**The Subnational Corruption Database (SCD)**

The Subnational Corruption Database (SCD) provides the levels of total, grand, and petty corruption for 1,473 subnational regions within 178 countries.

The SCD also offers subnational versions of existing corruption indices.

The SCD contains three different data files: “Baseline Dataset SCI”, “Comprehensive Dataset SCI”, and “SUBCPI and SUBCCI Dataset”. Each of these files is available in three different versions; for SPSS, STATA and MS Excel. The dataset also contains two Stata syntax files: “From Raw Dataset to Baseline Dataset” and “From Baseline Dataset to Comprehensive Datasets”. Below, we also indicate all variable names in closed brackets.

The *Baseline Dataset SCI* file contains the subnational and national values of the Subnational Corruption Index (SCI) [SCI] and its two components: the Subnational Petty Corruption Index (SPCI) [petty] and the Subnational Grand Corruption Index (SGCI) [grand]. All indices range from 0-100, where 0 is the largest level of corruption and 100 is the lowest level of corruption. These indices are most suitable for explanatory analyses. Further, the dataset contains predicted values for the SCI [SCI2], SPCI [petty2] and SGCI [grand2], which fill up missing values in years where either the SPCI or the SGCI is missing by exploiting the relationship between the SPCI and the SGCI. The *Baseline Dataset SCI* contains 6,701 area-years of corruption data based on 804 standardized and harmonized household surveys.

The *Comprehensive Dataset SCI* file contains the subnational and national values of the estimated SCI after applying methods of inter- and extrapolation. This dataset is most suitable for descriptive analyses such as within-year comparisons. The number of subnational regions varies between countries and can reach up to 37 regions for Nigeria, with an average of 8.3 regions. The years for which the SCI is available are similar to those of the World Bank’s Control of Corruption (CCI) (from 1995-2022) The total number of country-years for which SCI data is available is 4,930.

The *SUBCPI and SUBCCI Dataset* also contains subnational and national values for all countries and years (from 1995-2022). In this dataset, we attached the subnational variation of our SCI around two existing measures of corruption: the World Bank Control of Corruption Index (CCI) (1995-2022) [SUBCCI] and Transparency International’s Corruption Perceptions Index (CPI) (2012-2022) [SUBCPI].

The *From Raw Dataset to Baseline Dataset* Stata do-file shows the syntax steps made to get from the raw dataset to our *Baseline Dataset SCI*. The raw dataset contains the underlying questions of all sources at the individual level, after combining all subnational region definitions under one coding scheme. The raw dataset is not available given copyright. The *From Baseline Dataset to Comprehensive Datasets* Stata do-file shows the syntax required to move from our *Baseline Dataset SCI* to our *Comprehensive Dataset SCI* and *SUBCPI and SUBCCI Dataset*.

A detailed description of the SCD can be found in the data descriptor paper “The Subnational Corruption Database (SCD): Total, Petty and Grand Corruption in 1,473 Regions in 178 Countries Between 1995-2022”.

The most recent version of the SCD is available at [www.corruptionradar.com](http://www.corruptionradar.com). All data files contain region classifications that coincide with those of the Global Data Lab (www.globaldatalab.org), an independent and data research center with the largest subnational databases for LMICs on socioeconomic data.