

# Effective Tax Rates, Firm Size and the Global Minimum Tax

Code Repository Documentation

Pierre Bachas\*  
Anne Brockmeyer†  
Roel Dom‡  
Camille Semelet§

March 2025

---

\*Development Research Group, World Bank & EU Tax Observatory. Email: [pbachas@worldbank.org](mailto:pbachas@worldbank.org)

†IFS, UCL, World Bank and CEPR. Email: [abrockmeyer@worldbank.org](mailto:abrockmeyer@worldbank.org)

‡University of Antwerp and World Bank. Email: [roel.dom@uantwerpen.be](mailto:roel.dom@uantwerpen.be)

§University of Munich, ifo Institute and World Bank. Email: [semelet@ifo.de](mailto:semelet@ifo.de)

# Contents

<b>1</b>	<b>Overview</b>	<b>3</b>
<b>2</b>	<b>Data Availability and Provenance Statements</b>	<b>3</b>
2.1	Statement about Rights . . . . .	3
2.2	Summary of Availability . . . . .	3
2.3	Details on each Data Source . . . . .	3
<b>3</b>	<b>Datasets</b>	<b>6</b>
3.1	Input Datasets . . . . .	6
3.2	Datasets for Analysis . . . . .	9
<b>4</b>	<b>Computational Requirements</b>	<b>10</b>
<b>5</b>	<b>Instructions to replicators</b>	<b>11</b>
5.1	Folder structure . . . . .	11
<b>6</b>	<b>List of tables, figures and programs</b>	<b>12</b>
6.1	Mapping of tables and analysis code . . . . .	12
6.2	Mapping of figures and analysis code . . . . .	12
<b>7</b>	<b>Description of programs and code</b>	<b>14</b>
7.1	Creation of data sets for analysis . . . . .	14
7.2	Analysis . . . . .	19
7.3	Programs . . . . .	28
	<b>Bibliography</b>	<b>29</b>

# 1 Overview

The codes in this repository replicate the tables and figures from “Effective Tax Rates, Firm Size, and the Global Minimum Tax”, by Bachas, Brockmeyer, Dom, and Semelet. The replication folder contains the codes to go from the raw administrative data to the results in the paper.

This documentation is structured as follows. Section 2 describes the data sources and their availability. Section 3 describes the datasets used in the analysis. Section 4 provides details on the computational requirements. Section 5 provides instructions to replicators. Section 6 provides a mapping between the codes and the tables and figures of the paper. Finally, section 7 describes the codes, including data creation codes (section 7.1), analysis codes (section 7.2), and programs (section 7.3).

## 2 Data Availability and Provenance Statements

- ☐ This paper does not involve analysis of external data (i.e., no data are used or the only data are generated by the authors via simulation in their code).

### 2.1 Statement about Rights

- ☒ I certify that the author(s) of the manuscript have legitimate access to and permission to use the data used in this manuscript.
- ☐ I certify that the author(s) of the manuscript have documented permission to redistribute/publish the data contained within this replication package.

### 2.2 Summary of Availability

- ☐ All data **are** publicly available.
- ☐ Some data **cannot be made** publicly available.
- ☒ **No data can be made** publicly available.

The administrative data for this project is owned by different government tax authorities. The data were made available to us exclusively for the purpose of this research project through collaboration agreements between the corresponding author and the government agency.

Individuals interested in accessing the data for replication purposes can contact Pierre Bachas ([pbachas@worldbank.org](mailto:pbachas@worldbank.org)) or Anne Brockmeyer ([abrockmeyer@worldbank.org](mailto:abrockmeyer@worldbank.org)). We will provide reasonable assistance to requests for clarification and replication, and will support application for data access through an agreement with each government tax authority.

Researchers interested in obtaining the data for their own analyses can directly contact the tax authorities.

### 2.3 Details on each Data Source

The analysis is mainly based on de-identified administrative raw data provided by the national tax agencies. The main datasets, constructed annually at the firm level from corporate income tax returns, contain data on income, costs, assets, and tax obligations. Data coverage periods differ by country.

The tax administrative data was shared directly with us only for some of the countries involved in the project. In other cases, the data was preserved at corresponding country offices and the local staff conducted the analysis. The table below describes the specific situation in each country on a case-by-case basis.

Country	Public data?	Shared data?	Tax Administration Office	Date of access
Albania	No	Yes	General Directorate of Taxation (DPT)	May 2020
Colombia	No	No	Dirección de Impuestos y Aduanas Nacionales (DIAN)	July 2024
Costa Rica	No	Yes	Ministerio de Hacienda	August 2021
Dominican Republic	No	Yes	Dirección General de Impuestos Internos (DGII)	November 2019
Ecuador	No	Yes	Servicio de Rentas Internas (SRI)	August 2018
Eswatini	No	No	Eswatini Revenue Service (ERS)	July 2020
Ethiopia	No	No	Ethiopian Revenues and Customs Authority (ERCA)	June 2018
Greece	No	No	Independent Authority for Public Revenue (IAPR)	January 2021
Guatemala	No	Yes	Superintendencia de Administración Tributaria (SAT)	February 2016
Honduras	No	Yes	Servicio de Administración de Rentas (SAR)	April 2022
Jamaica	No	Yes	Tax Administration Jamaica (TAJ)	April 2024
Mexico	Yes <sup>a</sup>	Yes	Servicio de Administración Tributaria (SAT)	November 2018
Rwanda	No	Yes	Rwanda Revenue Authority (RRA)	June 2021
Senegal	No	Yes	Directorate General of Taxes and Domains (DGID)	April 2021
South Africa	No	No	South African Revenue Service (SARS)	October 2022
Uganda	No	No	Uganda Revenue Authority (URA)	June 2020

<sup>a</sup>This data was public at the time we accessed it, but the Mexican government later removed it.

Additional details follow for the 6 countries with unshared data:

- Colombia. This data is owned by the Colombian General Directorate of Taxation (DIAN). The data was shared in the context of a research collaboration with its director of studies at the time, David Suarez Castellanos. Please contact Pierre Bachas [pbachas@worldbank.org](mailto:pbachas@worldbank.org) for more information.
- Eswatini. This data is owned by the Eswatini Revenue Service (ERS). This data was made available thanks to collaboration with the Research, Strategy and Statistics Department of the Eswatini Revenue Service and the International Centre for Taxation and Development (ICTD). Contact Anne Brockmeyer [abrockmeyer@worldbank.org](mailto:abrockmeyer@worldbank.org) for more information.
- Ethiopia. This data is owned by the Ethiopian Revenues and Customs Authority (ERCA). The data was made available to the the Institute for Fiscal Studies (IFS) as part of their ongoing collaboration with the Tax Policy Directorate of the Ministry of Finance. IFS affiliated staff may be able to get access to the data. Contact Anne Brockmeyer [abrockmeyer@worldbank.org](mailto:abrockmeyer@worldbank.org) for more information.
- Greece. This data is owned by the Greek Authority for Public Revenue (IAPR). The data was accessed thanks to collaboration with Evangelos Koumanakos. Contact Pierre Bachas [pbachas@worldbank.org](mailto:pbachas@worldbank.org) for more information.

- South Africa. This data is owned by the South African Revenue Service (SARS). The data was accessed on site in a secure data facilities, in collaboration with the United Nations University World Institute for Development Economics Research (UNU-WIDER). Contact Camille Semelet [semelet@ifo.de](mailto:semelet@ifo.de) for more information.
- Uganda. This data is owned by the Uganda Revenue Authority (URA). This data was accessed as part of a collaboration with the Research and Innovation Lab of the Ugandan Revenue Authority and the Overseas Development Institute (ODI), and UNU-WIDER. Contact Anne Brockmeyer [abrockmeyer@worldbank.org](mailto:abrockmeyer@worldbank.org) for more information.

We provide below details of the additional supplementary data used throughout the study for various analyses:

- Our project relies on data extracted from Orbis BvD, a data platform that provides information on private and public companies around the world. We use this data to extract the necessary information to identify multinational groups and their linked firms in the study countries and to obtain ownership information. This data was last accessed in October 2024.
- Data on population, inflation, exchange rates, and other macroeconomic variables was obtained from the [World Development Indicators](#) through the [DataBank](#). For certain countries, we use exchange rates data from the OECD. GDP per capita data was last accessed in September 2023. The remaining data was last accessed in December 2024.
  - Exchange rate data for Greece was obtained from the [OECD](#) data explorer tool. This data was last accessed in July 2024.
- We utilize tax expenditure data from the Global Tax Expenditures Database for select analyses. This data can be download from the [GTED](#) website. This data was last accessed in September 2023.
- The study also uses aggregate country-by-country reports (CbCR) data compiled by the [EU Tax Observatory](#) and the [OECD](#). This data was last accessed in November 2023 and March 2024, respectively.
- We also use data from the [World Bank Entrepreneurship Database](#) for some countries (when available) for the number of firms and surplus data from the [World Inequality Database](#). This data was last accessed in September and October 2024.
- The study also draws on country-specific production data from national statistical offices and central bank databases ([INEGI](#) for Mexico, [INSTAT](#) for Albania, [Banco Central de Honduras](#), [Banco Central de Republica Dominicana](#), [Banco Central de Costa Rica](#), [Banco de Guatemala](#), [Banco de Ecuador](#), [CEPAL](#) for Jamaica, [ELSTAT](#) for Greece, [Government of Eswatini](#), [DANE](#) for Colombia). This data was last accessed in October 2024.
- For some countries we rely on specific tax expenditure data obtained from several sources. The first source is the [World Revenue Longitudinal Data \(WoRLD\)](#). This data on tax expenditures was last accessed in September 2023. The second source is the Government Revenue Dataset from the [UNU Wider](#). This data was accessed in March 2023. Finally, we also use data from Bachas et. al (2022). This data (accessed in September 2023) is available [here](#).

### 3 Datasets

#### 3.1 Input Datasets

Dataset	Description	Notes	Provided
renta.txt	Costa Rica corporate tax data.	Confidential	No
SIIAT_RENTA_‘i’.txt	Costa Rica corporate tax data.	Confidential	No
D101_Renta_SIIAT_PF2016.txt	Costa Rica corporate tax data.	Confidential	No
SAP_RENTA_1.txt	Costa Rica corporate tax data.	Confidential	No
SAP_RENTA_2.txt	Costa Rica corporate tax data.	Confidential	No
D101_Renta_SAP_PF2016.xlsx	Costa Rica corporate tax data.	Confidential	No
D101_Renta_SAP_PF‘i’.del	Costa Rica corporate tax data.	Confidential	No
RENTA_VENTAS_RUT_PDOS_2008_2016.xlsx	Costa Rica corporate tax data.	Confidential	No
vD101_Renta_‘i’.txt	Costa Rica corporate tax data.	Confidential	No
id_convertor_CRI	Costa Rica IDs data.	Confidential	No
c_clase.dta	Costa Rica entity type data.	Confidential	No
a_contribuyentes.dta	Costa Rica firm level data.	Confidential	No
Contribuyentes.txt	Costa Rica firm level data.	Confidential	No
Diccionario de tablas.xlsx	Costa Rica entity type data.	Confidential	No
2021-08-13 Datos identificativos_Final.xlsx	Costa Rica sectors data.	Confidential	No
orbis_ids_multinationals.txt	Costa Rica Orbis IDs data.	Confidential	No
CRI_Lasso_Estimates.txt	Costa Rica Lasso estimates.	Confidential	No
Listado_contribuyentes_zona_franca.xlsx	Costa Rica franc zones data.	Confidential	No
D152_‘i’.xlsx	Costa Rica withholding tax data.	Confidential	No
RENTA_TD_2017_2018.xlsx	Costa Rica corporate tax data.	Confidential	No
CIU4_firmsectors.dta	Sectors data (many countries).	Confidential	No
DR_TaxData_Excel.xlsx	Dom. Rep. corporate tax data.	Confidential	No
bal‘i’.dta	Ecuador corporate tax data.	Confidential	No
combined data.dta	Ethiopia corporate tax data.	Confidential	No
LOC_ETH_opencagegeo.csv	Ethiopia firm level data.	Confidential	No
taxreturns_noduplicates_jan2021.csv	Greece corporate tax data.	Confidential	No
GRC_Subs_Scope_GUO.csv	Greece Orbis IDs data.	Confidential	No

Dataset	Description	Notes	Provided
GRC_export_econ_var_2018.xlsx	Greece firm level data.	Confidential	No
legal status code description.xlsx	Greece entity type data.	Confidential	No
GRC_Lasso_Estimates.csv	Greece Lasso estimates.	Confidential	No
ISR_SOBRE_UTILIDADES_DC_26-92_PF2006-2012.txt	Guatemala corporate tax data.	Confidential	No
GPD_ISR_DC10-2012_P2013-2014.txt	Guatemala corporate tax data.	Confidential	No
ISR_ANUAL_2016_2018.txt	Guatemala corporate tax data.	Confidential	No
ISR_ANUAL_2019_2020.txt	Guatemala corporate tax data.	Confidential	No
PADRON_CONTRIBUYENTES.txt	Guatemala firm level data.	Confidential	No
CIT_Dataset_26_4_2022.dta	Honduras corporate tax data.	Confidential	No
GUO_ISRPJ_2014_2020_0305_WB.dta	Honduras Orbis IDs data.	Confidential	No
CIT_2018_2022.xlsx	Jamaica corporate tax data.	Confidential	No
CIT_Sch_1_2018_2022.xlsx	Jamaica corporate tax data.	Confidential	No
sch1_ally.dta	Jamaica corporate tax data.	Confidential	No
IT15_2018-2022.xlsx	Jamaica corporate tax data.	Confidential	No
WB_OECD_ORBIS_Match.xlsx	Jamaica Orbis IDs data.	Confidential	No
OECD_Req_Schedule A.xlsx	Jamaica payroll data.	Confidential	No
OECD_Req_Schedule C.xlsx	Jamaica payroll data.	Confidential	No
AT01_2018.xlsx	Jamaica assets data.	Confidential	No
AT02_2018.xlsx	Jamaica assets data.	Confidential	No
OECD_Req_AT01_2016-2017.xlsx	Jamaica assets data.	Confidential	No
OECD_Req_AT02_2016-2017.xlsx	Jamaica assets data.	Confidential	No
Anuales_ISR_PM_‘i’.csv	Mexico corporate tax data.	Confidential	No
Catalogo_RFCs_PM.txt	Mexico sectors data.	Confidential	No
CIT_‘i’.csv	Rwanda corporate tax data.	Confidential	No
IS_appended.dta	Senegal corporate tax data.	Confidential	No
Afristat_conversion.dta	Senegal firm level data.	Confidential	No
Liste_activite.xlsx	Senegal sector data.	Confidential	No
TVA_appended.dta	Senegal firm costs data.	Confidential	No
RAS_IRPP_appended.dta	Senegal firm costs data.	Confidential	No
Jul2019_CIT_2013_2018_Clean.dta	Eswatini corporate tax data.	Confidential	No
Registry_Location.dta	Eswatini location data.	Confidential	No
CITanalysis_Raw_Extravars2.dta	Uganda corporate tax data.	Confidential	No
citfp_2008_2022_e5_v1.dta	South Africa corporate tax data.	Confidential	No
CIT-IRP5_employee.dta	South Africa employee data.	Confidential	No
mne_ind_cbc.dta	South Africa CbCR data.	Confidential	No
EUR_USD_exchange_rate.dta	South Africa exchange rate data.	Confidential	No

Dataset	Description	Notes	Provided
revenue-forgone-of-gdp.csv	GTED tax expenditure data.	Public	No
share-of-total-tax-expen.csv	GTED tax expenditure data (computed as the ratio of Revenue Forgone from Corporate Income Tax to Total Revenue Forgone).	Public	No
UNUWIDERGRD_2022_Full.xlsx	UNU WIDER revenue data.	Public	No
WoRLD_timeSeries.csv	Corporate tax revenue data (in % of GDP) from IMF World Revenue Longitudinal Data.	Public	No
globalETR_bfjz.csv	Harmonized tax revenue data from <a href="#">Bachas et al. (2022)</a> .	Public	No
GDPPC.csv	GDP per capita data.	Public	No
EUTO_CbCR_data_2023-11-09	MNEs ownership data	Confidential	No
Export_COUNTRY	Subsidiaries Orbis data	Confidential	No
Export_GUO_New_Alt_YEAR	MNEs Orbis data	Confidential	No
Export_GUO_New_Alt_YEAR_t500	MNEs Orbis data	Confidential	No
Count_Turnover	MNEs turnover data extracted from Orbis	Confidential	No
MNE_Count	MNEs data extracted from Orbis	Confidential	No
Subsidiaries_Count	Subsidiaries data extracted from Orbis	Confidential	No
TableA1_ETR	Production data sourced from national data agencies	Public	No
Total LLCs	World Bank Entrepreneurship dataset information	Confidential	No
WDI_all_correlation_	WDI macroeconomic data for all countries	Confidential	No
WID_Data_12092024-181642	Surplus data (Gross operating surplus and mixed income for all ages and individuals) sourced from the World Inequality Database	Public	No



### 3.2 Datasets for Analysis

Dataset	Description	Notes	Provided
WDI_vars_ <i>COUNTRY</i> .dta	WDI macroeconomic data by country.	Public	No
OECD_GRC_EXCHANGE_RATES_2024.csv	Exchange rate data (Greece)	Public	No
index_2019.csv	US CPI data.	Public	No
<i>COUNTRY</i> _withvars.csv	Clean corporate tax datasets at the firm level.	Confidential	No
CBCR_TABLEI_06032024193031092.csv	CbCr Aggregate Data	Public	No

## 4 Computational Requirements

### Software requirements

The data creation and analysis code is written in **Stata** and R. Version 16 of Stata is used.

### Memory and runtime requirements

We suggest the use of a processor with at least 8 cores and a RAM of 16 GB. With those settings, the replicator can expect full data creation to take about 60 minutes. Around 40 minutes can be expected to run the full data analysis.

## 5 Instructions to replicators

### 5.1 Folder structure

For full replication of the project, we suggest the following, complete folder structure.

- Auxillary data
  - GTED
  - orbis
    - Final data
    - Intermediate data
    - Raw data
- codes
  - Auxillary codes
    - orbis
  - cleaning
- input
  - prep
  - WDI
- output
  - figures
  - GMT
  - metadata
  - regressions
  - tables

This folder structure is reflected in the file `0_master_ETR_GMT.R` (which we provide in the code repository). We provide the code files to create the datasets used in the project and the subsequent analysis. The `0_master_ETR_GMT.R` R script and the remaining R and Stata files correspond to the above folder tree. The codes described in section 7 rely on these folder paths. The whole project analysis can be reproduced by just executing the `0_master_ETR_GMT.R` R script.

## 6 List of tables, figures and programs

The provided code reproduces:

- ✓ All numbers provided in text in the paper
- ✓ All tables and figures in the paper
- Selected tables and figures in the paper, as explained and justified below.

### 6.1 Mapping of tables and analysis code

Table 1 provides a mapping between all tables of the paper (including the online appendix) and the codes producing these results. The codes generate *.tex* files containing the content of the tables.

Table 1: Mapping of tables and analysis code

Table	Code	File Name
Table 1	2.3_output_ETR.R	T1_desc_stats.tex
Table 2	2.3_output_ETR.R	T2_regression_ETR_Top1.tex
Table A1	<i>Not generated by code</i>	TA1_literaturenew.tex
Table A2	2.3_output_ETR.R	TA2_coverage_tax_data.tex
Table A3	2.3_output_ETR.R	TA3_panelA_bins_all.tex TA3_panelB_bins_profitable.tex
Table A4	2.3_output_ETR.R	TA4_top1_stats.tex
Table A5	2.3_output_ETR.R	TA5_regression_ETR_B90
Table A6	2.3_output_ETR.R	TA6_regression_robustness.tex
Table A7	2.3_output_ETR.R	TA7_regression_A_Top1_All.tex TA7_regression_B_B90_All
Table A8	<i>Not generated by code</i>	TA8_exemption.tex
Table A9	2.3_output_ETR.R	TA9_Top1_SEZ.tex
Table A10	<i>Not generated by code</i>	<i>Not generated by code</i>
Table A11	05_Counts_GMT	TA11_Count_Table.tex
Table A12	04_Count_OrbisData_on_Firms	TA12_Count_Turnover_Present.tex
Table B1	2_4_GMT_output.do	TB1_GMT_Stats.tex
Table B2	<i>Not generated by code</i>	TB2_data_availability.tex

### 6.2 Mapping of figures and analysis code

Table 2 provides a mapping between all figures of the paper (including the online appendix) and the code producing these results. The codes generate *.pdf* or *.png* files containing the graphs.

Table 2: Mapping of figures and analysis code

Figure	Panels	Code	File Name
Figure 1		<i>Not generated by code</i>	F1_CITbreakdown231024.pdf
Figure 2		2.3_output_ETR.R	F2_ETR_size_all.pdf
Figure 3		2.3_output_ETR.R	F3_ETR_minus_STR.pdf
Figure 4	A, B, C	2.3_output_ETR.R	A: F4_a_ETR_country_avg.pdf B: F4_b_ETR_sector_avg.pdf C: F4_c_ETR_lifetime_avg.pdf
Figure 5	A, B, C, D	2.3_Minimum_Tax_Top1_output.R	A: F5_a_share_less15_nodat.pdf B: F5_b_ETR_below15_nodat.pdf C: F5_c_profit_less15_nodat.pdf D: F5_d_revchange_nodat.pdf
Figure 6	A, B	2_4_GMT_output.do	A: F6_GMT_Share_Tax_Orbis_CbC_topup.png B: F6_GMT_Share_N_G_E_Orbis_CbC_topup.png
Figure 7	A, B, C, D, E	2_4_GMT_output.do	A: F7_CRI_GMT_Share_Tax_Cbc_topup.png B: F7_GRC_GMT_Share_Tax_Cbc_topup.png C: F7_HND_GMT_Share_Tax_Cbc_topup.png D: F7_JAM_GMT_Share_Tax_Cbc_topup.png E: F7_ZAF_GMT_Share_Tax_Cbc_topup.png
Figure A1		<i>Not generated by code</i>	FA1_taxform_example_RWA.png
Figure A2	A, B	Code_FA2.do	A: FA2_a.png B: FA2_b.png
Figure A3		2.3_output_ETR.R	FA3_ETR_size_profitable.pdf
Figure A4		2.3_output_ETR.R	FA4_unprofitable.pdf
Figure A5		2.3_output_ETR.R	FA5_regressions_robustness_STR.pdf
Figure A6		2.3_output_ETR.R	FA6_ETR_size_robust.pdf
Figure A7	A, B	2.3_output_ETR.R	A: FA7_panelA_correlation.pdf B: FA7_panelB_missing.pdf
Figure A8		2.3_output_ETR.R	FA8_ETR_foreigncredit.pdf
Figure A9	A, B, C, D	2.3_Minimum_Tax_Top1_output.R	A: FA9_a_share_less15.pdf B: FA9_b_ETR_below15.pdf C: FA9_c_profit_less15.pdf D: FA9_d_revenue_change.pdf
Figure B1	A, B, C, D, E	2_4_GMT_output.do	A: FB1_CRI_GMT_Share_Tax_Orbis_topup.png B: FB1_GRC_GMT_Share_Tax_Orbis_topup.png C: FB1_HND_GMT_Share_Tax_Orbis_topup.png D: FB1_JAM_GMT_Share_Tax_Orbis_topup.png E: FB1_ZAF_GMT_Share_Tax_Orbis_topup.png
Figure B2	A, B	2_4_GMT_output.do	A: FB2_GMT_Share_Tax_Orbis_CbC_topup_Y10.png B: FB2_GMT_Share_N_G_E_Orbis_CbC_topup_Y10.png
Figure B3		2_4_GMT_output.do	FB3_GMT_Share_Tax_Group_Entity.png
Figure B4	A, B, C, D	2_4_GMT_output.do	A: FB4_a_GMT_N_Groups_Orbis_V.png B: FB4_b_GMT_Share_Profits_Orbis_V.png C: FB4_c_GMT_N_Groups_CbCR_V_topup.png D: FB4_d_GMT_Share_Profits_CbCR_V_topup.png

## 7 Description of programs and code

The project is organized as follows. First, we use the raw data files to create the datasets used in the analysis. These codes are described in section 7.1. Second, we describe the data analysis code in section 7.2.

Some of the code rely on programs downloaded from the Statistical Software Components (SSC) archive. These programs are described in section 7.3.

### 7.1 Creation of data sets for analysis

#### 0\_master\_cleaning.do

This do-file executes centrally all the individual cleaning do-files.

---

#### CIT\_ALB.do

This do-file creates the dataset ALB\_withvars.csv.

Input data:

```
$data/input/Albania/15.05.2020 TF 2015-A-2019-A.xlsx  
$data/input/Albania/NACE_REV2_US_NAICS_2017.xlsx  
$data/input/Albania/NAICS_2-6 digit_2017_Codes.xlsx  
$data/input/Albania/"Shutdown" sectors as defined by Vavra (2020).xlsx
```

Output data:

```
$data/output/Albania/ALB_withvars.csv
```

---

#### CIT\_COL.do

This do-file creates the dataset COL\_withvars.csv.

Input data:

```
$data/input/Colombia/df_panel_110_otras.csv
```

Output data:

```
$data/output/Colombia/COL_withvars.csv
```

---

#### CIT\_CRI.do

This do-file creates the dataset CRI\_withvars.csv.

Input data:

```
$data/input/Costa Rica/renta.txt  
$data/input/Costa Rica/SIIAT_RENTA_‘i’.txt  
$data/input/Costa Rica/D101_Renta_SIIAT_PF2016.txt  
$data/input/Costa Rica/id_convertor_CRI  
$data/input/Costa Rica/SAP_RENTA_1.txt  
$data/input/Costa Rica/SAP_RENTA_2.txt  
$data/input/Costa Rica/D101_Renta_SAP_PF2016.xlsx
```

```
$data/input/Costa Rica/D101_Renta_SAP_PF'i'.del
$data/input/Costa Rica/c_clase.dta
$data/input/Costa Rica/RENTA_VENTAS_RUT_PDOS_2008_2016.xlsx
$data/input/Costa Rica/a_contribuyentes.dta
$data/input/Costa Rica/Contribuyentes.txt
$data/input/Costa Rica/vD101_Renta_'i'.txt
$data/input/Costa Rica/Diccionario de tablas.xlsx
$data/input/Costa Rica/2021-08-13 Datos identificativos_Final.xlsx
$data/input/Costa Rica/orbis_ids_multinationals.txt
$data/input/Costa Rica/CRI_Lasso_Estimates.txt
$data/input/Costa Rica/Listado_contribuyentes_zona_franca.xlsx
```

Output data:

```
$data/output/Costa Rica/CRI_withvars.csv
```

---

### **CRI\_Lasso\_Estimates.do**

This do-file creates the dataset CRI\_Lasso\_Estimates.dta.

Input data:

```
$data/input/Costa Rica/D152_'i'.xlsx
$data/input/Costa Rica/RENTA_TD_2017_2018.xlsx
$data/input/Costa Rica/Diccionario de tablas.xlsx
$data/input/Costa Rica/2021-08-13 Datos identificativos_Final.xlsx
$data/input/Costa Rica/CRI_withvars_temp.dta
```

Output data:

```
$data/output/Costa Rica/CRI_Lasso_Estimates.dta
```

---

### **CIT\_DOM.do**

This do-file creates the dataset DOM\_withvars.csv.

Input data:

```
$data/input/Dominican Republic/DR_TaxData_Excel.xlsx
$data/input/Dominican Republic/CIIU4_firmsectors.dta
```

Output data:

```
$data/output/Dominican Republic/DOM_withvars.csv
```

---

### **CIT\_ECU.do**

This do-file creates the dataset ECU\_withvars.csv.

Input data:

```
$data/input/Ecuador/bal'i'.dta
$data/input/Ecuador/CIIU4_firmsectors.dta
```

Output data:

\$data/output/Ecuador/ECU\_withvars.csv

---

### **CIT\_ETH.do**

This do-file creates the dataset ETH\_withvars.csv.

Input data:

\$data/input/Ethiopia/combined data.dta  
\$data/input/Ethiopia/LOC\_ETH\_opencagegeo.csv

Output data:

\$data/output/Ethiopia/ETH\_withvars.csv

---

### **CIT\_GRC.R**

This R-file creates the dataset GRC\_withvars.csv.

Input data:

\$data/input/Greece/taxreturns\_noduplicates\_jan2021.csv  
\$data/input/Greece/GRC\_Subs\_Scope\_GUO.csv  
\$data/input/Greece/GRC\_export\_econ\_var\_2018.xlsx  
\$data/input/Greece/legal status code description.xlsx  
\$data/input/Greece/GRC\_Lasso\_Estimates.csv

Output data:

\$data/output/Greece/GRC\_withvars.csv

---

### **GRC\_Lasso.do**

This R-file creates the dataset GRC\_Lasso\_Estimates.csv.

Input data:

\$data/input/Greece/GRC\_withvars.csv

Output data:

\$data/output/Greece/GRC\_Lasso\_Estimates.csv

---

### **CIT\_GTM.do**

This do-file creates the dataset GTM\_withvars.csv.

Input data:

\$data/input/Guatemala/ISR\_SOBRE\_UTILIDADES\_DC\_26-92\_PF2006-2012.txt  
\$data/input/Guatemala/GPD\_ISR\_DC10-2012\_P2013-2014.txt  
\$data/input/Guatemala/ISR\_ANUAL\_2016\_2018.txt  
\$data/input/Guatemala/PADRON\_CONTRIBUYENTES.txt  
\$data/input/Guatemala/CIU4\_firmsectors.dta  
\$data/input/Guatemala/ISR\_ANUAL\_2019\_2020.txt

Output data:



\$data/output/Guatemala/GTM\_withvars.csv

---

### **CIT\_HND.do**

This do-file creates the dataset HND\_withvars.csv.

Input data:

\$data/input/Honduras/CIT\_Dataset\_26\_4\_2022.dta

\$data/input/Honduras/GUO\_ISRPJ\_2014\_2020\_0305\_WB.dta

Output data:

\$data/output/Honduras/HND\_withvars.csv

---

### **CIT\_JAM.do**

This do-file creates the dataset JAM\_withvars.csv and the dataset sch1\_ally.dta.

Input data:

\$data/input/Jamaica/CIT\_2018\_2022.xlsx

\$data/input/Jamaica/CIT\_Sch\_1\_2018\_2022.xlsx

\$data/input/Jamaica/WB\_OECD\_ORBIS\_Match.xlsx

\$data/input/Jamaica/JAM\_SEZ.dta

\$data/input/Jamaica/SO2.dta

\$data/input/Jamaica/AT01\_AT02\_Assets\_16\_18.dta

Output data:

\$data/output/Jamaica/JAM\_withvars.csv

\$data/input/Jamaica/sch1\_ally.dta

---

### **CIT\_JAM\_SEZ.do**

This do-file creates the dataset JAM\_SEZ.dta. This file cleans tax form data for firms in Jamaica in special economic zones and is used as input in the do-file CIT\_JAM.do.

Input data:

\$data/input/Jamaica/IT15\_2018-2022.xlsx

\$data/input/Jamaica/sch1\_ally.xlsx

\$data/input/Jamaica/WB\_OECD\_ORBIS\_Match.xlsx

Output data:

\$data/input/Jamaica/JAM\_SEZ.dta

---

### **AT01\_AT02\_JAM.do**

This do-file creates the datasets AT01\_AT02\_Assets\_16\_18.dta and AT01\_AT02\_Assets.dta. This file cleans asset tax form data. The datasets generated are used as input in the do-files CIT\_JAM.do and CIT\_JAM\_SEZ.do.

Input data:

```
$data/input/Jamaica/AT01_2018.xlsx
$data/input/Jamaica/AT02_2018.xlsx
$data/input/Jamaica/OECD_Req_AT01_2016-2017.xlsx
$data/input/Jamaica/OECD_Req_AT02_2016-2017.xlsx
```

Output data:

```
$data/input/Jamaica/AT01_AT02_Assets_16_18.dta
```

---

### **SO2\_JAM.do**

This do-file cleans payroll data and creates the dataset SO2.dta. This dataset is used as input in the do-files CIT\_JAM.do.

Input data:

```
$data/input/Jamaica/OECD_Req_Schedule A.xlsx
$data/input/Jamaica/OECD_Req_Schedule C.xlsx
```

Output data:

```
$data/input/Jamaica/SO2.dta
```

---

### **CIT\_MEX.do**

This do-file creates the dataset MEX\_withvars.csv.

Input data:

```
$data/input/Mexico/Anuales_ISR_PM_‘i’.csv
$data/input/Mexico/Catalogo_RFCs_PM.txt
```

Output data:

```
$data/output/Mexico/MEX_withvars.csv
```

---

### **CIT\_RWA.do**

This do-file creates the dataset RWA\_withvars.csv.

Input data:

```
$data/input/Rwanda/CIT ‘i’.csv
```

Output data:

```
$data/output/Rwanda/RWA_withvars.csv
```

---

### **CIT\_SEN.do**

This do-file creates the dataset SEN\_withvars.csv.

Input data:

```
$data/input/Senegal/IS_appended.dta
$data/input/Senegal/Afristat_conversion.dta
$data/input/Senegal/Liste_activite.xlsx
$data/input/Senegal/TVA_appended.dta
```

```
$data/input/Senegal/RAS_IRPP_appended.dta
```

Output data:

```
$data/output/Senegal/SEN_withvars.csv
```

---

### **CIT\_SWZ.do**

This do-file creates the dataset ESW\_withvars.csv.

Input data:

```
$data/input/Eswatini/Jul2019_CIT_2013_2018_Clean.dta
```

```
$data/input/Eswatini/Registry_Location.dta
```

Output data:

```
$data/output/Eswatini/ESW_withvars.csv
```

---

### **CIT\_UGA.do**

This do-file creates the dataset UGA\_withvars.csv.

Input data:

```
$data/input/Uganda/CITanalysis_Raw_Extravars2.dta
```

Output data:

```
$data/output/Uganda/UGA_withvars.csv
```

---

### **CIT\_ZAF.do**

This do-file creates the dataset ZAF\_withvars.csv.

Input data:

```
$data/input/South Africa/citfp_2008_2022_e5_v1.dta
```

```
$data/input/South Africa/CIT-IRP5_employee.dta
```

```
$data/input/South Africa/mne_ind_cbc.dta
```

```
$data/input/South Africa/EUR_USD_exchange_rate.dta
```

Output data:

```
$data/output/South Africa/ZAF_withvars.csv
```

---

## **7.2 Analysis**

### **0\_master\_ETR\_GMT.R**

This R-file centrally executes the project analysis.

Input files:

```
$data/output/Albania/ALB_withvars.csv
```

```
$data/output/Colombia/COL_withvars.csv
```

```
$data/output/Costa Rica/CRI_withvars.csv
```

```
$data/output/Dominican Republic/DOM_withvars.csv
```

```
$data/output/Ecuador/ECU_withvars.csv
$data/output/Eswatini/ESW_withvars.csv
$data/output/Ethiopia/ETH_withvars.csv
$data/output/Greece/GRC_withvars.csv
$data/output/Guatemala/GTM_withvars.csv
$data/output/Honduras/HND_withvars.csv
$data/output/Jamaica/JAM_withvars.csv
$data/output/Mexico/MEX_withvars.csv
$data/output/Rwanda/RWA_withvars.csv
$data/output/Senegal/SEN_withvars.csv
$data/output/South Africa/ZAF_withvars.csv
$data/output/Uganda/UGA_withvars.csv
```

---

## 1\_prepare\_data.R

This R-file runs a pre-cleaning procedure before starting the analysis.

Input files:

```
$data/output/'COUNTRY'/'COUNTRY'_withvars.csv
$output/GMT/GRC_Groups_cbr_TA9.xlsx
$output/GMT/CRI_Groups_cbr_TA9.xlsx
$output/GMT/JAM_Groups_cbr_TA9.xlsx
$output/GMT/HND_Groups_cbr_TA9.xlsx
$input/WDI/WDI_vars_COUNTRY.xlsx
$input/prep/OECD_GRC_EXCHANGE_RATES_2024.csv
$input/prep/index_2019.csv
```

---

## 2.1\_GMT\_simulations.R

This do-file runs the GMT analysis corresponding for the subset of five countries (Costa Rica, Greece, Honduras, Jamaica and South Africa) for which we simulate a QDMTT scenario and creates input files that are used in the analysis do-file 2\_4\_GMT\_output.do.

Input files:

```
$input/WDI/WDI_vars_COUNTRY.xlsx
$input/prep/CBCR_TABLEI_06032024193031092.csv
$input/prep/OECD_GRC_EXCHANGE_RATES_2024.csv
$input/prep/index_2019.csv
$data/output/'COUNTRY'/'COUNTRY'_withvars.csv
```

Output files:

```
$output/metadata/GMT_15/'c'15_step2_topup_orbis.csv
$output/metadata/GMT_15/'c'15_step2_topup_cbr_topup.csv
$output/metadata/GMT_15/'c'15_step3_topup_nocarv_cbr_topup.csv
$output/metadata/GMT_15/'c'15_step3_topup_nocarv_orbis.csv
```

```
$output/metadata/GMT_15/'c'15_step5_topup_carv_orbis.csv
$output/metadata/GMT_15/'c'15_step5_topup_carv_cbr_topup.csv
$output/metadata/GMT_15/'c'15_step5_topup_carv_cbr_topup_QRTC.csv
$output/metadata/GMT_15/'c'20_step5_topup_carv_cbr_topup.csv
$output/metadata/GMT_15/'c'15_step5_topup_carv_orbis_QRTC.csv
$output/metadata/GMT_15/'c'20_step5_topup_carv_orbis.csv
$output/metadata/GMT_15/'c'15_step5_topup_carv_orbis_nogroup.csv
$output/metadata/GMT_15/'c'15_aggregate_statistics.csv
$output/metadata/GMT_15/'c'15_1pct_revenue_simul.csv
$output/metadata/GMT_15/'c'15_1pct_revenue_simul_nodat.csv
$output/metadata/GMT_15/'c'15_List_Carveouts_Final_cbr_topup10.csv
```

---

## 2.2\_Minimum\_Tax\_Top1.R

This R script runs the top 1% GMT analysis with and without deferred tax assets for the remaining countries of the study sample.

Input files:

```
$input/WDI/WDI_vars_COUNTRY.xlsx
$input/prep/OECD_GRC_EXCHANGE_RATES_2024.csv
$input/prep/index_2019.csv
$data/output/'COUNTRY'/'COUNTRY'_withvars.csv
```

Output files:

```
$output/metadata/'c'_aggregate_statistics.csv
$output/metadata/'c'_1pct_revenue_simul.csv
$output/metadata/'c'_1pct_revenue_simul_nodat.csv
```

---

## 2.3\_Minimum\_Tax\_Top1\_output.R

This R script runs the top 1% simulation graphs that correspond to Figures 5 and A9.

Input files:

```
$output/metadata/'c'_1pct_revenue_simul.csv
$output/metadata/'c'_1pct_revenue_simul_nodat.csv
$output/metadata/GMT_15/'c'15_1pct_revenue_simul.csv
$output/metadata/GMT_15/'c'15_1pct_revenue_simul_nodat.csv
$output/metadata/df.ETR.size.p.RDS
```

Output files:

```
$output/figures/FA9_a_share_less15.pdf
$output/figures/F5_a_share_less15_nodat.pdf
$output/figures/FA9_b_ETR_below15.pdf
$output/figures/F5_b_ETR_below15_nodat.pdf
$output/figures/FA9_c_profit_less15.pdf
$output/figures/F5_c_profit_less15_nodat.pdf
$output/figures/FA9_d_revenue_change.pdf
```

```
$output/figures/F5_d_revchange_nodat.pdf
$output/figures/F5_Data_Below15.csv
```

---

## 2\_4\_GMT\_output.do

This do-file runs the GMT analysis corresponding for the subset of five countries (Costa Rica, Greece, Honduras, Jamaica, and South Africa) for which we simulate a QDMTT scenario. This do-file uses as input the files generated by the simulation R script 2.1\_GMT\_simulations.R and creates Table B1 and Figures 6, 7, B1, B2, B3, and B4.

Input files:

```
$output/metadata/GMT_15/'c'15_step2_topup_orbis.csv
$output/metadata/GMT_15/'c'15_step2_topup_cbcr_topup.csv
$output/metadata/GMT_15/'c'15_step3_topup_nocarv_cbcr_topup.csv
$output/metadata/GMT_15/'c'15_step3_topup_nocarv_orbis.csv
$output/metadata/GMT_15/'c'15_step5_topup_carv_orbis.csv
$output/metadata/GMT_15/'c'15_step5_topup_carv_cbcr_topup.csv
$output/metadata/GMT_15/'c'15_step5_topup_carv_cbcr_topup_QRTC.csv
$output/metadata/GMT_20/'c'20_step5_topup_carv_cbcr_topup.csv
$output/metadata/GMT_15/'c'15_step5_topup_carv_orbis_QRTC.csv
$output/metadata/GMT_20/'c'20_step5_topup_carv_orbis.csv
$output/metadata/GMT_15/'c'15_step5_topup_carv_orbis_nogroup.csv
$output/metadata/GMT_15/'c'15_aggregate_statistics.csv
$output/metadata/GMT_15/'c'15_List_Carveouts_Final_cbcr_topup10.csv
```

Output files:

```
$output/figures/F7_'c'_GMT_Share_Tax_Cbc_topup.png
$output/figures/FB1_'c'_GMT_Share_Tax_Orbis_topup.png
$output/figures/FB4_GMT_N_Groups_CbCR_V_topup.png
$output/figures/FB4_GMT_N_Groups_Orbis_V.png
$output/figures/F6_GMT_Share_N_G_E_Orbis_CbC_topup.png
$output/figures/FB2_GMT_Share_N_G_E_Orbis_CbC_topup_Y10.png
$output/figures/FB4_d_GMT_Share_Profits_CbCR_V_topup.png
$output/figures/FB4_GMT_Share_Profits_Orbis_V.png
$output/figures/FB3_GMT_Share_Tax_Group_Entity.png
$output/figures/F6_GMT_Share_Tax_Orbis_CbC_topup.png
$output/figures/FB2_GMT_Share_Tax_Orbis_CbC_topup_Y10.png
$output/tables/TB1_GMT_Stats.tex
$output/metadata/Entities_Taxed_TA9.xlsx
$output/metadata/List_Groups_cbcr_TA9.xlsx
$output/metadata/'c'_Groups_cbcr_TA9.xlsx
```

---

## 2.1\_descriptive\_ETR.R

The script processes corporate tax data from multiple countries loaded with the 0\_master\_ETR\_GMT.R-file plus the additional data pre-cleaned in the 1\_prepare\_data.R script and combines them into

standardized formats for downstream analysis.

Output files:

```
$output/metadata/df.stat.decile.usd.RDS
$output/metadata/df.perc.turnover.RDS
$output/metadata/df.perc.distribution.RDS
$output/metadata/df.p90.stat.RDS
$output/metadata/df.ETR.gmt.RDS
$output/metadata/df.ETR.descr.RDS
$output/metadata/df.percentile.overtime.RDS
$output/metadata/df.ETR.size.p.rob.RDS
$output/metadata/df.ETR.size.p.RDS
$output/metadata/df.ETR.size.p.div.RDS
$output/metadata/df.ETR.panel.cross.RDS
$output/metadata/df.ETR.corr.asset.RDS
$output/metadata/df.ETR.avg.table.RDS
$output/metadata/df.taxgap.cap.RDS
$output/metadata/df.table_A1.RDS
$output/metadata/df.ETR.size.num.p.RDS
$output/metadata/df.ETR.share15.p.RDS
$output/metadata/df.ETR.sector.p.RDS
$output/metadata/df.ETR.payroll.RDS
$output/metadata/df.ETR.panel.bal.RDS
$output/metadata/df.ETR.nbemployee.RDS
$output/metadata/df.ETR.largesector.p.RDS
$output/metadata/df.ETR.largesector.dec.RDS
$output/metadata/df.ETR.GDP.RDS
$output/metadata/df.ETR.assets.RDS
```

---

## 2.2 \_regression\_ETR.R

This R-file creates a second set of metadata files to be used as input in the analysis using the a similar procedure as the 2.1\_descriptive\_ETR.R script.

Output files:

```
$output/metadata/df.ETR.fit.T20.RDS
$output/metadata/df.ETR.fit.T20.D5.RDS
$output/metadata/df.ETR.fit.T20.D3.RDS
$output/metadata/df.ETR.fit.T20.D2.RDS
$output/metadata/df.ETR.fit.T20.D01.RDS
$output/metadata/df.ETR.fit.T20.C.RDS
$output/metadata/df.ETR.fit.T10.RDS
$output/metadata/df.ETR.fit.T10.rbst.RDS
$output/metadata/df.ETR.fit.T10.D5.RDS
$output/metadata/df.ETR.fit.T10.D3.RDS
$output/metadata/df.ETR.fit.T10.D2.RDS
```

```
$output/metadata/df.ETR.fit.T10.D01.RDS
$output/metadata/df.ETR.fit.T10.C.RDS
$output/metadata/df.ETR.fit.FE.T10.RDS
$output/metadata/df.ETR.fit.FE.B90.RDS
$output/metadata/df.ETR.fit.B90.RDS
$output/metadata/df.ETR.fit.B90.rbst.RDS
```

---

### **merge\_metadata\_ETR.R**

This R-file appends the metadata files created in the R scripts 2.1\_descriptive\_ETR.R and 2.3\_regression\_ETR to use prepare the input data used in the R script 2.4\_output\_ETR.

---

### **2.3\_output\_ETR**

This R-file uses the metadata files created in the R scripts 2.1\_descriptive\_ETR.R and 2.3\_regression\_ETR as input and generates the analysis output listed below, which corresponds to Tables 1, 2, A2, A3, A4, A5, A6, A7, and A9 and Figures 2, 3, 4, A3, A4, A5, A6, A7, AND A8.

Input files:

```
$output/metadata/df.stat.decile.usd.RDS
$output/metadata/df.perc.turnover.RDS
$output/metadata/df.perc.distribution.RDS
$output/metadata/df.p90.stat.RDS
$output/metadata/df.ETR.gmt.RDS
$output/metadata/df.ETR.descr.RDS
$output/metadata/df.percentile.overtime.RDS
$output/metadata/df.ETR.size.p.rob.RDS
$output/metadata/df.ETR.size.p.RDS
$output/metadata/df.ETR.size.p.div.RDS
$output/metadata/df.ETR.panel.cross.RDS
$output/metadata/df.ETR.corr.asset.RDS
$output/metadata/df.ETR.avg.table.RDS
$output/metadata/df.taxgap.cap.RDS
$output/metadata/df.table_A1.RDS
$output/metadata/df.ETR.size.num.p.RDS
$output/metadata/df.ETR.share15.p.RDS
$output/metadata/df.ETR.sector.p.RDS
$output/metadata/df.ETR.payroll.RDS
$output/metadata/df.ETR.panel.bal.RDS
$output/metadata/df.ETR.nbemployee.RDS
$output/metadata/df.ETR.largesector.p.RDS
$output/metadata/df.ETR.largesector.dec.RDS
$output/metadata/df.ETR.GDP.RDS
$output/metadata/df.ETR.assets.RDS
$output/metadata/df.ETR.fit.T20.RDS
```



```

$output/metadata/df.ETR.fit.T20.D5.RDS
$output/metadata/df.ETR.fit.T20.D3.RDS
$output/metadata/df.ETR.fit.T20.D2.RDS
$output/metadata/df.ETR.fit.T20.D01.RDS
$output/metadata/df.ETR.fit.T20.C.RDS
$output/metadata/df.ETR.fit.T10.RDS
$output/metadata/df.ETR.fit.T10.rbst.RDS
$output/metadata/df.ETR.fit.T10.D5.RDS
$output/metadata/df.ETR.fit.T10.D3.RDS
$output/metadata/df.ETR.fit.T10.D2.RDS
$output/metadata/df.ETR.fit.T10.D01.RDS
$output/metadata/df.ETR.fit.T10.C.RDS
$output/metadata/df.ETR.fit.FE.T10.RDS
$output/metadata/df.ETR.fit.FE.B90.RDS
$output/metadata/df.ETR.fit.B90.RDS
$output/metadata/df.ETR.fit.B90.rbst.RDS
$output/metadata/df.ETR.fit.B90.rbst.RDS
$input/WDI_all_correlation.csv
$input/WDI/WDI_vars_COUNTRY.csv
$input/WID_Data_12092024-181642.xlsx
$input/TableA1_ETR.xlsx
$input/Total LLCs.xlsx

```

Output files:

```

$output/metadata/metadata_taxgap.csv
$output/tables/T1_desc_stats.tex
$output/tables/T2_regression_ETR_Top1.tex
$output/tables/TA2_coverage_tax_data.tex
$output/tables/TA3_panelA_bins_all.tex
$output/tables/TA3_panelB_bins_profitable.tex
$output/tables/TA4_top1_stats.tex
$output/tables/TA5_regression_ETR_B90
$output/tables/TA6_regression_robustness.tex
$output/tables/TA7_regression_A_Top1_All.tex
$output/tables/TA7_regression_B_B90_All
$output/tables/TA9_Top1_SEZ.tex
$output/figures/F2_ETR_size_all.pdf
$output/figures/F3_ETR_minus_STR.pdf
$output/figures/F4_a_ETR_country_avg.pdf
$output/figures/F4_b_ETR_sector_avg.pdf
$output/figures/F4_c_ETR_lifetime_avg.pdf
$output/figures/FA3_ETR_size_profitable.pdf
$output/figures/FA4_unprofitable.pdf
$output/figures/FA5_regressions_robustness_STR.pdf
$output/figures/FA6_ETR_size_robust.pdf

```

\$output/figures/FA7\_panelA\_correlation.pdf  
\$output/figures/FA8\_ETR\_foreigncredit.pdf

---

### **function\_ETR.R**

This R-file builds the ETR functions used in the analysis codes to compute the different ETR measures. This auxiliary file is used in the codes 2.1\_descriptive\_ETR.R and 2.2\_regression\_ETR.R.

---

### **ggtheme.R**

This R-file sets a customized ggplot2 function. This auxiliary file is used in the codes 2.1\_GMT\_simulations.R, 2.2\_Minimum\_Tax\_Top1.R.R, and 2.3\_output\_ETR.R.

---

### **Code\_FA2.do**

This do-file creates panels A and B of paper Figure A2.

Input files:

\$Auxillary data/GTED/revenue-forgone-of-gdp.csv  
\$Auxillary data/GTED/share-of-total-tax-expen.csv  
\$Auxillary data/GTED/UNUWIDERGRD\_2022\_Full.xlsx  
\$Auxillary data/GTED/WoRLD\_timeSeries.csv  
\$output/metadata/metadata\_taxgap.csv  
\$Auxillary data/GTED/globalETR\_bfjz.dta  
\$Auxillary data/GTED/GDPPC.csv

Output files:

\$output/figures/FA2\_a.png  
\$output/figures/FA2\_b.png

---

### **Code\_TA9.R**

This R-file creates Table A9.

Input files:

\$output/metadata/df.ETR.descr  
\$output/metadata/df.ETR.avg.table  
\$output/metadata/df.ETR.size.p  
\$output/metadata/Entities\_Taxed\_TA9.xlsx  
\$output/metadata/F5\_Data\_Below15.csv

Output files:

\$output/tables/TA9\_Top1\_SEZ.tex

---

## **Orbis\_Main.R**

This is master file to create Orbis Tables A11 and A12. This do-file sources the five Orbis related R scripts to generate the aforementioned tables.

---

### **01\_MNEs.R**

This R-file processes MNEs' data.

Input files:

\$Auxillary data/orbis/Raw data/UO/Export\_GUO\_New\_Alt\_YEAR.xlsx

\$Auxillary data/orbis/Raw data/UO/Export\_GUO\_New\_Alt\_YEAR\_t500.xlsx

Output files:

\$Auxillary data/orbis/Intermediate data/MNEs\_2015\_2018.csv

\$Auxillary data/orbis/Intermediate data/MNEs\_2015\_2018\_t500.csv

---

### **02\_Clean\_Subsubs.R**

This R-file processes subsidiaries' data.

Input files:

\$Auxillary data/orbis/Raw data/COUNTRY/Export\_COUNTRY.csv

\$Auxillary data/orbis/Raw data/UO/Export\_GUO\_New\_Alt\_YEAR\_t500

Output files:

\$Auxillary data/orbis/Intermediate data/COUNTRY\_Subsubs.csv

---

### **03\_Subsubs\_Scope.R**

This R-file merges MNEs and subsidiaries data to select those subsidiaries in scope.

Input files:

\$Auxillary data/orbis/Intermediate data/MNEs\_2015\_2018.csv

\$Auxillary data/orbis/Intermediate data/MNEs\_2015\_2018\_t500.csv

\$Auxillary data/orbis/Intermediate data/COUNTRY\_Subsubs.csv

Output files:

\$Auxillary data/orbis/Final data/COUNTRY\_Subsubs\_Scope\_GUO.csv

\$Auxillary data/orbis/Final data/COUNTRY\_Subsubs\_t500\_Scope\_GUO.csv

---

### **04\_Count\_OrbisData\_on\_Firms.R**

This R creates Table A12.

Input files:

\$Auxillary data/orbis/Raw data/General Extractions/Count\_Turnover.xlsx

\$Auxillary data/orbis/Raw data/General Extractions/Subsidiaries\_Count.xlsx

\$Auxillary data/orbis/Intermediate data/COUNTRY\_Subsubs.csv

Output files:

```
$output/tables/TA12_Count_Turnover_Present.csv  
$output/tables/TA12_Count_Turnover_Present.tex
```

---

## 05\_Counts\_GMT.R

This R creates Table A11.

Input files:

```
$output/tables/TA12_Count_Turnover_Present.csv  
$Auxillary data/orbis/Raw data/General Extractions/MNE_Count.xlsx  
$Auxillary data/orbis/Raw data/EUTO_CbCR_data_2023-11-09.csv  
$Auxillary data/orbis/Intermediate data/MNEs_2015_2018.csv  
$Auxillary data/orbis/Intermediate data/MNEs_2015_2018_t500.csv  
$Auxillary data/orbis/Final data/COUNTRY_Subs_Scope_GUO.csv  
$Auxillary data/orbis/Final data/COUNTRY_Subs_t500_Scope_GUO.csv
```

Output files:

```
$output/tables/TA11_Count_Table.tex
```

## 7.3 Programs

### Programs installed via SSC

Our code utilizes the following commands installed via SSC manually in `0_master_cleaning.do`:

- **egenmore:**  
Nicholas J. Cox, 2000. “EGENMORE: Stata modules to extend the generate function,” Statistical Software Components S386401, Boston College Department of Economics.
  - **gtools:**  
Mauricio Caceres Bravo, 2018. “GTOOLS: Stata module to provide a fast implementation of common group commands,” Statistical Software Components S458514, Boston College Department of Economics, revised 03 Apr 2019.
  - **unique:**  
Tony Brady, 1998. “UNIQUE: Stata module to report number of unique values in variable(s),” Statistical Software Components S354201, Boston College Department of Economics, revised 18 Jun 2020.
  - **wid:**  
Blanchet, T. (2021). “WID: Stata module to download data from the World Inequality Database (WID. world)”.
  - **rscript**  
Molitor, D., & Reif, J. (2023). “RSCRIPT: Stata module to call an R script from Stata.”.
-

## Bibliography

**Bachas, Pierre Jean, Matthew Fisher-Post, Anders Jensen, and Gabriel Zucman**, “Capital Taxation, Development, and Globalization: Evidence from a Macro-Historical Database,” *NBER Working Paper*, 2022, (w29819).