1. The Cross Country Mean Firm Size Dataset

Andrew Kerr[1] and Nic Forster[2]

November 2023

1.1. Introduction

This document explains the creation of the "Cross country mean firm size" dataset. This dataset was created as part of a project entitled "Firm size and Economic Growth in sub-Saharan Africa", which was funded by the Structural Transformation and Economic Growth (STEG) initiative. The dataset is created from several sources. These include IPUMS population census microdata, United Nations Statistical Division (UNSD) aggregated population census data, aggregated household survey data from the International Labour Organization (ILO), replication data from papers by Bento and Restuccia (2017) and Poschke (2018), microdata from the World Bank Enterprise Surveys (WBES) and GDP per capita information from the Penn World Table.

Data Creation

This document is separated into 9 short sections, each of which describes how a particular data source was used to create a final dataset. Each section corresponds to a Stata do file containing the code necessary to create the data from publicly available data sources.

1.2. Country Names and Codes

To enable the easy merging in of any data file in the Cross Country Mean Firm Size Dataset, a comprehensive dataset with all the names, ISO Alpha-3 codes and numerical codes of all countries and territories was obtained from the United Nations Statistics Division (UNSD, 2023). This dataset also contains regional indicators for every country. The only changes necessary were the additions of country names, codes and regions for Taiwan and Kosovo.

[1] School of Economics and DataFirst, University of Cape Town. andrew.kerr@uct.ac.za
[2] School of Economics, University of Cape Town

1.3. IPUMS

Following this, population census data from the Minnesota Population Centre's IPUMS microdata repository was downloaded. The size of the data files made the extraction a laborious exercise, as each country-year sample had to be downloaded individually. A list of each of the country-year samples and the labelling format can be found at the beginning of the relevant do file. 263 country-year samples were downloaded from IPUMS. This is a smaller sample than available on the IPUMS website due to the unavailability of some important variables in the rest of the data.

Once each country-year sample was downloaded, a loop was run through each of the country-year observations to calculate the mean firm size. To do this, the class of worker variables (classwk and classwkd) were used to identify the status in employment of each worker. A person was defined to be engaged if they were in category 2 – wage/salary worker, category 3 – unpaid worker, or category 4 – other. Own-account workers and employers were identified using the detailed version of the class of worker variable. Employers were identified using category 110 – Employer, while own-account workers were identified using category 100 – self-employed, category 120 – working on own-account and category 124 – own-account, other.

The formula below was then used to calculate the mean firm size in manufacturing. Several other measures were created using slight variations of the formula below. A version excluding own-account workers was calculated, as well as a version excluding public-sector workers.

$$mean\ firm\ size_1 = \frac{persons\ engaged + employers + own-account}{employers + own-account} \quad (1)$$

The data was then collapsed by the mean of the relevant variables to output a data file with the country, year, mean firm size, and the number of people in each status in employment category. Country names and codes were merged in from the data discussed in the preceding section.

1.4. United Nations Statistics Division (UNSD) Data

The next data source used was aggregated population census data obtained from the UNSD (2023). It includes totals in employment types by industry in each country-year observation. The data includes multiple observations for each status-country-year observation, as different revisions of the International Standard Industrial Classification of All Economic Activities (ISIC) are included. Only the most recent revision available in each country is used in the creation of the data. Categories for employers and own-account workers are included, while we use the employee and contributing family members categories to create a persons engaged category. Using the formula above, mean firm size measures including and excluding own-account workers are created. Totals for each of the employment categories are also included. These country-year observations were then appended to the IPUMS data.

We treat the IPUMS and UN data as one dataset in our analysis. In the paper, we refer to the IPUMS/UN estimate. Where an estimate existed for both datasets, the IPUMS estimate was given preference due to the availability of the micro-data. The UN data contains 139 country-year observations, 53 of which were already present in the IPUMS data. There are therefore 349 country-year observations when the two datasets are combined. The IPUMS/UN estimate variables are labelled “unip” in the dataset.[3]

1.5. Bento & Restuccia (2017)

The next dataset to be included is the mean firm size data created by Bento & Restuccia (2017).[4] The authors made this data publicly available, which therefore enabled comparisons across the datasets. Bento & Restuccia (2017) created a single country estimate for mean firm size between 2000 to 2012. To ensure ease of comparability, the data was merged to the IPUMS/UN year closest to 2006, the midpoint of Bento & Restuccia’s (2017) time range. However, a maximum difference between the IPUMS/UN year and Bento & Restuccia's (2017) year of 12 years was allowed.

[3] The mean firm size variable for the IPUMS/UN data including own-account workers is named ‘mfs_unip_oa’ in the dataset.
[4] This data can be found at:
https://www.openicpsr.org/openicpsr/project/116406/version/V1/view.

1.6. GEM in Poschke (2018)

A similar process was followed to merge in the mean firm size estimates created by Poschke (2018) using the Global Entrepreneurship Monitor (GEM).[5] Observations were merged to the closest IPUMS/UN year to 2002, the midpoint of the period that Poschke (2018) used to create a single country mean firm size estimate.

1.7. ILO

The fourth dataset used was obtained from the International Labour Organisation which was used by Salas-Fumás & Sanchez-Asin (2019). The Labour Force Statistics database[6] was used, which contains aggregated household survey data on labour-related outcomes. The dataset allows users to select two classifications by which to aggregate the data. Status in Employment and Economic Activity were selected. As in the UN data, the categories for own-account workers and employers are used straightforwardly in the mean firm size formula. Once again, to create a persons engaged estimate, family workers are added to employees. 1159 country-year observations were merged in, with only 69 observations matching exactly to an IPUMS/UN country-year observation.

1.8. PENN

Data for real GDP per capita was obtained from the Penn World Table V10.0.[7] The Purchasing Power Parity method of calculating real GDP across countries was used, which is labelled "rgdpo" in the Penn data.

[5] This data can be found at:
https://www.openicpsr.org/openicpsr/project/114100/version/V1/view

[6] This data can be found at: Download ILO data from: https://ilostat.ilo.org/data/. The following points are instructions for how to download:
* Select Database: Labour Force Statistics
* Select 1st Classification: Status in Employment
* Select 2nd Classification: Economic Activity
* Then download the Zipped CSV for: Employment by sex, status in employment and economic activity

[7] This data can be found at: https://www.rug.nl/ggdc/productivity/pwt/pwt-releases/pwt100

1.9. WBES

Lastly, the World Bank Enterprise Surveys[8] were used to calculate estimates for the number of firms in each country that would be unlikely to have any owner report themselves as an employer in a household survey or census. Firms with dispersed shareholder ownership, whether public or private, are unlikely to have any owner report themselves as an employer, as well as firms that are large partnerships. The process for identifying such firms is not perfect, and thus the principle of overestimating the number of firms is adopted, to ensure that the maximum amount of possible bias is accounted for. This is done for manufacturing firms only.

Using the available variable in the WBES dataset, a firm is said to have owners who would not report themselves as employers in a household survey or population census if:

1. It falls into the category of a "Shareholding company with shares traded in the stock market",

OR

2. It falls into the category of a "Shareholding company with shares traded privately",

AND

a. The establishment is part of a larger firm,

OR

b. The largest owner does not own the whole firm,

OR

c. More than 90% of the firm is owned by a foreign individual,

OR

d. The owner is not the top manager.

For the identification of partnerships, a firm is said to be a partnership if:

The largest owner does not own 100% of the firm

AND

The establishment falls into the category of a "Partnership" or a "Limited Partnership".

[8] This data can be found here: https://login.enterprisesurveys.org/content/sites/financeandprivatesector/en/library/combineddata.html.

From the WBES estimates dummy variables are created that indicate the closest IPUMS/UN and ILO country-year observation. These estimates are then added to the denominator of the mean firm size formula to produce an adjusted estimate of mean firm size. The adjusted IPUMS/UN and ILO variables are thus only created for the closest year to the WBES year.

1.10. Other Variable Creation

In all the IPUMS, UN and ILO datasets, the shares of each status in employment category as a proportion of total employment were calculated. A decade and 5-year average of all mean firm size and GDP per capita estimates were created, as well as a decade and 5-year dummy to ensure the use of only one of the averaged observations.

2. Reference List

Bento, P., and Restuccia, D. 2017. Replication data for: Misallocation, Establishment Size, and Productivity. *American Economic Journal: Macroeconomics*, 9(3), pp. 267–303. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor]. Available: https://doi.org/10.3886/E116406V1.

Feenstra, Robert C., Robert Inklaar and Marcel P. Timmer (2015), "The Next Generation of the Penn World Table" American Economic Review, 105(10), 3150-3182, available for download at www.ggdc.net/pwt

International Labour Organisation. 2023. Labour Force Statistics. Available: https://ilostat.ilo.org/data/.

Minnesota Population Center. Integrated Public Use Microdata Series, International: Version 7.3 [dataset]. Minneapolis, MN: IPUMS, 2020. https://doi.org/10.18128/D020.V7.3.

Poschke, Markus. Replication data for: The Firm Size Distribution across Countries and Skill-Biased Change in Entrepreneurial Technology. *American Economic Journal: Macroeconomics*,

10(3), pp. 1–41. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 2019-10-12. https://doi.org/10.3886/E114100V1

United Nations Statistics Division. Demographic Statistics Database: Employed population by status in employment, industry and sex. Available: http://data.un.org/Data.aspx?d=POP&f=tableCode:323.

United Nations Statistics Division. Standard country or area codes for statistical use (M49). Available: https://unstats.un.org/unsd/methodology/m49/overview. [Accessed: 31 January 2023].

World Bank. Enterprise Surveys. Available: https://login.enterprisesurveys.org/content/sites/financeandprivatesector/en/library/combineddata.html.