Producing regional aggregates: ILO perspectives

Inter-agency Meeting on Preparation for the 2018 SDG Reports

Steven Kapsos
Data Production and Analysis Unit
ILO Department of Statistics
February 2018

Main points

1. Garbage in, garbage out – keep input data clean
2. Identify and address non-response bias
3. Data missingness patterns should dictate methodolgy and scope of aggregation
4. Transparency is crucial
   - Methodologies should be well documented and publicly disseminated
   - Extent of imputation should be indicated to users

SDG indicators: ILO custodian or partner

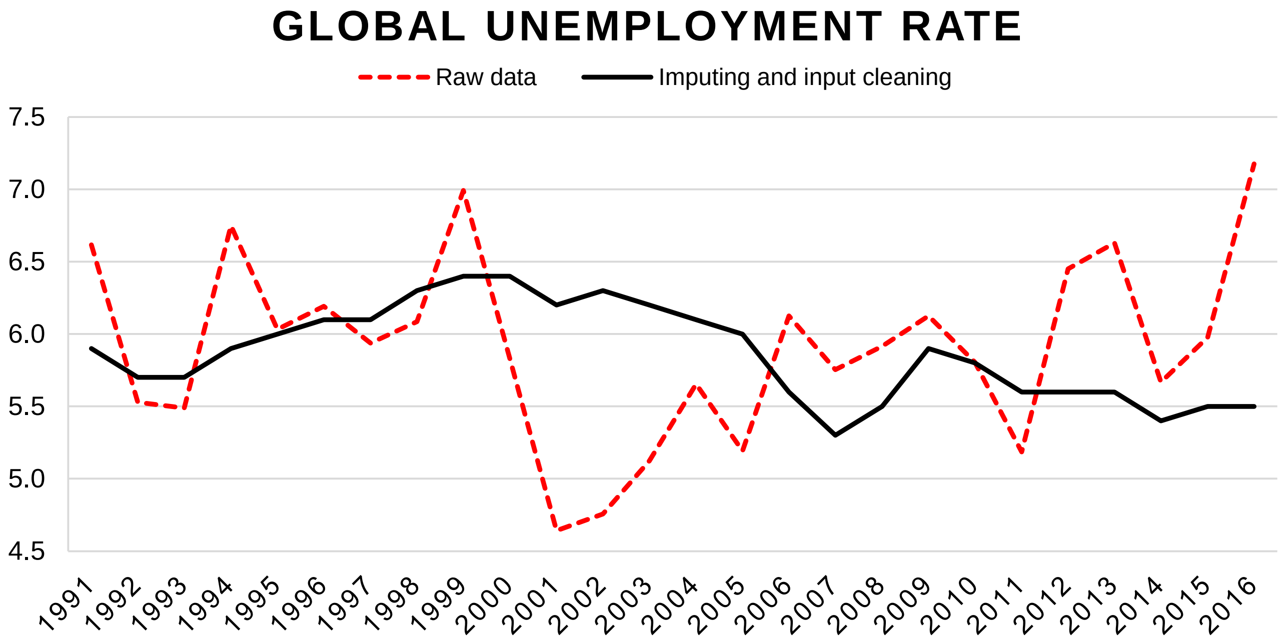
SDG Indicator (Tier I and II)	Custodian	Partner	Tier	Country data	Global and regional data
1.1.1 Proportion of population below the international poverty line, by sex, age, employment status and geographical location (urban/rural)	World Bank	ILO	I	Yes	Yes
1.3.1 Proportion of population covered by social protection floors/systems by sex, distinguishing children, unemployed persons, older persons, persons with disabilities, pregnant women, new-borns, work-injury victims and the poor and the vulnerable	ILO	World Bank	II	Yes	Yes*
5.5.2 Proportion of women in managerial positions	ILO		I	Yes	No
8.2.1 Annual growth rate of real GDP per employed person	ILO	World Bank, UNSD	I	Yes	Yes
8.3.1 Proportion of informal employment in non-agricultural employment, by sex	ILO		II	Yes	Yes*
8.5.1 Average hourly earnings of female and male employees, by occupation, age and persons with disabilities	ILO		II	Yes	No
8.5.2 Unemployment rate, by sex, age and persons with disabilities	ILO		I	Yes	Yes
8.6.1 Proportion of youth (aged 15-24 years) not in education, employment or training	ILO		I	Yes	No
8.7.1 Proportion and number of children aged 5-17 years engaged in child labour, by sex and age	ILO, UNICEF		II	Yes	Yes*
8.8.1 Frequency rates of fatal and non-fatal occupational injuries, by sex and migrant status	ILO		II	Yes	No
10.4.1 Labour share of GDP, comprising wages and social protection transfers	ILO	IMF	II	Yes	No

Garbage in, garbage out: the importance of input data cleaning

Input data cleaning and **harmonization** avoid erroneous and non-comparable data entry

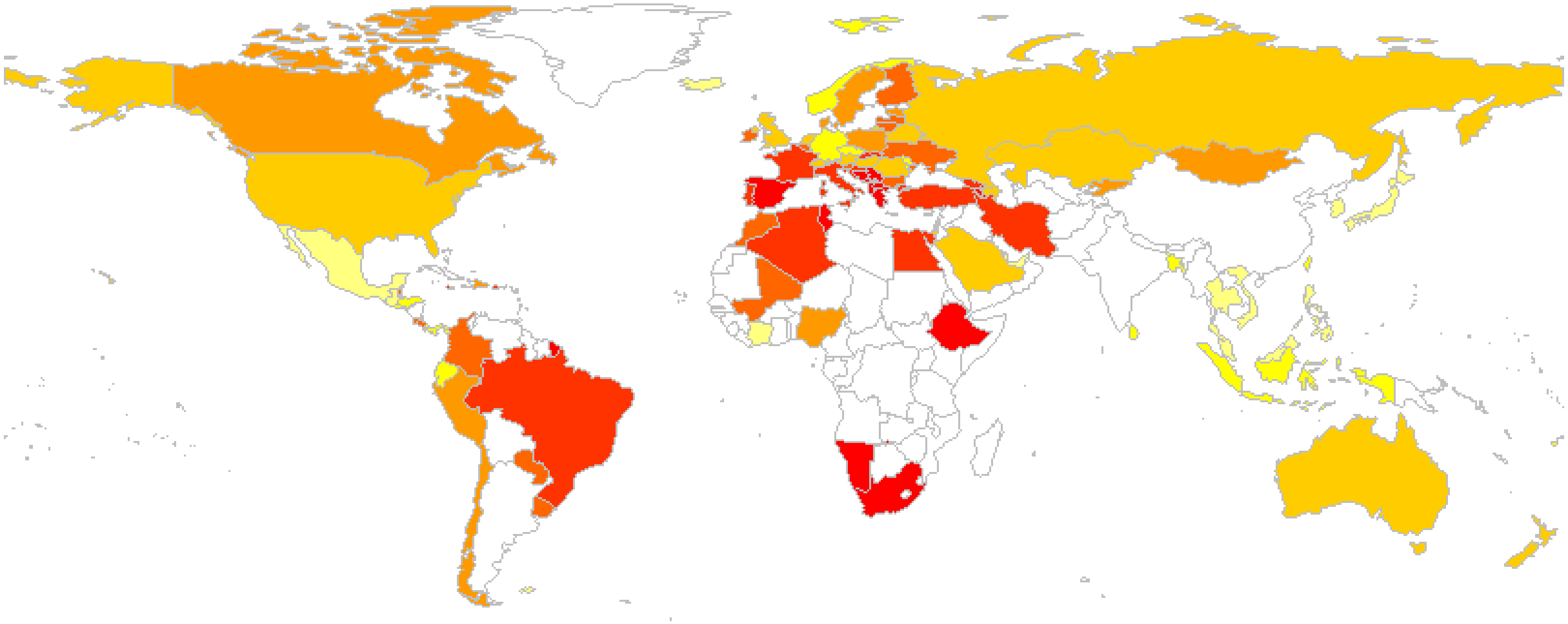
Imputation deals with non-response and differential response

Thus both reduce the **bias and volatility** in global and regional aggregates



Non-response bias can be a major issue

Unemployment rate %, 2016



Data availability/pattern of missing data should drive model selection

- High data availability
  - SDGs: 1.1.1, 8.2.1, 8.5.2
  - Strict data selection and harmonization
    - Removal or treatment of non-comparable data
    - Direct production of comparable data (ex: ILO micro data)
  - Production of balanced panel data
- Limited data availability
  - SDG: 1.3.1, 8.3.1, 8.7.1
  - More flexible data selection
  - Production of cross-section, representative time period

Recommended practice: case study on unemployment rates (I)

Input preparation

- Data processing
  - Outlier and erroneous data discarding
  - Source homogenization: labour force surveys

    Very restrictive use of household surveys, or population census
  - Coverage homogenization: nationally representative data
  - Age-group homogenization: standard age bands
- Harmonized data production (labour and data intensive)
  - Ensuring all international standards are enforced
  - Use the raw source (micro) data and process all relevant indicators

Recommended practice: case study on unemployment rates (II)

Estimation procedures

- Country-level imputation: Compute aggregates from estimated country data

  Advantages: Flexible groupings, offsetting of non-systematic errors, higher data availability

- Modelling non-response and differential response

  Estimates have to deal data not missing randomly:

  Related to degree of development, demographic factors, etc.

- Model selection, (pseudo) out of sample performance

  Rigorous choice, highest performing models

- Uncertainty management, estimating the confidence in results

Transparency

- Users should be informed about
  - The precise methodologies used to produce global and regional aggregates
  - The extent of imputation globally and by region (% of countries reporting data, % of population covered)
  - If model design allows, a confidence interval for the estimates
- Ideally this information would be available not only from agencies, but also through the SDG global database & portal

Thank you