"More Reliable Labor Market Analysis through Quality Checks: An Application to the Global Jobs Indicators Database (JoIn)"

Authors: Elena Casanovas, Jörg Langbein, and Michael Weber

Extended Abstract

In recent years, there has been a substantial increase in the availability of large datasets for labor market analysis in developed economies. While a more and more digitized world provides an abundance of data, researchers in developing countries are usually confined to either indicators from household surveys from national statistics offices (secondary data) or self-collected, primary microdata. The secondary data can give representative information on a national level but may suffer from bad data quality. The primary data, in contrast, is often collected from smaller samples and fails to provide representative information on labor market developments. For answering broader question, relying only on self-collected data is, hence, typically not feasible.

In this paper we present a methodology for checking the quality of (secondary) labor market data provided by the World Bank's Global Jobs Indicators database (JoIn) with a focus on developing countries. The jobs indicators are derived from the World Bank's International Income Database (I2D2) which covers approximately 120 million observations in 1,250 surveys for 130 countries in the world. For all the surveys 74 Job related indicators were generated. The indicators are built from both household and labor surveys from the I2D2 database and its variable definitions. They can be subsumed into the following topics: socio-demographics, labor force and employment status, employment by sector and occupation, labor market outcomes, including earnings. The indicators are disaggregated by urban vs. rural, gender, young workers vs old workers and low vs high educated workers dimensions. In order to ensure the reliability of the data, we have applied a series of automatized quality checks to both the indicators and the micro data, at the cross-sectional survey level and at the survey time-series level. Results are, for example, corroborated using statistics provided by the ILO or the World Bank's World Development Indicators as well as newly developed techniques. In total, we run 115 quality checks, each of whom may result in a flag indicating a potential issue in the data. The checks are applied both to individual surveys and to series of surveys. In the individual surveys, we test for differences with other data sources, internal coherence of the indicators, realistic values of the data and missing values in the micro-data. At the survey time-series level, we test for different types of outliers. In order to raise a flag, we compare the result of the test to the distribution of results in our database and mark the results in four different levels. The outcomes of the secondary data quality checks process are explicitly reported in the Global Jobs Indicators (JoIn) database and can be publicly accessed.

The results are relevant for policymakers and researchers on three levels: First, by deriving a novel methodology on data quality testing for a huge set of observations, surveys and countries. The approach allows to distinguish between different sources of error and provide both indication on how to treat the resulting data as well as useful feedback for the design and implementation of the surveys. Second, making the so-checked results publicly available allows other researcher and policy makers immediately to identify the flaws in the datasets. Third, the validated job indicators can provide more reliable insights into the changing nature of jobs and structural transformation in the respective (developing) countries. This complements information provided by international organizations such as the ILO with additional new indicators. Finally, the checks represent a novel methodology on data

quality testing that can potentially become a reference for labor market analysis of household survey data.