps’
  - All obs are imputed
  - 100 MI data sets in Main file (with Gini means in Summary file)

LSE

World Income Inequality Databases have advantages and disadvantages

Advantages

- Global coverage of countries
- Long time period covered

Disadvantages

- Data non-comparabilities
- Data quality, more generally
- Missing data (WIID)

My paper:

- Takes the advantages as given
- Comments on file content and documentation (not today)
- Reviews the disadvantages in detail, with illustrations
- Advises users how to minimize their impact
  - Nature of WIID and SWIID implies different approaches

Headline conclusions

1. Comparability and quality issues raised by Atkinson & Brandolini (2001, 2009) w.r.t. WIID-predecessor (Deininger-Squire data set) remain very relevant
2. WIID users must report the details of their country-year selection algorithms and justify the choices made
3. WIID regression-based adjustments to account for non-comparabilities need to be more sophisticated than the commonly-used simple dummy variable approach
4. SWIID "provides plausible data but not sufficiently credible data"
   - Concerns about the imputation model per se (bias issue)
   - But ignoring the MI nature of the data appears not to lead to big differences in SEs (precision issue)
5. Overall, I recommend WIID over SWIID
   - Support is conditional on proper attention being given to data issues

LSE

Data issues when comparing Ginis

Non-comparabilities in definitions of distributions

- Resource measure
  - e.g. income vs consumption vs earnings
- Reference period
  - e.g. month vs year
- Sharing unit
  - e.g. household, family, person
- Equivalisation
  - e.g. per capita, OECD scales
- Unit of analysis
  - e.g. distribution among individuals or households

Nature of data source and pre-calculation adjustments

- Source type
  - e.g. survey, admin records
- Coverage of people
  - e.g. population vs prime-aged
- Coverage of areas
  - e.g. country vs urban or rural
- Representativeness and other quality of collection issues
- Treatment of data
  - e.g. continuous vs banded; top-coding; trimming; Gini formulae

LSE

The problematic Quality-Coverage trade-off

- The more global the coverage, the greater the prevalence of poorer quality data that are included

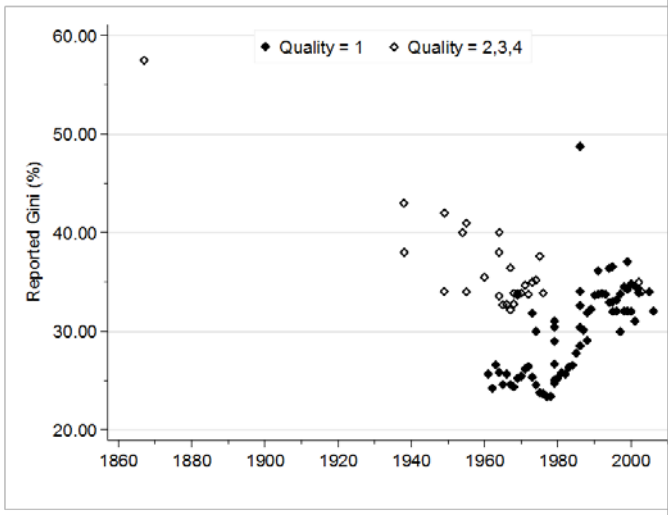
Table 1. WIID: number of country-year observations, by geographical region and year

Region	Period							
	1867 –1899	1900 –1959	1960 –1969	1970 –1979	1980 –1989	1990 –1999	2000 –2006	*Total*
	All observations							
Africa	0	28	61	56	67	140	26	*378*
Western Europe (EU15)	1	54	98	141	235	342	182	*1,053*
Other Europe, Turkey, Russia	0	11	68	72	185	483	231	*1,050*
North America	0	17	25	35	53	51	10	*191*
Central & South America	0	34	154	177	197	424	124	*1,110*
Central, East, & South East Asia	1	96	188	210	280	288	85	*1,148*
Oceania	0	42	42	43	45	55	11	*238*
Middle East	0	20	19	30	22	23	9	*123*
Total	*2*	*302*	*655*	*764*	*1,084*	*1,806*	*678*	*5,291*
	Observations with Quality = 1							
Africa	0	0	0	0	3	2	0	*5*
Western Europe (EU15)	0	2	19	72	163	293	170	*719*
Other Europe, Turkey, Russia	0	4	5	10	17	135	95	*266*
North America	0	14	16	28	44	42	9	*153*
Central & South America	0	0	0	2	15	40	8	*65*
Central, East, & South East Asia	0	0	5	15	39	53	8	*120*
Oceania	0	0	0	0	18	28	7	*53*
Middle East	0	0	0	2	2	13	3	*20*
Total	*0*	*20*	*45*	*129*	*301*	*606*	*300*	*1,401*

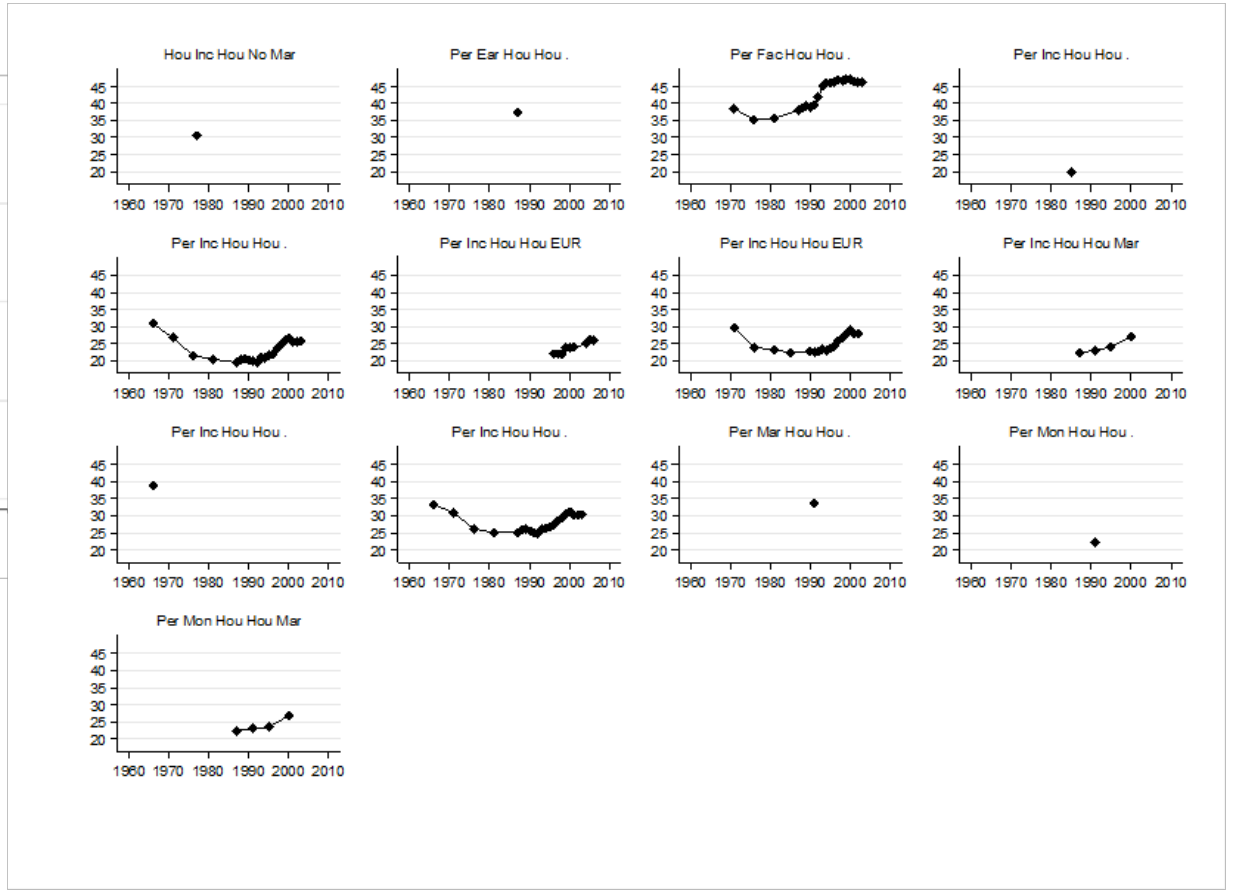
Notes. The classification excludes 22 country-year observations with multi-year 'year' values. All observations classified in the table have non-missing observations on Reported Gini. 'Quality = 1' refers to the highest WIID data quality classification. See main text for details.

Multiple data series (different definitions) and multiple observations per country-year cell ⇒ selection algorithms needed

WIID: United Kingdom

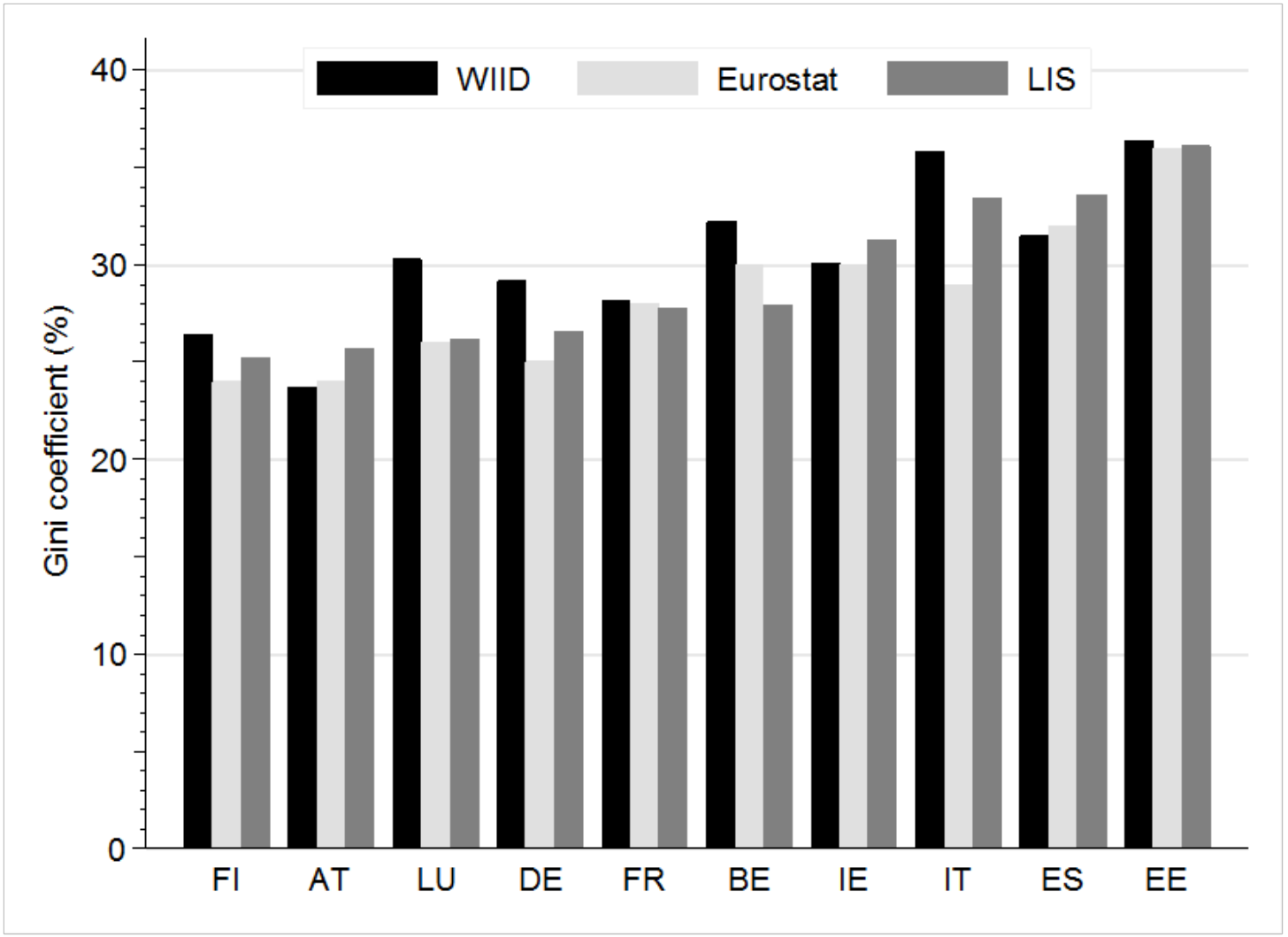


WIID: Finland (*Quality* = 1 obs)



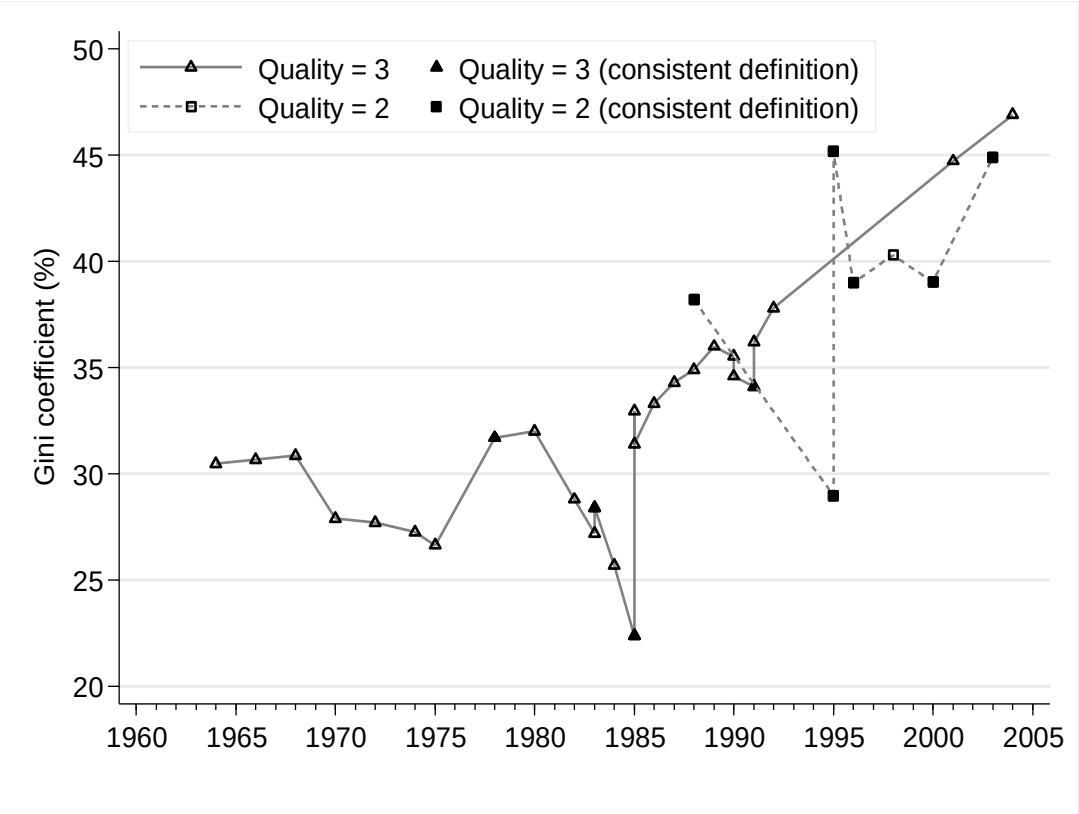
Benchmarking WIID: cross-sectional

- Even with tight selections focusing on obs with relatively homogeneous definitions and for same year (2000), some quite large differences levels and country-rankings appear:



WIID: assessing trends (China example)

- The Quality-Coverage conundrum again
  - Long series available only for poor(er) quality obs
- Multiple obs per year, even when income definitions apparently the same (e.g. 1995!)
- Differences between WIID and official series and – for recent years – several other household surveys (Xie & Zhou, *PNAS* 2014)



Reported Gini. All observations with *AreaCovr* = 'All'.
Subsets of observations with 'consistent definition' are those for which, in addition, *UofAnala* = 'Person' and *IncDefn* = 'Income, Disposable'.

SWIID: imputation model

- Selections and exclusions (e.g. drop pre-1960 WIID obs)
- Imputation procedure: key idea summarized:
  - Suppose there are two data series for the Gini coefficient available for a large number of country-year observations, one based on gross income and the other on net income, but some estimates are missing for the net income Gini
  - If the **ratio of Ginis** for net income to gross income were **constant within some group** g **of country-year observations**, and one had an estimate of that ratio, call it R_g, then one could impute the missing values
  - The net income Gini imputation for a particular country-year observation within group g is equal to its observed gross income Gini multiplied by ratio R_g
  - Repeating multiple times $\rightarrow$ **multiple imputations** (multiple distributions of estimated Ginis)
- Imputation procedure: much more complicated than this, e.g.:
  - Regression-based
  - c. 20 data 'types' (many series of Gini ratios)
  - definition of 'group' varies (and unclear)
  - various other steps as well (including MA smoothing)
  - also yields estimates of 'share of richest 1%'

SWIID's imputations: basic problem

- Assumes constancy of ratios of Ginis across data series within groups of country-year observations
  - NB Multiplicative version of the "dummy variable adjustment" procedure that assumes constant absolute differences between series (used a lot by WIID analysts)
- Two competing demands that cannot both be met
  1. Country-year observations have to be grouped in order to have donor observations to provide the values to be imputed to the missing observations and, other things being equal, the larger the group size, the more reliable is the within-group mean used for the imputation. But, …
  2. Need as many groups as possible to allow for the acknowledged variation in Gini ratios but, other things being equal, having more groups means a smaller average group size and, in the limit, no potential donor observations.
  - Given available source data, groups are relatively broadly defined in SWIID, and so the assumption of within-group constancy in Gini ratios is very likely to be compromised
    - NB The same is, of course, likely to be true for Gini differences, which means that regression-based adjustments to WIID data for differences in variable definitions need to more sophisticated than simple intercept shifts
    - Regression-based adjustments can be more transparent and also adapted to context (SWIID provides a general all-purpose solution, and not transparent)

LSE

SWIID's imputations: other issues

Including …

- Imposition of 5-year moving-average smooth
- Definitions of data 'types' (series)
- Bug in calculation of 'share of top 1%' series
  - Don't use these data (see Figure 11)

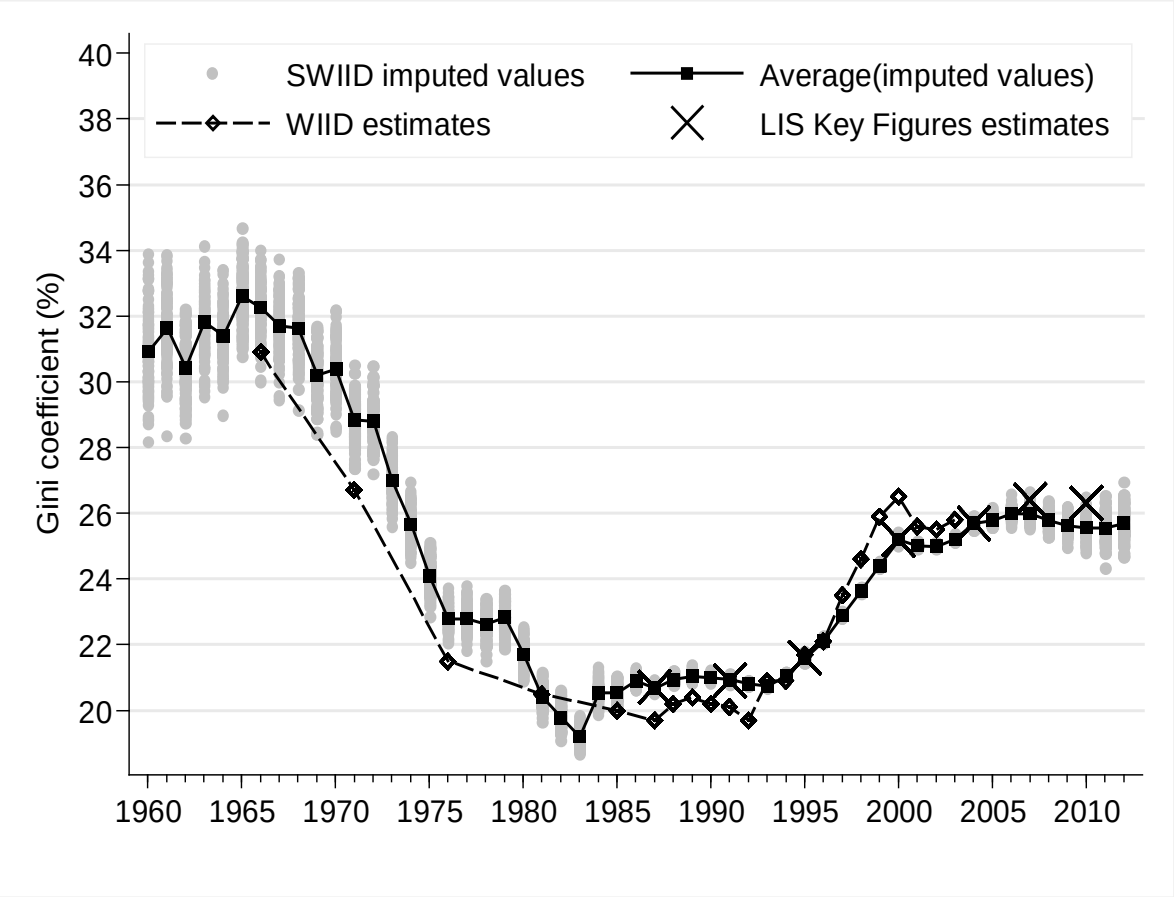
See paper for further details

- Also applaud Frederick Solt's provision of "replication script"

SWIID compared to other estimates: Finland

‘Net income Gini’

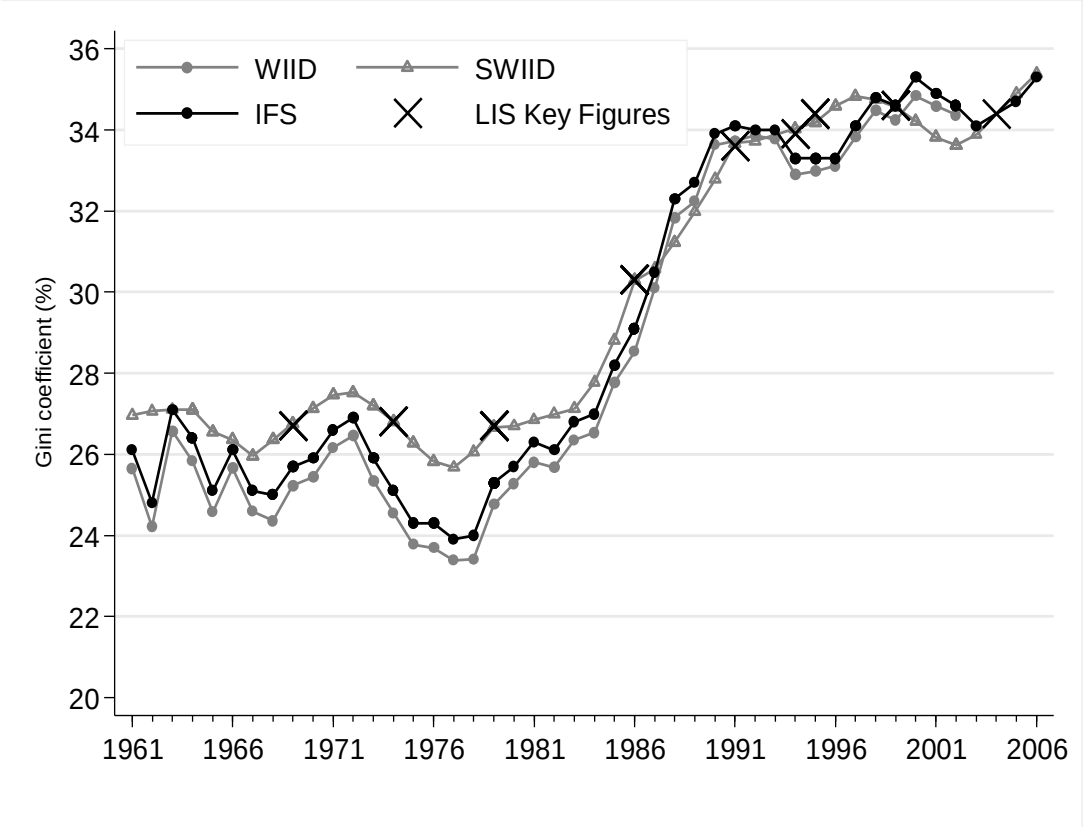
- Compare high quality external estimates from WIID and LIS Key Figures with SWIID
- Note differences in levels and trends



SWIID compared to other estimates: UK

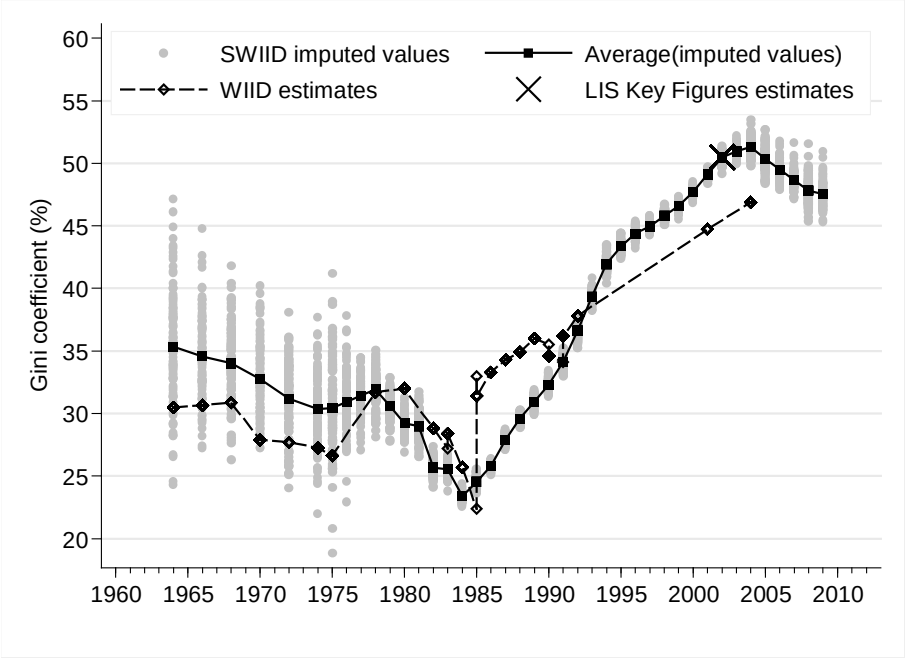
'Net income Gini'

- Compare high quality external estimates from WIID and IFS (both are UK 'official' series) with SWIID
- SWIID estimates are mean values (full range not shown for legibility)
- Note differences in levels and trends

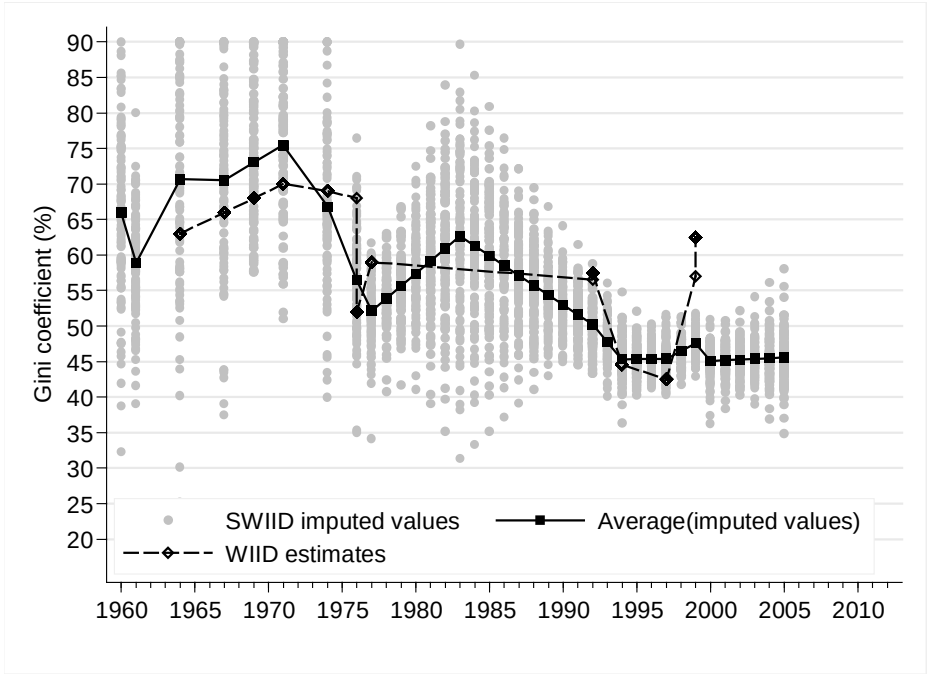


SWIID compared to other estimates: CN, KE

China



Kenya



Note. The WIID estimates shown for each country are based on all observations with *Quality* = 3 and *AreaCvr* = 'All'. All other WIID observations for Kenya are of lower quality. The shorter *Quality* = 2 WIID series for China is shown in Figure 6.

- Range of imputed values for a given year can be huge!
- Differences across series

Regression illustrations (Tables 4–6)

1. Regress Gini on unemployment rate, inflation rate, time trend (cf. 'Blinder-Esaki' literature): various samples pooling countries and years; various sources (WIID, SWIID, Eurostat, LIS)
2. Regress Gini on decade dummies for each of number of countries

- Changing the source for the Gini (and using different samples of countries) can lead to big differences in estimated coefficients and statistical significance
- WIID-based and SWIID-based (and other) estimates are similar if one uses homogenous sample (EU-15)
- SWIID: can't assess bias (no external sources by definition)
- SWIID: SEs of coefficients much the same if (a) use mean Gini and ignore MI; or (b) take proper account of MI variability (`mi estimate:` in Stata)

LSE

Headline conclusions

1. Comparability and quality issues raised by Atkinson & Brandolini (2001, 2009) w.r.t. WIID-predecessor (Deininger-Squire data set) remain very relevant
2. WIID users must report the details of their country-year selection algorithms and justify the choices made
3. WIID regression-based adjustments to account for non-comparabilities need to be more sophisticated than the commonly-used simple dummy variable approach
4. SWIID "provides plausible data but not sufficiently credible data"
   - Concerns about the imputation model per se (bias issue)
   - But ignoring the MI nature of the data appears not to lead to big differences in SEs (precision issue)
5. Overall, I recommend WIID over SWIID
   - Support is conditional on proper attention being given to data issues

LSE