

## **Filling Data Gaps in the Measurement of Income Inequality. A Complete Dataset of National Gini Coefficients 1995-2019**

Ivo Mossig, Hannes Lehmann (University of Bremen)

Income inequalities are a major societal challenge and research on their causes, consequences, and policy solutions has become central to social science (Grusky 2018; Polacko 2021). Despite criticism, the Gini coefficient – especially income-based Ginis – remain the most important indicators for measuring the extent and development of income inequality within a country. Common sources include the OECD, World Bank, Luxembourg Income Study (LIS), and World Income Inequality Database (WIID). Unfortunately, Gini coefficients based on comparable methodologies are only available to a very limited extent. To remedy this lack of data, we create a complete dataset of Gini coefficients from 1995 to 2019 for 160 countries through the combination of machine learning methods and a large dataset of social-policy indicators.

From a geographical perspective, data gaps are especially prevalent for countries in the Global South. Furthermore, pre-2005 data are often incomplete or missing even for the Global North. The most comprehensive data set available with consistent definitions for net income is the WIID Gini. With around 900 data points, this data set covers only 22% of the possible country-year combinations for the selected sample of 160 countries in the chosen observation period.

Therefore, we pursue two objectives: (1) to close existing data gaps through statistical imputation thereby creating a consistent and plausible dataset of Gini coefficients for 160 countries with over 1 Mio. inhabitants from 1995 to 2019 and (2) to identify the socioeconomic and political indicators that most strongly influence these imputations. To achieve this, missing data are estimated using a gradient boosting machine (GBM) drawing on over 1400 socioeconomic and political indicators from the WeSIS database ([www.wesis.org](http://www.wesis.org); Mossig and Obinger 2023). The results are contextualized on the basis of an extensive literature review on the causes of income inequality.

According to our methodology, the following indicators are particularly suitable for predicting income inequality: (a) Gender Inequality Index, (b) Proportion of people living below 50% of median income, (c) Total population ages 65 and above, (d) Codetermination and information/consultation of workers, (e) De facto coverage of children by the child benefit for

citizens/residents as well as (f) Capital city coordinates (Latitude). With this novel dataset, we enable researchers to broaden their inquiry into causes and effects of socio-economic inequality on a formerly unachievable scale.

JEL Codes:

C55 Large Data Sets: Modeling and Analysis

D31 Personal Income, Wealth, and Their Distributions