Replication Code for "China in Tax Havens": README File

Christopher Clayton* Antonio Coppola† Amanda Dos Santos‡
Matteo Maggiori§ Jesse Schreger¶

May 2023

This README file describes the data required and the overall structure of the replication package for Clayton et al. (2023). The master executable file (`Master_Build.sh`), contained in the `'code'` folder, runs a pipeline comprising two parts. First, the code reads in a number of publicly and commercially available data and prepares them for analysis. Second, it uses the processed files to construct the exhibits in the paper.

1 Input Data

The code produces the results in the paper starting from a number of input data files. These enter the file structure in the folder `$ccdms1/raw`. Due to restrictions on sharing commercially available data, we are only able to include files that are publicly available. For the rest, we provide details on how they can be commercially acquired. The input files are as follows:

- **Restatement Matrices:** We use restatement matrices constructed as in Coppola et al. (2021). These files are publicly available at our website globalcapitalallocation.com, and therefore are also included as part of this replication package.
- **Country-Level Bilateral External Portfolios:** We use data on bilateral external portfolios at the country level constructed as in Coppola et al. (2021), based on publicly available data from U.S. Treasury International Capital (TIC) data and the IMF Coordinated Portfolio Investment Survey (CPIS), as well as commercially available holdings data. These files are publicly available at our website globalcapitalallocation.com, and therefore are also included as part of this replication package.
- **Holdings Data:** The code makes use of estimates based on the Morningstar holdings data for mutual funds and ETFs, as described in Maggiori et al. (2020) and Coppola et al. (2021),

*Yale School of Management; christopher.clayton@yale.edu.
†Stanford University, Graduate School of Business; acoppola@stanford.edu.
‡Columbia Business School; amanda.dossantos@gsb.columbia.edu.
§Stanford University, Graduate School of Business, NBER, and CEPR; maggiori@stanford.edu.
¶Columbia Business School, NBER, and CEPR; jesse.schreger@columbia.edu.

as well as based on data on U.S. insurance companies holdings from S&P Global Services, as described in Coppola et al. (2021). Please see those paper for details and code on the construction of the holdings samples. Since the Morningstar and S&P data are commercially available, they cannot be included in this replication package, however they can be obtained by contacting the respective data providers.

- **Issuance Data:** To compute amounts outstanding for the securities that we focus on in the analysis, we use issuance estimates constructed as in Coppola et al. (2021), based on data from Worldscope, Dealogic and Factset. Please see that paper for details and code on the construction of the issuance samples. Since the Worldscope, Dealogic and Factset data are commercially available, they cannot be included in this replication package, however they can be obtained by contacting the respective data providers.

2 File Structure and Technical Notes

The replication code is written to be executed on a high-performance computing (HPC) cluster equipped with the Slurm Workload Manager. The folder containing the replication code (`'code'`) contains the following Bash script files:

- `Master_Build.sh` is the primary executable script, which runs the replication from start to finish. It defines the compute parameters for each of the files and submits them as separate, parallel jobs in the HPC cluster. The file can be launched via command line as `bash Master_Build.sh`. The following options should be specified prior to running the file:
  - `<CODE_PATH>`: The path to the root folder containing the replication code.
  - `<DATA_PATH>`: The path to the root folder containing the replication data.
  - `<LOGS_PATH>`: The path where log files should be written.
  - `<SLURM_PARTITION>`: The name of the compute partition to be used by the Slurm scheduler for job execution.
- `Master_Controller.sh` is the file that is submitted to the Slurm scheduler for Stata jobs. Options passed to `Master_Controller.sh` define which child Stata do-file is then run for each job.
- `Master_Controller_R.sh` is the equivalent of `Master_Controller.sh` for running jobs in R.

The individual jobs are split into two types, which are executed in the following sequence:

1. **Reading and Cleaning Data.** These files contains jobs that read in the input files described above and prepare the data for analysis. These are:

- `MS_Data.do`: Produces the necessary filtering to the Morningstar holdings estimates.
- `Append_MS_Data.do`: Appends the filtered Morningstar estimates.
- `Firm_Level.do`: Constructs the firm-level restatements.

2. **Processing Data and Constructing Figures.**

   - `Issuance.do`: Produces Figures 1a and 1b.
   - `Holdings.do`: Produces Figures 1c and 1d.
   - `Currency.do`: Produces Figure 2.
   - `Sankey_Plot.R`: Produces Figure 3. The option `<DATA_PATH>` should also be specified in this file prior to code execution.

In addition to the main files described above, this replication package also includes the following auxiliary files:

- `Project_Globals.do`: Defines global variables used in the analysis, including the list of tax havens. The option `<DATA_PATH>` should also be specified in this file prior to code execution.
- `scheme-gcap_solid.scheme`: Defines the graphical scheme to be used by Stata for plots.

References

Clayton, Christopher, Antonio Coppola, Amanda Dos Santos, Matteo Maggiori, and Jesse Schreger, "China in Tax Havens," *AEA Papers and Proceedings*, January 2023.

Coppola, Antonio, Matteo Maggiori, Brent Neiman, and Jesse Schreger, "Redrawing the map of global capital flows: The role of cross-border financing and tax havens," *The Quarterly Journal of Economics*, 2021, *136* (3), 1499–1556.

Maggiori, Matteo, Brent Neiman, and Jesse Schreger, "International currencies and capital allocation," *Journal of Political Economy*, 2020, *128* (6), 2019–2066.