

Globalization and Taxation Dataset

Pierre Bachas Matthew Fisher-Post Anders Jensen Gabriel Zucman

This README describes the datasets that are available for download at the website <https://globaltaxation.world/>. We provide an overview of the main sources as well as a list of the variables contained in the dataset. More details can also be found in Online Appendices A and B of the study *Globalization and Factor Income Taxation* (downloadable on the website).

Sources

We construct a new tax revenue dataset that includes disaggregated tax revenue data by type of tax. Our database includes all taxes---personal income taxes, corporate income taxes, Social Security payroll taxes, property taxes, wealth taxes, estate and inheritance taxes, consumption, and other indirect taxes ---at all levels of government. Our tax revenue dataset derives from four main sources: (i) the OECD Government Revenue Statistics; (ii) the ICTD/UNU-WIDER Government Revenue dataset; (iii) the IMF Government Finance Statistics; and, (iv) digitized archival records of thousands of government documents.

When available, OECD tax revenue data is our preferred source, because it covers and classifies all types of tax revenues, usually back to 1965 for OECD countries. OECD data accounts for 41% of the country-year observations in our tax dataset. Its drawback is its limited coverage of non-OECD countries (it covers 93 countries in total) and only over the most recent two decades. To increase coverage, we augment the OECD data with the tax revenue data from ICTD/UNU-WIDER (17% of observations). This dataset achieves near worldwide coverage but, for our purposes, faces limitations: it only starts in the 1980s; it does not follow the tax classification of the OECD; it sometimes mixes personal and corporate income taxes; and it often lacks payroll taxes and decentralized taxes. To address these shortcomings, we use historical public finance data from government reports, primarily from the Harvard Library archives (30% of our sample) and from IMF GFS offline historical database (12% of observations). For each country-year, the variable *'source_tax'* (see list below) records the data-source that is being used.

To estimate the macro-economic effective tax burdens on labor and capital (the variables *'ETR_K'* and *'ETR_L'*), we combine this tax revenue database with new measures of labor and capital factor shares in national income. These measures are derived by combining data from the UN System of National Accounts, the World Inequality Database, and other sources. From the World Inequality Database, we retrieve United Nations (UN) System of National Accounts (SNA) data that covers approximately 4,000 country-years. These data come from the production and income accounts of the online SNA's ``Main Aggregates and Detailed Tables.'' In addition, the UN

Statistics Division provided us access to their archival data on the components of GDP, with over 2,000 country-year observations from the 1960s and 1970s, presented following the 1968 System of National Accounts. To ensure comparability with the more recent data, we recast the historical series into the 2008 System of National Accounts framework. To our knowledge, this is the first factor income shares dataset that harmonizes data from the 2008 and 1968 System of National Accounts. In countries and years when the two systems overlap (typically in the 1970s, when countries transitioned from the old to the new framework), the series match well.

List of variables

The variables contained in the downloadable datasets are the following:

variable	definition
<i>country_name</i>	Country name as specified by the United Nations
<i>country</i>	ISO code
<i>year</i>	year
<i>region</i>	OECD or continent (excl. OECD members)
<i>wb_inc</i>	modified WB income class (LIC & LMIC = 'developing')
<i>ndp_usd</i>	net domestic product (constant 2019 USD)
<i>ETR_L</i>	effective tax rate on labor [excl. NFI and NIT, noting PIT thresholds & dual PIT systems]
<i>ETR_K</i>	effective tax rate on capital [excl. NFI and NIT, noting PIT thresholds & dual PIT systems]
<i>Tau_L</i>	tax revenue (% of NDP at factor prices), taxes on labor [noting PIT thresholds & dual PIT systems]
<i>Tau_K</i>	tax revenue (% of NDP at factor prices), taxes on capital [noting PIT thresholds & dual PIT systems]
<i>Lsh_ndp</i>	labor share of (factor-price) net domestic product
<i>Ksh_ndp</i>	capital share of (factor-price) net domestic product
<i>pct_tax</i>	total tax revenue, as % of NDP
<i>pct_1000</i>	total income tax (1000 series) revenue, as % of NDP
<i>pct_1100</i>	PIT (1100 series) revenue, as % of NDP
<i>pct_1200</i>	CIT (1200 series) revenue, as % of NDP
<i>pct_2000</i>	social contributions (2000 series) revenue, as % of NDP
<i>pct_4000</i>	property and wealth tax (4000 series) revenue, as % of NDP
<i>pct_5000</i>	indirect tax (5000 series) revenue, as % of NDP
<i>pct_6000</i>	other tax (6000 series) revenue, as % of NDP
<i>source_tax</i>	tax revenue data source (HA = historical archive, local, and scholarly sources)
<i>source_sna</i>	source of SNA components (SNA1968 UNSD archive; or WID via SNA online, etc.)