Winter School on Inequality and Social Welfare Theory – IT19

Alba di Canazei, 7-9 January 2026.

Using the Forbes Billionaires List as a Research Dataset: Methods, Challenges, and Insights on the Global Super-Rich

Lidia Ceriani

Department of Economics, University of Verona

Email: lidia.ceriani@univr.it

Web: https://www.dse.univr.it/?ent=persona&id=97660

Description

Research on economic inequality increasingly relies on unconventional data sources to illuminate the dynamics of wealth concentration at the very top of the distribution. Among these, the Forbes list of billionaires has become a prominent input for empirical analysis, despite being neither administratively generated nor statistically harmonized. This laboratory-based lecture introduces young scholars to the methodological challenges and opportunities involved in working with such data, combining hands-on data cleaning techniques with the responsible use of large language models (LLMs) as research assistants. Using Stata, participants learn how to diagnose and correct common data quality issues in the Forbes dataset, with particular emphasis on string manipulation functions for cleaning names, harmonizing country identifiers, and parsing unstructured textual information on sources of wealth.

The session then turns to the systematic problem of missing demographic information—most notably gender and age—which limits distributional and intersectional analyses of extreme wealth. Rather than treating these gaps as purely technical nuisances, the laboratory frames them as analytically meaningful and introduces transparent workflows for augmenting the data using ChatGPT.

Participants are shown how to design reproducible prompts, request sources and uncertainty assessments, and integrate AI-generated information back into Stata while preserving auditability through explicit flags and documentation. The lab concludes with a discussion of validation strategies, ethical considerations, and best practices for disclosure in inequality research. Overall, the session aims to equip early-career researchers with practical skills at the intersection of data cleaning, extreme-wealth analysis, and critical AI literacy, highlighting how new tools can enhance empirical work without obscuring its limitations.

Prerequisites

- Basic familiarity with Stata (use, gen/replace, merge)
- No prior experience with LLMs required