Income Database of Harmonized Microdata: LIS- Cross National Data Center in Luxembourg

Think thanks Summit
Zurich, 22 January 2019

Outline of the presentation

I. LIS, the institution

II. The LIS/LWS Databases

III. Research possibilities

IV. Data dissemination & Documentation

V. LIS golden rules for harmonization

Part I:
LIS, the institution

LIS History

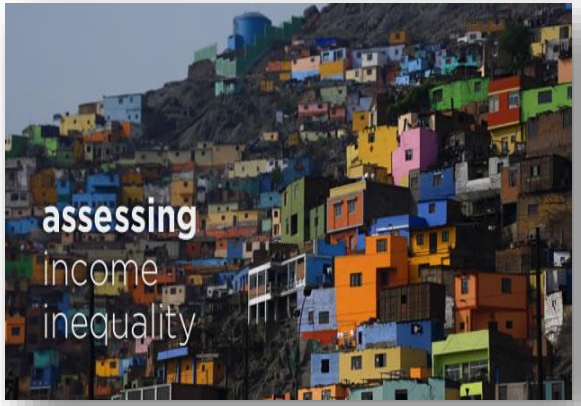
- LIS was founded in 1983 by two US academics (Professors Tim Smeeding and Lee Rainwater) and a team of multi-disciplinary researchers in Europe. It began as a “study”, which later grew and was institutionalized as “LIS”. (Note the “oral history” video on LIS website.)

- For nearly 20 years, LIS was part of a Luxembourg-based research institute known as CEPS (recently renamed LISER). In 2002, LIS became an independent non-profit institution (an ASBL).

- LIS is supported by the Luxembourg government (25%), by the national science foundations and other funders in many of the participating countries (50%), and by several supranational organizations – supplemented by project-specific grants and some private philanthropy (25%).

LIS Mission

To enable, facilitate, promote, and conduct cross-national comparative research on socio-economic outcomes and on the institutional factors that shape those outcomes.



LIS
CROSS-NATIONAL
DATA CENTER
in Luxembourg

LIS: new leadership structure

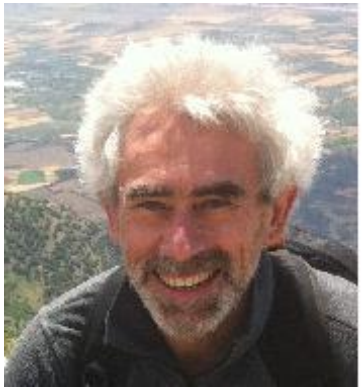
launched 1 September, 2016

Prof. Daniele Checchi

Director of Luxembourg Office of LIS

* He is Professor of Economics currently on leave from the University of Milan.

* He is currently working at the Italian National agency for the Evaluation of the University system.



Prof. Janet Gornick

Director of US Office of LIS

* She is Professor of Political Science and Sociology, at the Graduate Center of the City University of New York.

* The “LIS Center” has been renamed “The Stone Center on Socio-Economic Inequality”.

* The “US Office of LIS” is now an entity within the enlarged Center.

LIS structure

LIS - Cross-National Data Center
François Bourguignon, President
LIS Executive Committee
Daniele Checchi, Secretary General

MAIN OFFICE: Luxembourg Office of LIS Daniele Checchi, Director	SATELLITE OFFICE: US Office of LIS Janet Gornick, Director
• parent organization • located in Luxembourg, together with the University and LISER • independent, chartered non-profit organization • cross-national, participatory governance • acquires, harmonizes, and disseminates data for research • venue for research, visiting scholars, conferences, and user training	• satellite office • located at the Graduate Center of the City University of New York • one pillar of *Stone Center on Socio-Economic Inequality* • administrative, managerial, development support to parent office • collaborative public programs (lectures, conversations) • venue for research, teaching, and graduate student supervision

LIS Senior Scholars

LIS: who's who?

LIS - Cross-National Data Center
François Bourguignon, President
LIS Executive Committee
Daniele Checchi, Secretary General

MAIN OFFICE: Luxembourg Office of LIS **Daniele Checchi, Director**	SATELLITE OFFICE: US Office of LIS **Janet Gornick, Director**
Thierry Kruten, Director of Operations and IT Lucie Scapoli, Administrator Officer Benjamin Gérard, System and Network Administrator Data team: Teresa Munzi, Data Team Manager Paul Alkemade, Heba Omar Andrej Cupak, Piotr Paradowski Jörg Neugschwender, Carmen Petrovici	Caroline Batzdorf, Assistant Director Mei-Ling Israel, Financial Manager Laurie Maldonado, Senior Administrator

LIS Senior Scholars	
Prof. Louis Chauvel Prof. Daniele Checchi Prof. Conchita D'Ambrosio Prof. Markus Jäntti Prof. Frank Cowell	Prof. Janet Gornick Prof. Paul Krugman Prof. Leslie McCall Prof. Branko Milanovic

LIS' partners

Our partners include data providers, data users, and funders, in more than 50 countries …

and in major supranational organizations, including:

Financial contributors:

World Bank (WB)
Organization for Economic Cooperation and Development (OECD)
International Monetary Fund (IMF)
International Labor Organization (ILO)
European Union (InGRID)

Dataset exchange; joint research projects; joint fundraising:

Economic Research Forum (ERF)
European Central Bank (ECB)
French Development Agency (AFD)
LISER
University of Luxembourg

Part II:
The *LIS/LWS Databases*

Overview of the LIS data

Deliverables

Two cross-national harmonised databases that allow international comparative research using micro-data:

- LIS (focus on income): 339 datasets – this presentation will focus on LIS
- LWS (focus on wealth): 39 datasets

Scope

Initial focus on high-income countries, successively extended to middle-income countries

Time span

From the late 1960s to up-to-date

Geographical coverage

World-wide, but some regions are less covered (Africa, EECCA…)

Main contents

- household composition and characteristics
- socio-demographic characteristics of household members
- extensive set of labour market data
- detailed breakdown of household and individual income data
- household consumption data
- a detailed set of wealth and behavioural variables (LWS only)

Luxembourg Income Study Database (LIS)

- First and largest available database of harmonized income data, available at the household and person levels
- In existence since 1983
- Data mostly start in 1980, some go back to the 1960s (recollected every 3-5 years)
- Started with six countries; now 50 countries
- 300+ datasets (repeated cross sections)
- Used to study: poverty; income inequality; labor market outcomes; policy effects

Luxembourg Wealth Study Database (LWS)

- First available database of harmonized wealth data, available at the household level
- In existence since 2007
- 52 datasets from 16 countries – up or in process
- Revised and updated in 2016 (in coordination with Euro system's Household Finance and Consumption Survey - HFCS)
- Used to study: *household assets, debt, and expenditures; wealth portfolios; policy effects.*

The LIS Databases: *LIS* and *LWS*

1983

2007

2018

LIS Database

339 datasets

LWS Database

39 datasets

LIS

Time coverage

- Historical database
- Wave I around 1980
- Wave II around 1985
- Wave III around 1990
- Wave IV around 1995
- Wave V around 2000
- Wave VI around 2004
- Wave VII around 2007
- Wave VIII around 2010
- Wave IX around 2013
- Wave X around 2016

Geographical coverage

- **Europe**: 23 EU and 6 non EU
- **America**: 2 North, 3 Central, 1 Caribbean, 7 South
- **Asia**: 4 East, 1 South , 3 West
- **Africa**: 2 countries
- **Oceania**: 1 country

LWS

Time coverage

- Wave IV around 1995
- Wave V around 2000
- Wave VI around 2004
- Wave VII around 2007
- Wave VIII around 2010
- Wave IX around 2013
- Wave X around 2016

Geographical coverage

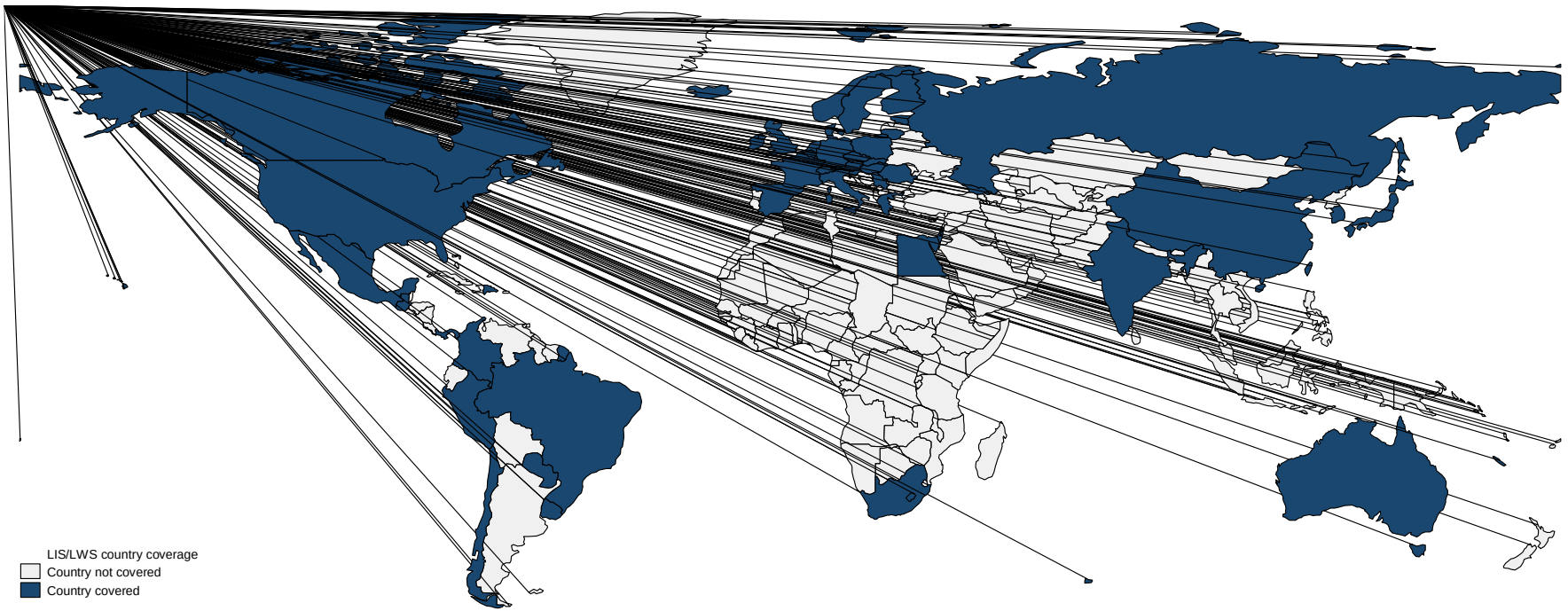
- **Europe**: 10 EU and 1 non EU
- **America**: 2 North
- **Asia**: 1 East
- **Africa**: 1 country
- **Oceania**: 1 country

Countries in LIS and LWS Databases (N=50)

approximately 65% of world population and 85% of world GDP

High-income countries (N=34):			Upper-middle-income countries (N=12):		Lower-middle-income countries (N=4):
Australia	Hungary	Slovak Republic	Brazil	Paraguay	Cote D'Ivoire (q1 '19)
Austria	Iceland	Slovenia	China	Peru	Egypt
Belgium	Ireland	South Korea	Colombia	Romania	Georgia
Canada	Israel	Spain	Dom. Republic	Russia	India
Chile	Italy	Sweden	Guatemala	Serbia	
Czech Republic	Japan	Switzerland	Mexico	South Africa	
Denmark	Lithuania	Taiwan			
Estonia	Luxembourg	United Kingdom			
Finland	Netherlands	United States			
France	Norway	Uruguay			
Germany	Panama				
Greece	Poland				

LIS country coverage by end of 2018



What we do at LIS

Core activity: **Data work**

Step 1. Data acquisition

We identify appropriate datasets (*reliable, and high-quality data)*
We negotiate with each data provider

Step 2. Data harmonisation

Common cross-national template
Comprehensive documentation

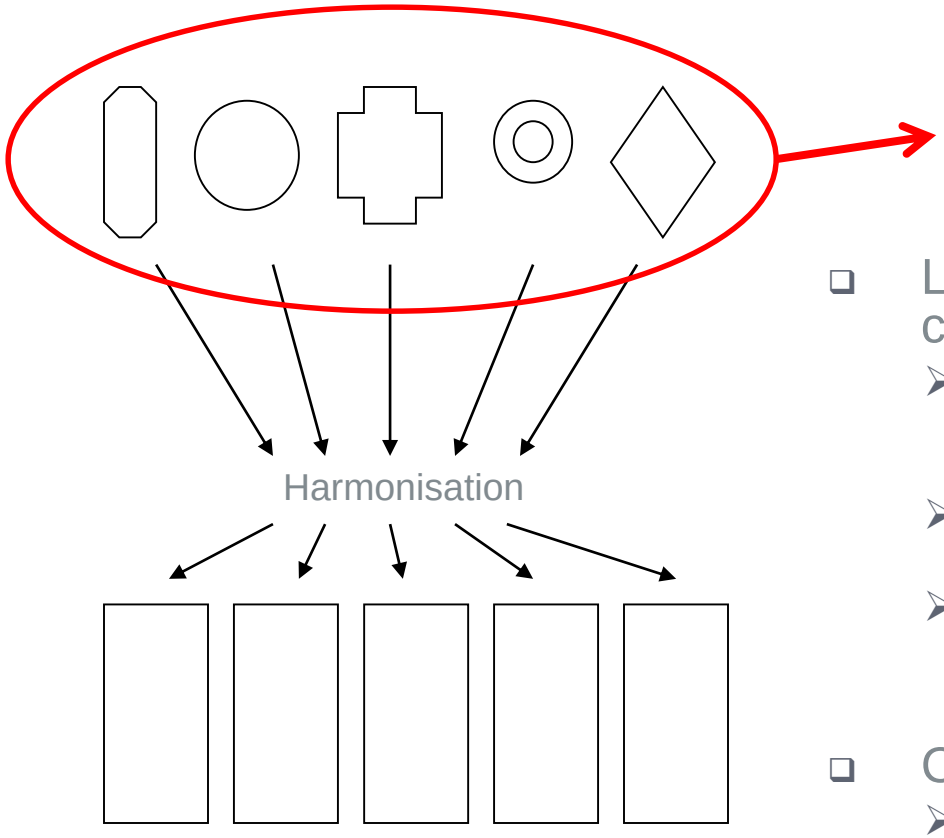
Step 3. Data dissemination

We create national-level indicators (*LIS Key Figures*)
We provide harmonized microdata to researchers via *remote execution*

Other activities

Research-promotion activities (conferences, work on methodological issues, collaborations with networks/users/journalists, newsletter, individual research)
Support (user support, research visits)

Ex-post harmonisation at LIS

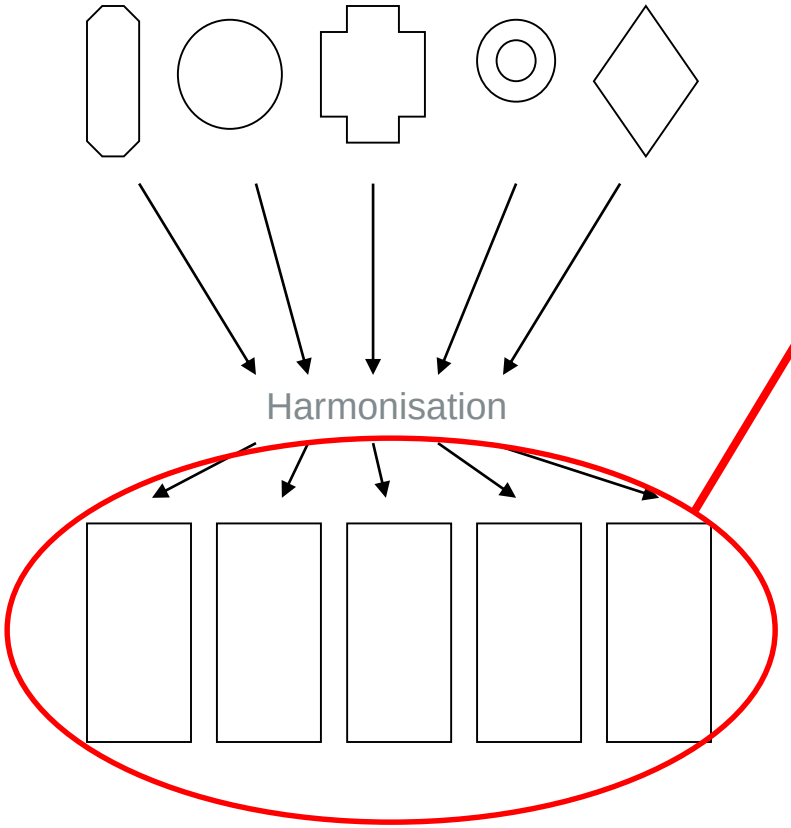


The origins of the LIS data

- LIS does not organise surveys but collects data from existing data sources:
  - *Survey data: income, household budget, living conditions, multipurpose, human development*
  - *Administrative records: tax records, employers records, social security records*
  - *Any mix of the above*

- Common denominator:
  - *microdata (household and individual level)*
  - *representative of the whole population*
  - *good quality income/wealth data*
  - *main demographic and (possibly) labour market information*

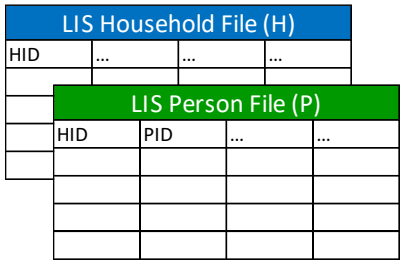
Ex-post harmonisation at LIS



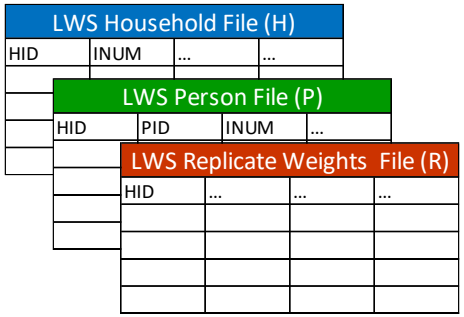
Final output:
the LIS/LWS datasets (CCYY)

- Technical harmonisation: same file structure, same variables

LIS files



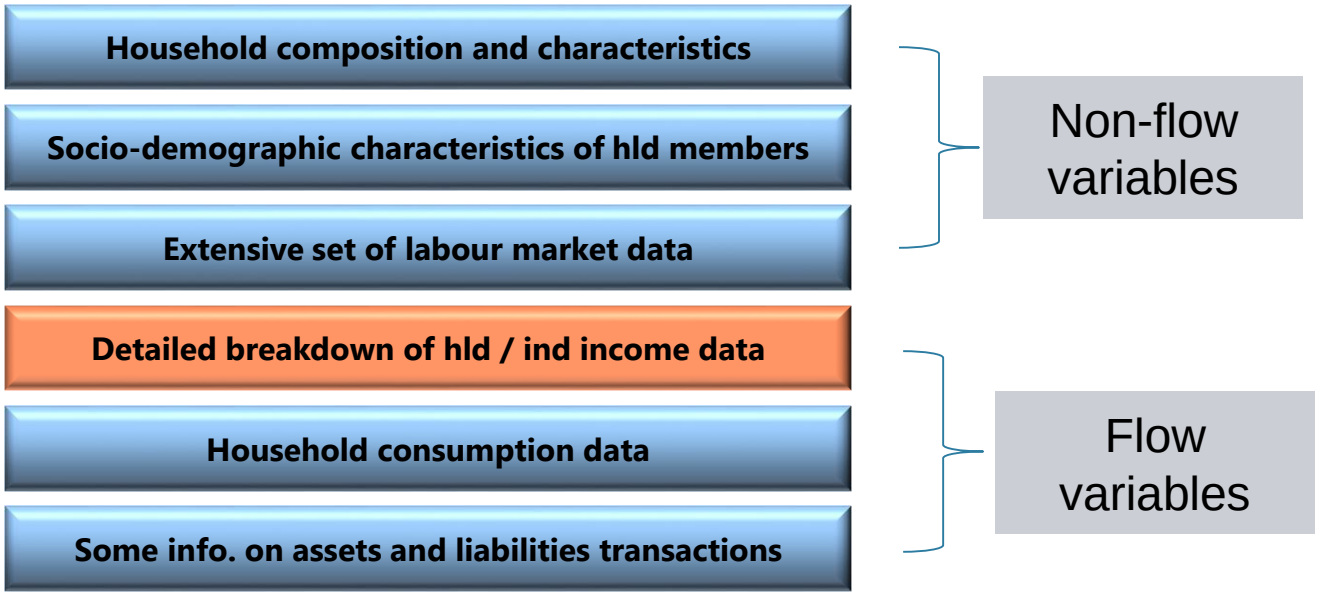
LWS files



- Conceptual harmonisation
  - Based on the same definitions
  - Comparable concepts

Harmonisation allows LIS users to eliminate many of the potential sources of technical and conceptual non-comparability

LIS Database content



Household disposable income

LIS Flow variables

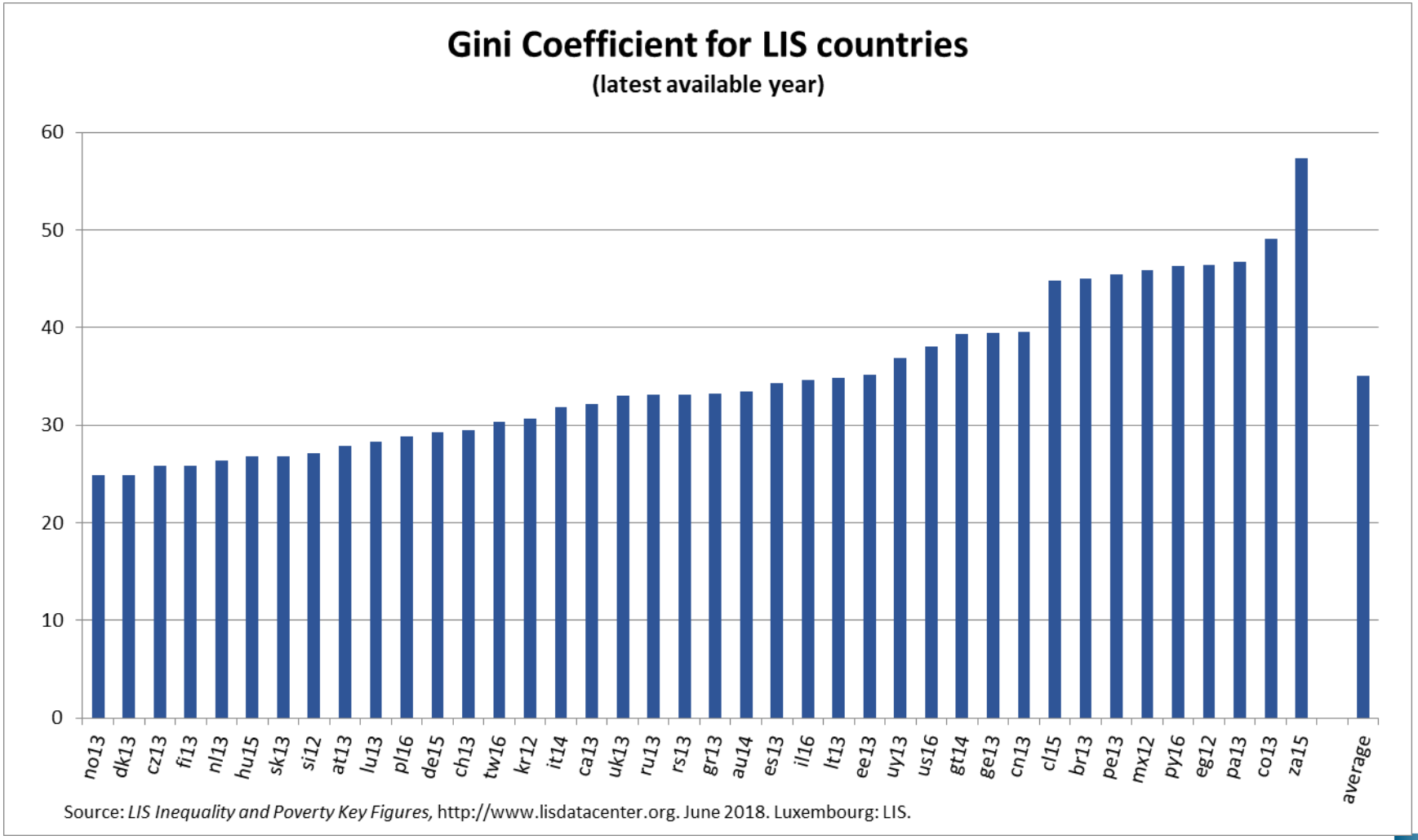
Current income	• Monetary payments as well as (the value of) goods and services received by the household or by individual members of the household at periodic intervals (annual or smaller), that are available for current consumption and that do not reduce the net worth of the household.
Windfall income	• Windfall gains and other such irregular and typically one-time receipts
Non-Consumption expenditure	• Monetary expenditures (i.e. paid directly by the household and/or its members) and nonmonetary expenditures (paid on behalf of the household and/or its members) on non-consumption goods and services (such as taxes, contributions, donations, inter-household transfers and interest paid).
Consumption	• Monetary and non-monetary consumption items.
Assets/Liabilities transactions	• Monetary inflows that do not constitute income (sales of real estate, financial products, durables or inflows from loans) and outflows that do not represent consumption (purchase of real estate financial products of outflows from loans)

LIS Household Disposable Income

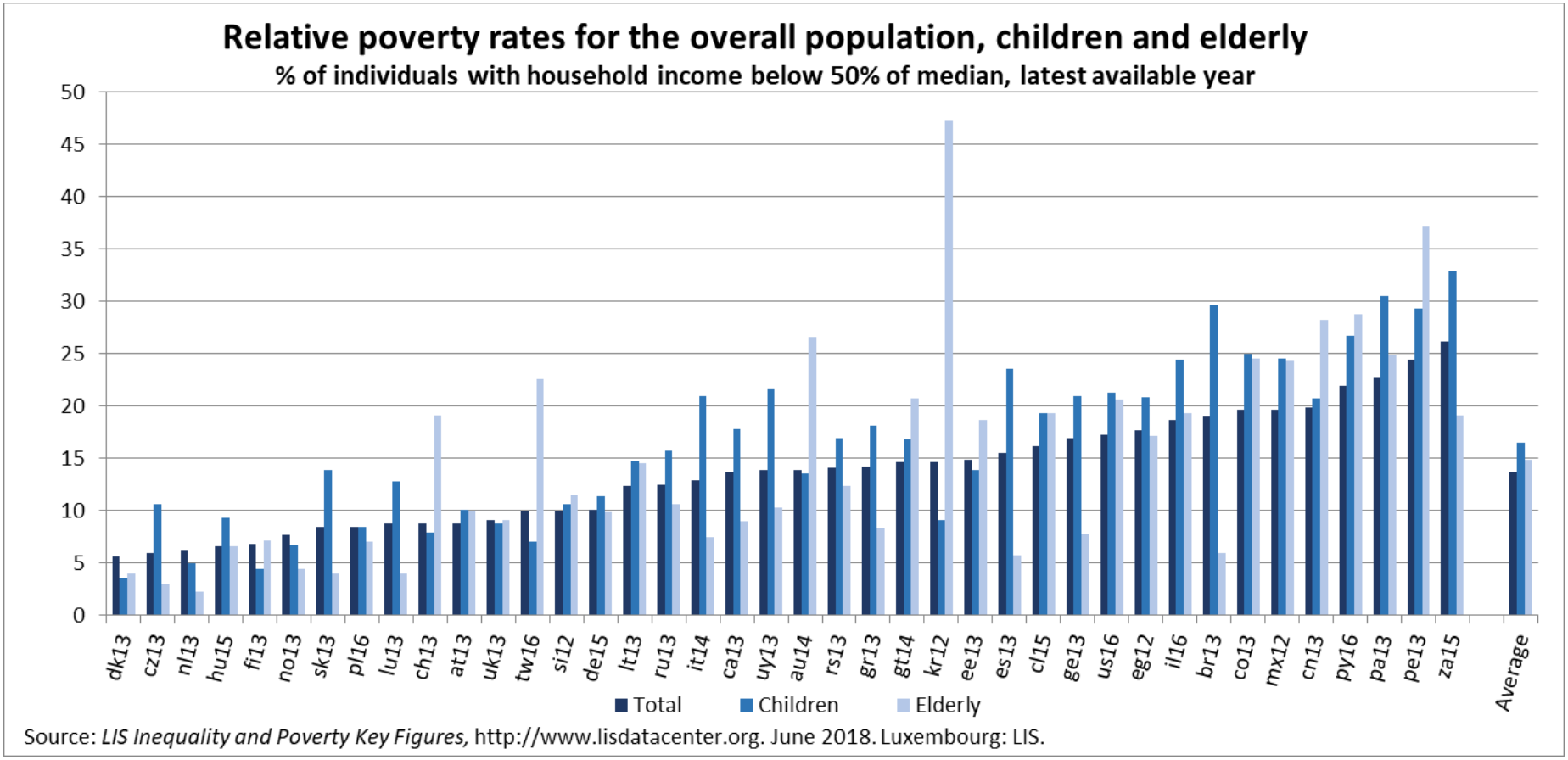
		MONETARY	NON-MONETARY
LABOUR	Dependent employment	Wages, salaries, bonuses	In-kind earnings
	Self-employement	Profits and losses	Own consumption
+ CAPITAL	Financial investment	Interest and dividends	-
	Real estate investment	Rental income	Imputed rent
+ TRANSFERS	Social security transfers		
	- *Work-related insurance*	Insurance pensions and wage-replacement benefits	-
	- *Universal transfers*	Universal pensions and universal benefits	STIK
	- *Social assitance transfers*	Minimum income guarantee	In-kind social assistance
	Private transfers	Inter-household transfers, transfers from charity	In-kind benefits from privates
= TOTAL GROSS INCOME			
- DEDUCTIONS		Income taxes Social security contributions	
= HOUSEHOLD DISPOSABLE INCOME			

Part III:
Research possibilities

Measuring Inequality Across Households

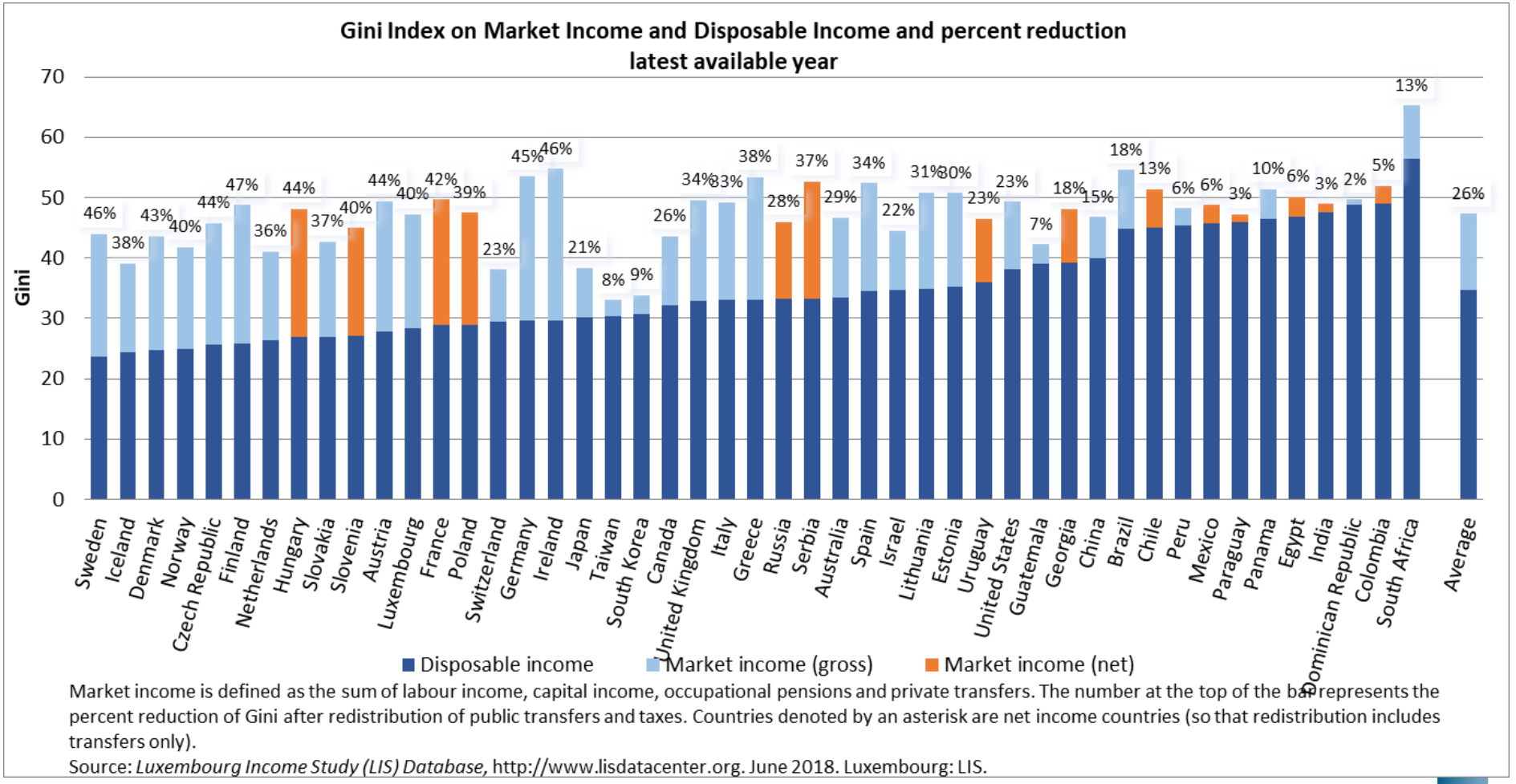


Measuring Poverty
Household Poverty Rates



Source: *LIS Inequality and Poverty Key Figures,* http://www.lisdatacenter.org. June 2018. Luxembourg: LIS.

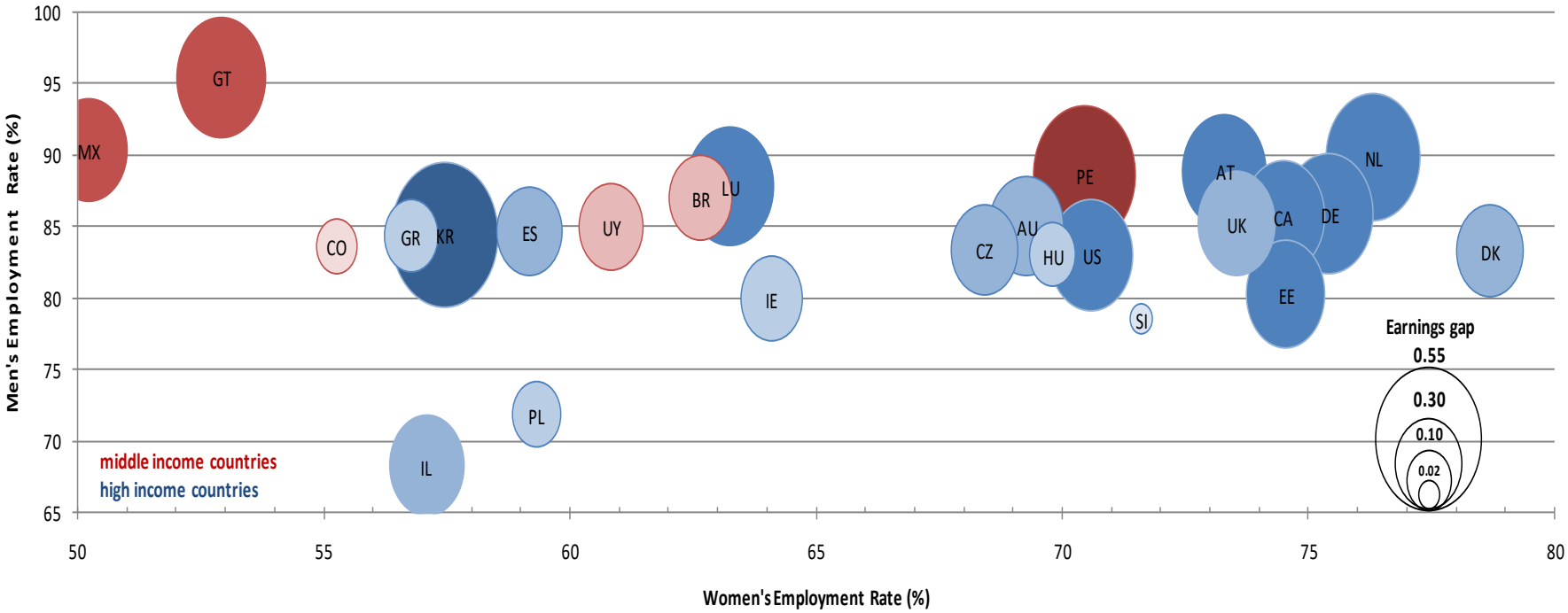
Researching Policy Impacts
Inequality reducing effect of Redistribution



Market income is defined as the sum of labour income, capital income, occupational pensions and private transfers. The number at the top of the bar represents the percent reduction of Gini after redistribution of public transfers and taxes. Countries denoted by an asterisk are net income countries (so that redistribution includes transfers only).

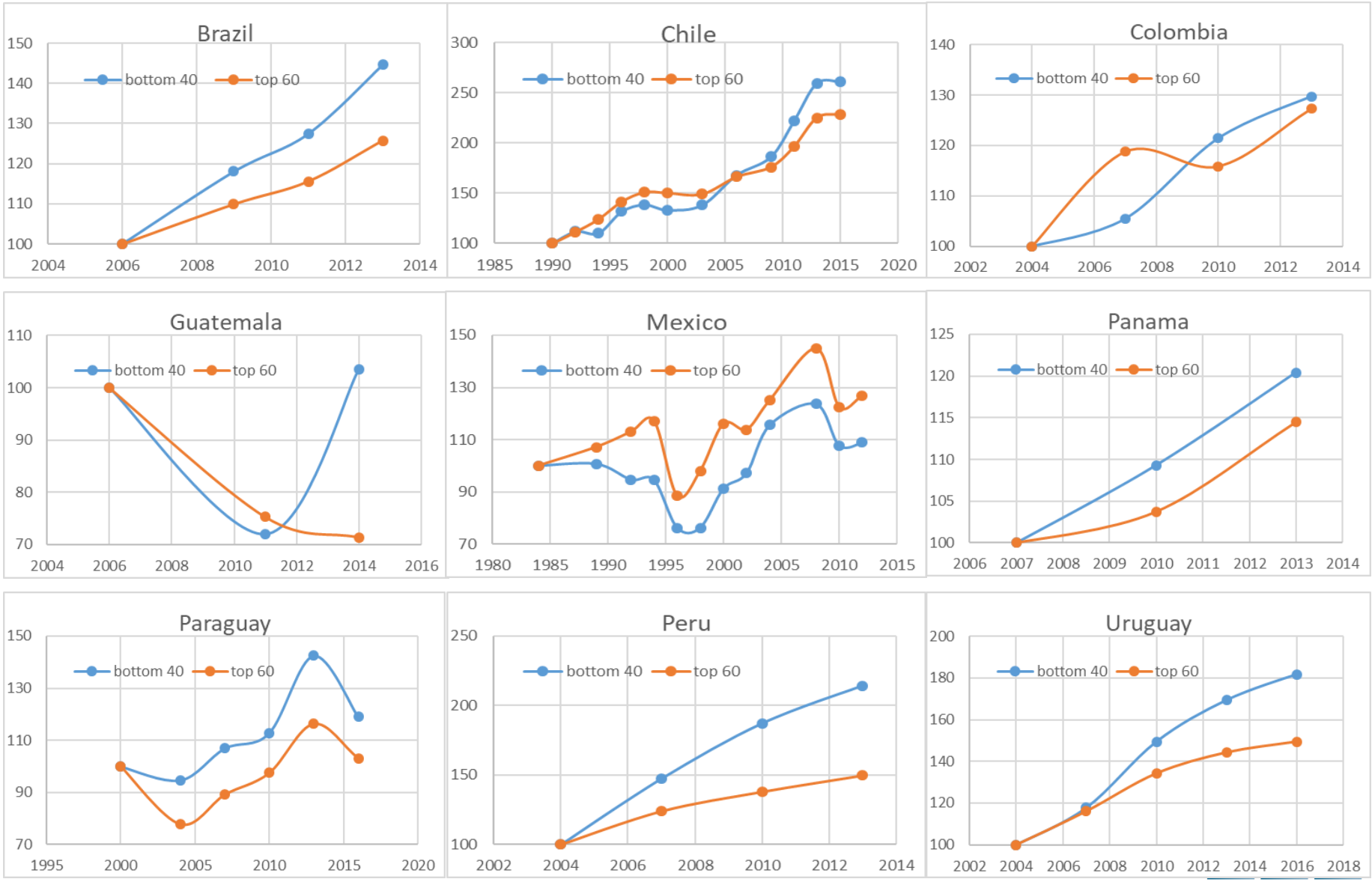
Source: *Luxembourg Income Study (LIS) Database,* http://www.lisdatacenter.org. June 2018. Luxembourg: LIS.

Comparing Employment Outcomes
Earnings Equality between Women and Men



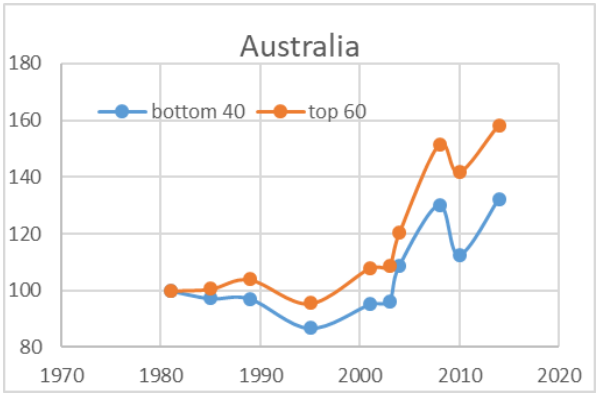
Measuring SDGs: goal 10.1

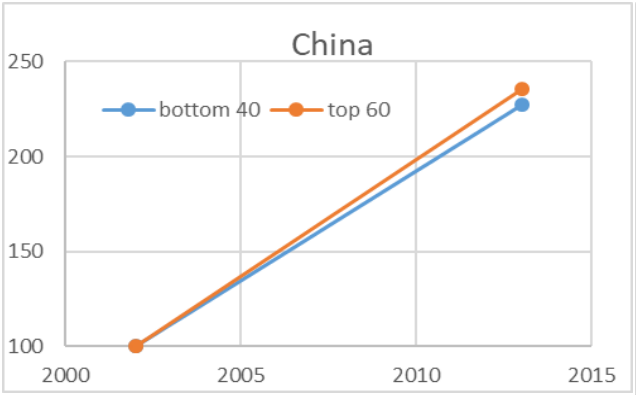
By 2030, progressively achieve and sustain income growth of the bottom 40 per cent of the population at a rate higher than the national average

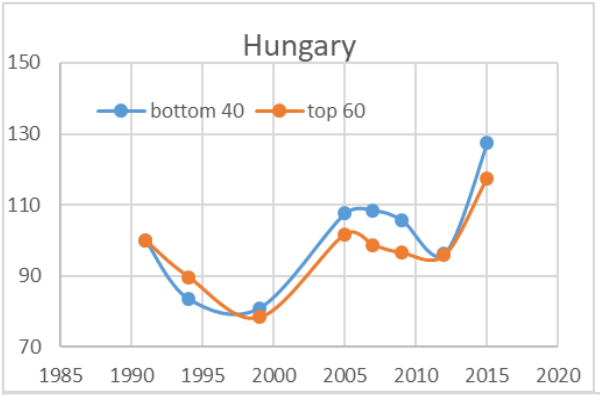


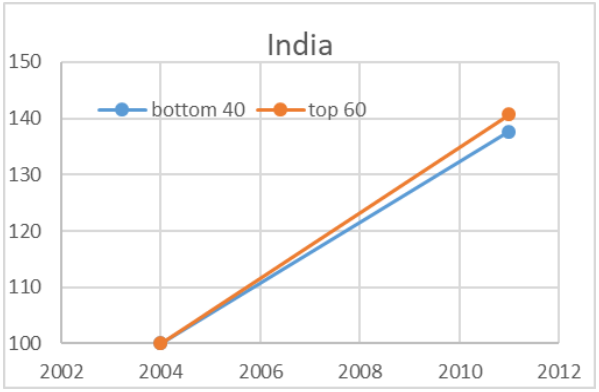
Measuring SDGs: goal 10.1

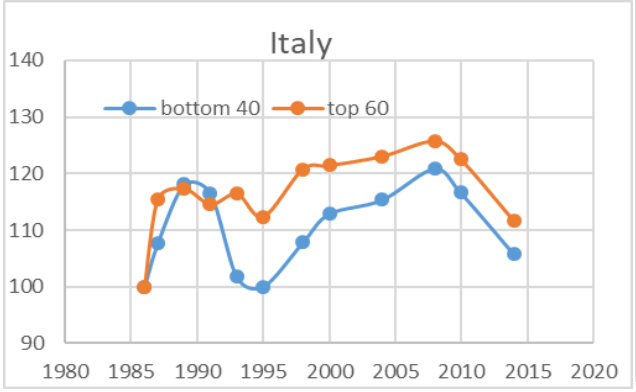
By 2030, progressively achieve and sustain income growth of the bottom 40 per cent of the population at a rate higher than the national average

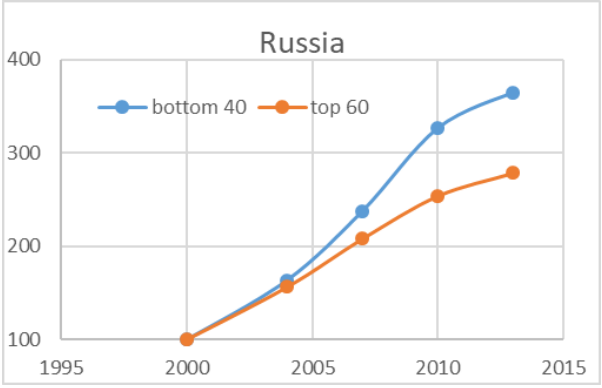


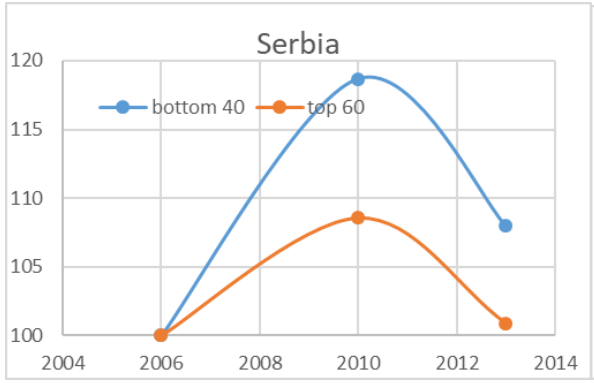


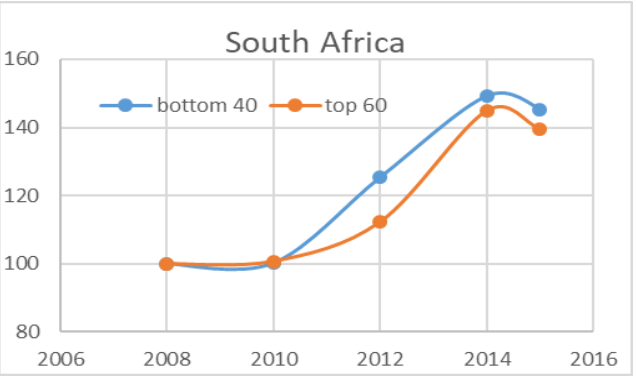


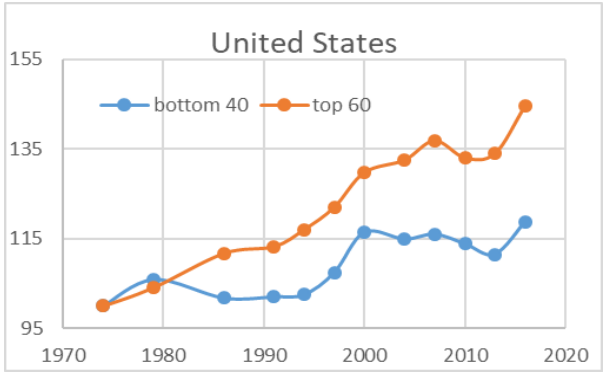






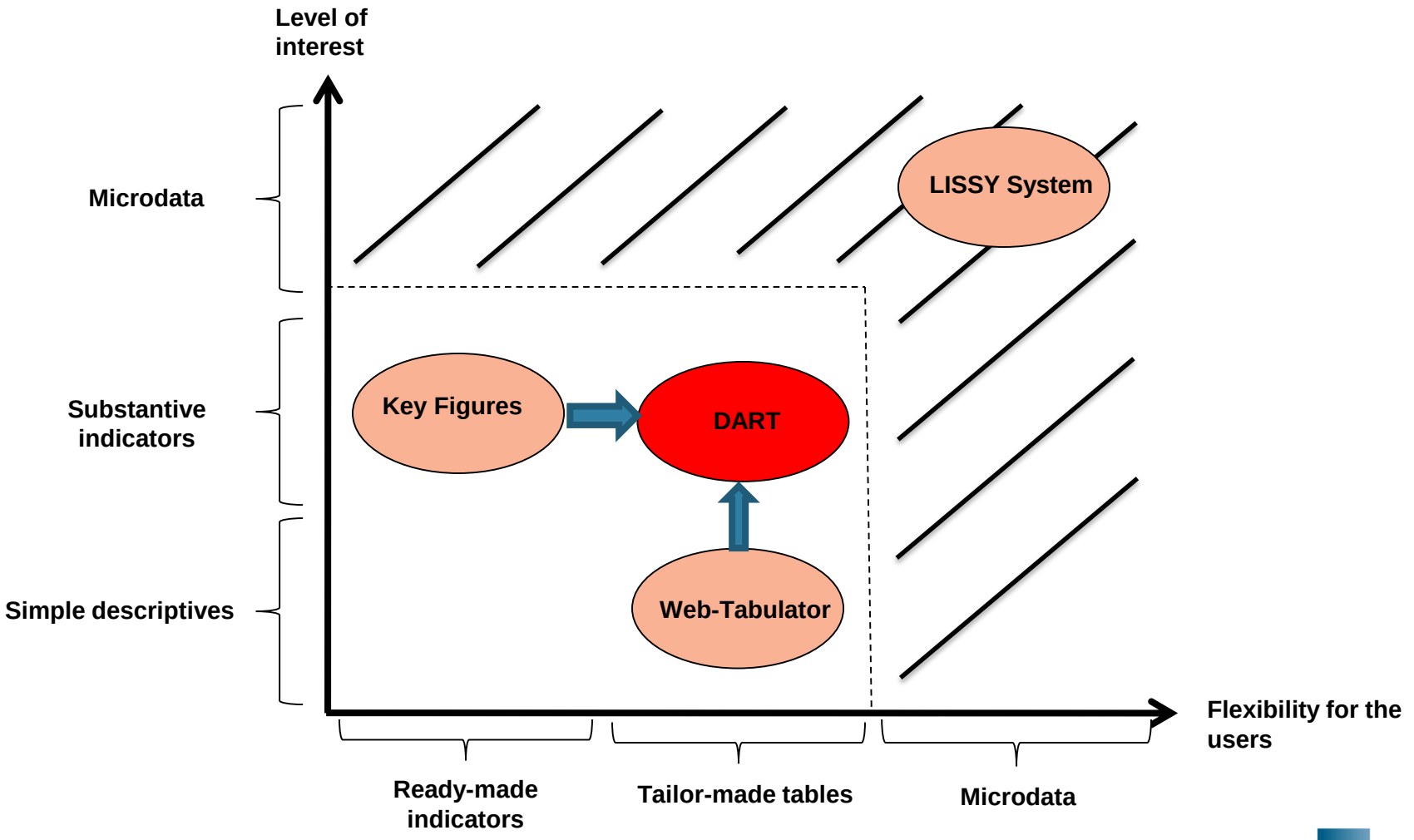






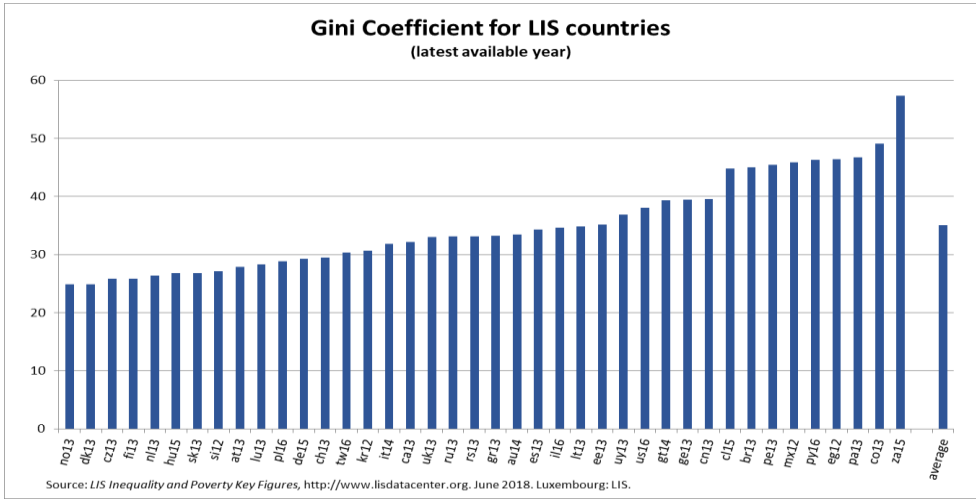
Part IV :
Data dissemination & Documentation

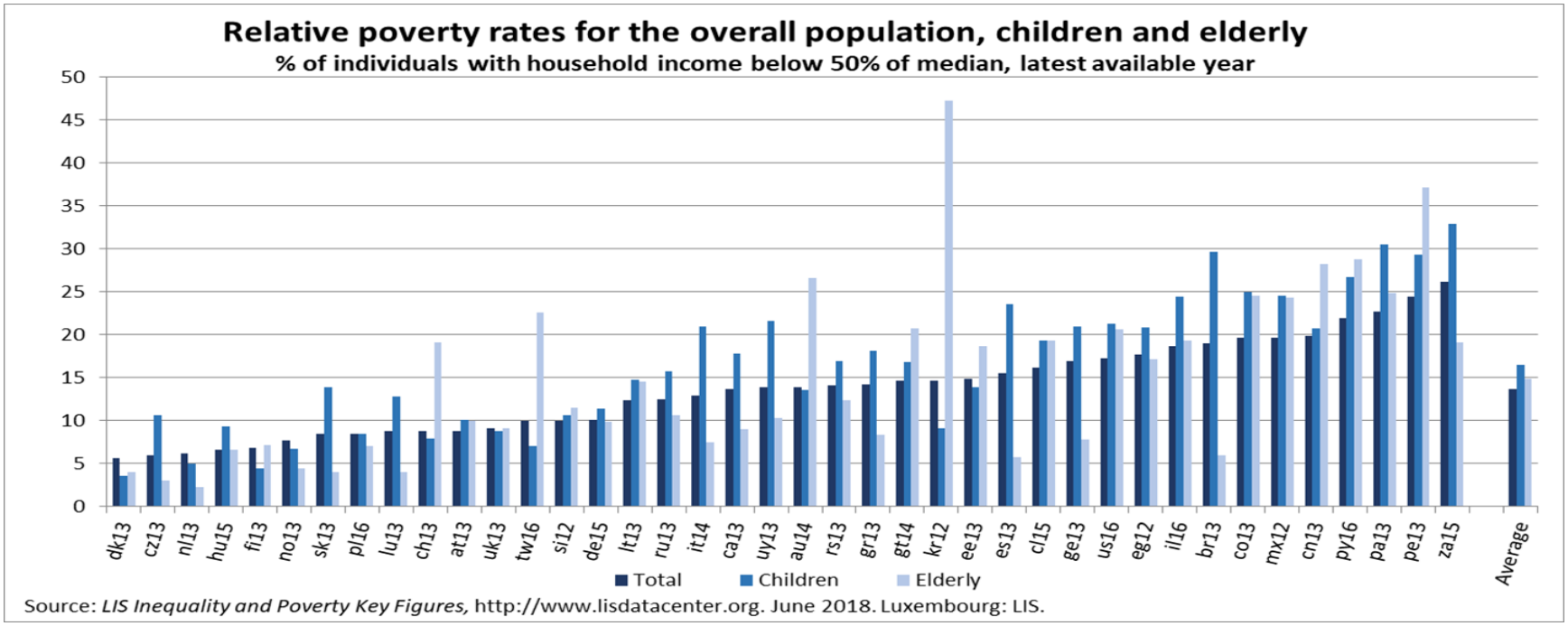
Data access


Level of interest
Microdata
Substantive indicators
Simple descriptives
LISSY System
Key Figures
DART
Web-Tabulator
Flexibility for the users
Ready-made indicators
Tailor-made tables
Microdata

LIS Key Figures

- Multiple country-level **inequality** measures (e.g., Gini and Atkinson coefficients, percentile ratios)
- Relative **poverty** rates for various demographic groups
- Median and mean disposable household income

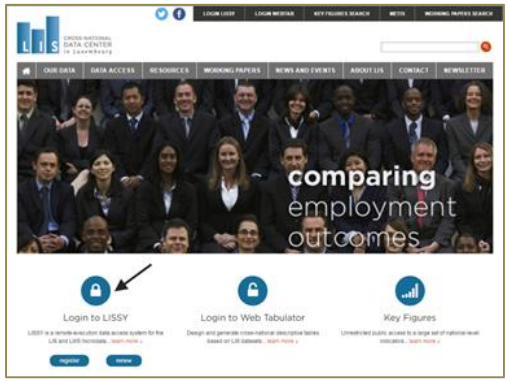




LISSY: remote-execution system

- *Fully automated, running 24 hours/day and 7 days/week*
- *Job Submission Interface (**JSI**): send statistical batch programs (SAS, SPSS, Stata or R) automatically processed and reports back aggregated results*
- *Micro-databases **cannot be downloaded** and no direct access to the data is permitted. **Only aggregated results** from statistical requests are sent back to the users.*
- *Access to LISSY is granted to researchers, incl. students, working for an academic, government or non-profit organization under the condition that use of the micro-data is restricted to **research purposes** only*

How LISSY works 1/2



LISSY WEB USERINTERFACE

submit job

job session | today jobs | job library

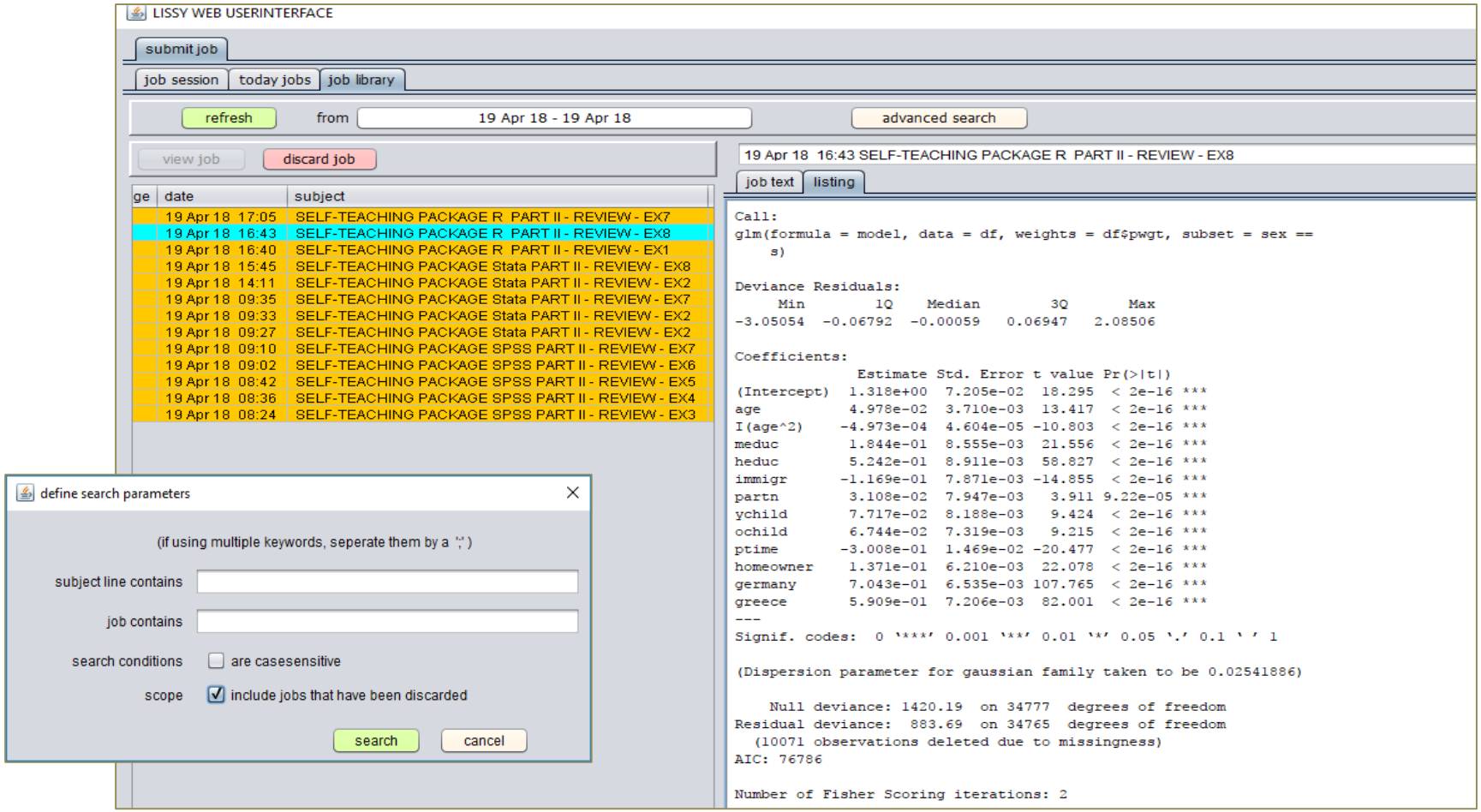
project: LIS
statistical package: R
subject: SELF-TEACHING PACKAGE R PART II - REVIEW - EX8

```
* user = kruten
* password = ****
* project  = LIS
* package  = R
```

job	project	processor	date	subject	status

```
get_stack <- function(datasets, varp, varh, subset) {
 # READ DATASETS
 pp     <- read.LIS(paste(datasets, 'p', sep = ''), labels = FALSE, vars = varp, subset = subset)
 hh     <- read.LIS(paste(datasets, 'h', sep = ''), labels = FALSE, vars = varh)
 df     <- merge(pp, hh,  by = c("dname", "hid"))
 # MAP NEW VARIABLES
 df$homeowner <- ifelse(df$own %in% 100:199, 1, ifelse(df$own %in% 200:299, 0, NA))
 df$child    <- ifelse(df$children %in% c(140, 200)   , 0, ifelse(df$children == 110 ,  1, ifelse(is.na(df$children), NA, 2)))
 df$ychild   <- ifelse(df$child == 1              , 1, ifelse(is.na(df$child)   , NA, 0))
 df$ochild   <- ifelse(df$child == 2              , 1, ifelse(is.na(df$child)   , NA, 0))
 df$partn    <- ifelse(df$partner %in% c(100, 120, 200), 0, ifelse(is.na(df$partner) , NA, 1))
 df$meduc    <- heduc <- 0
 df$meduc    <- ifelse(df$educ == 2               , 1, ifelse(is.na(df$educ)   , NA, 0))
 df$heduc    <- ifelse(df$educ == 3               , 1, ifelse(is.na(df$educ)   , NA, 0))
 df$ppp      <- ifelse(df$dname == 'de04', 0.74, ifelse(df$dname == 'gr04', 0.62, 1))
 df$hrwg     <- df$gross1
 df$hrwg     <- ifelse(df$hrwg <= 0, NA, df$hrwg)
 df$germany  <- ifelse (df$dname == 'de04', 1, 0)
 df$greece   <- ifelse (df$dname == 'gr04', 1, 0)
 for (i in 1:length(datasets)) {
  df$hrwg <- df$hrwg / df$ppp
  topline <- with(df[!is.na(df$hrwg) & df$dname==datasets[i],], 10 * wNtile(hrwg,ppopwgt,0.5))
  df$hrwg <- with(df[!is.na(df$hrwg) & df$dname==datasets[i],],
          ifelse(df$hrwg>topline,topline,df$hrwg))
 }
 return(df)
}
```

How LISSY works 2/2



METadata Information System (METIS)

METIS enables browsing, aggregating, and exporting LIS/LWS Databases documentation tailored to the users' needs

1. **Overview of datasets and variables**

Overview of the contents of the LIS/LWS databases in terms of datasets and variables

2. **Compare datasets**

Select datasets and compare generic information among them (original survey, Social Security, Key Figures)

3. **Compare variables**

Select variables and compare generic information among them (variables definitions and standard labels)

4. **Cross-compare** – main functionality

View the availability of the selected variables in the selected datasets and compare dataset-specific information (statistics and notes)

METIS by LIS CROSS-NATIONAL DATA CENTER in Luxembourg

DATABOX
8 dataset(s)
0 variables(s)
LIS

SELECTION RESULTS

Select datasets | Select variables

COUNTRY	WAVE X	WAVE IX	WAVE VIII	WAVE VII	WAVE VI	WAVE V	WAVE IV	WAVE III	WAVE II	WAVE I	HISTORICAL WAVE
Australia			2010	2008	2003	2001	1995	1989	1985	1981	
Austria		2013	2010	2007	2004	2000	1994 1995 1997		1987		
Belgium						2000	1995 1997	1988 1992	1985		
Brazil		2013	2009 2011	2006							
Canada		2013	2010	2007	2004	1998 2000	1994 1997	1991	1987	1981	1971 1975
Chile	2015	2013	2009 2011	2006	2003	1998 2000	1994 1996	1990 1992			
China		2013		2007		2002					
Colombia		2013	2010	2007	2004						
Czech Republic		2013	2010	2007	2004	2002	1996	1992			
Denmark		2013	2010	2007	2004	2000	1995	1992	1987		
Dominican Republic				2007							
Egypt		2012									
Estonia		2013	2010	2007	2004	2000					
Finland		2013	2010	2007	2004	2000	1995	1991	1987		
France			2010		2005	2000	1994	1989	1984	1978	
Georgia		2013	2010								
Germany	2015	2012 2013 2014	2009 2010 2011	2006 2007 2008	2003 2004 2005	1998 2000 2001 2002	1994 1995	1989 1991	1983 1984 1987	1978 1981	1973
Greece		2013	2010	2007	2004	2000	1995				
Guatemala		2014	2011	2006							

LIS/LWS Working papers

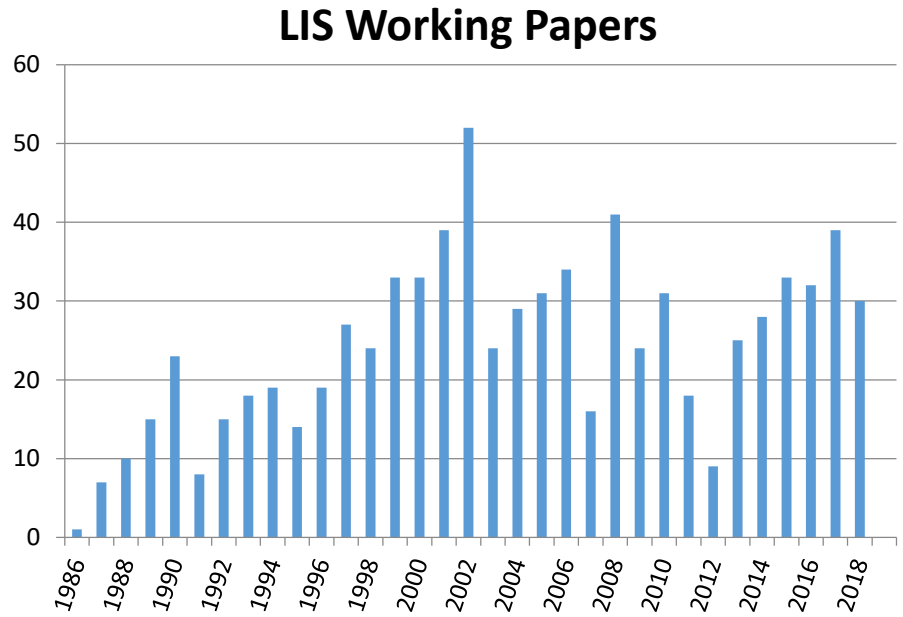
Working Paper Series

- *813 WP of which over 400 published in peer reviewed journals*
- *Aldi Award: best paper by an under 40 author*
- *included on RePec (Research Papers in Economics)*

- ***New WP:*** *69 new WP added in 2017 and 2018.*

- ***RePEC:*** *Since May 2017, LIS/ LWS working papers have been downloaded more than 9,000 times*

- ***H-index:*** *LIS = 166*
  LWS = 50



How do we respond to those challenges?

V. LIS golden rules for harmonisation

- Set clear definitions for LIS variables
  - *Maximise comparability by setting clear definitions for each variable (and trying to stick to them as much as possible)*
  - *Document very well any deviation from the general definition*
- Complement easiness of use with flexibility of use
  - *Enhance user-friendliness by providing fully standardised variables (standard variables, recodes, dummies, aggregate variables)*
  - *Allow users the flexibility to create other concepts by leaving a large amount of detailed information*
- Adapt the LIS template to the changing environment (over time and space)
  - *Template revisions*
  - *Revisions of previous datasets*

→ Overall guiding principle: **OPERATIONAL COMPARABILITY**

Final remarks

- Sample coverage
- Individual level income (high non response affecting overall household income).
- Constructing upper level aggregated variables to help in the consistency checks
- Clearer documentation and definitions