1. Introduction

This document describes how to replicate the results of "Mind the Data Gaps: An Examination Of Women-Owned Enterprise Representation". All of the sub-Saharan Africa data used in this paper is distributed by the World Bank and can be publicly accessed. Researchers interested in obtaining the datasets will need to both create an account on the World Bank Enterprise Survey portal, and apply for individual datasets on the World Bank Microdata Library. Further instructions on how to download these datasets is detailed later in the Data availability and Provenance Statement section. Except for the raw datasets, the replication material in this repository includes all necessary .do files to replicate the final dataset used for analysis and all the results. For completeness, this README describes the individual datasets as well.

The software used is STATA. All code was last run on Stata 17.0 SE. The total time required to run the code that creates the final dataset should not exceed 15 minutes.

2. Data availability and Provenance Statement

This section provides information on each data source, the purpose for which we use it, whether it is private or public, and gives data citation. The primary sources of data are World Bank Enterprise Surveys and the Living Standards Measurement Surveys. These are public data that can be accessed on websites hosted by the World Bank. Given that some key questions asked on most surveys were changed starting on the year 2006, this project only considers all World Bank surveys conducted in 2006 or later. Instructions on how to obtain the data are reported below.

1. World Bank Enterprise Survey (WBES Regular)

The World Bank Enterprise Survey is the most common and frequently collected
survey on enterprises by the World Bank that aims to generate a sample of enterprises representative of the whole non-agricultural private economy in each given country. Sampling protocol for the WBES Regular targets the formal private sector, and explicitly only includes enterprises with five or more employees. The sampling protocols specifically stratify on industry sector and geographic location; however, they do not explicitly mention or discuss gender. As of June 2022, there are a total of 85 WBES regular datasets in sub-Saharan Africa, and all of them are compiled into the final dataset.

In order to gain access to the World Bank enterprise Surveys, an account must be created in https://login.enterprisesurveys.org/content/sites/financeandprivatesector/en/signin.html

After doing so, datasets can be downloaded individually or in bulk. The approval for account creation can take up to 2 days. All the country and years that need to be downloaded from this source for this project can be found in Column 1 of Table 1 in the Appendix.

2. Micro-Enterprise Survey (WBES Micro)

The second type of data acquired from the "Enterprise Survey" efforts is a survey of micro enterprises, known as the Micro-Enterprise Survey. The WBES Micro survey targets registered establishments with less than five employees, with sampling techniques and questionnaires being the same as the WBES regular. For WBES micro, the regions sampled are selected based on the number of establishments, contribution to employment, and value added. In most cases these regions are metropolitan areas and reflect the largest centers of economic activity in a country. Similar to the WBES Regular, sampling protocols for the WBES Micro stratify on industry sector and geographic location but do not explicitly mention or discuss gender. As of June 2022, there are a total of 26 WBES micro datasets in Sub-Saharan Africa, and all of them are compiled into the final dataset.

The WBES micro can be accessed in the same portal as the WBES regular after creating an account: https://login.enterprisesurveys.org/content/sites/financeandprivatesector/en/signin.html

All the country and years that need to be downloaded from this source for this project can be found in Column 2 of Table 1 in the Appendix.

3. Informal Firm Survey (WBES Informal)

The third type of data used in the analysis is the Informal Firm Survey (IFS) collected and distributed as part of the overall WBES effort. These Informal Firm Surveys aim to provide information on a sample of informal private sector enterprises in selected urban centers within a country. The goal of the WBES Informal is to generate information about the reasons for informality. WBES informal surveys have employed two different sampling methodologies over time and are always implemented on urban areas. These two methodologies are further explained in the paper, but the core questions asked in the surveys remained relatively constant. As of June 2022, there are a total of 18 WBES informal datasets in Sub-Saharan Africa, and all of them are compiled into the final dataset.

The WBES informal can be accessed in the same portal as the WBES regular and micro after creating an account: https://login.enterprisesurveys.org/content/sites/financeandprivatesector/en/signin.html

All the country and years that need to be downloaded from this source for this project can be found in Column 3 of Table 1 in the Appendix.

4. Multi-Topic Household Surveys (HHS)

We use multi-topic household surveys which contain a survey module on non-farm enterprises. Many of these HHS are Living Standard Measurement Studies (LSMS), the flagship household survey program of the World Bank. These surveys aim to provide nationally representative estimates on household characteristics and outcomes from the country and year of survey. Sample size varies from 2,000 to 5,000 households depending on country and survey year. The samples are usually representative of the country as a whole and they are large enough to allow consideration of certain subgroups, such as rural vs. urban, or a few major agro-climatic zones.

We use all of the publicly available HHS data for Sub-Saharan Africa that is made available

by the World Bank, which totals up to 39 datasets. Many HHS include a survey module on non-agricultural enterprises that are owned by household members over the past 12 months. From this survey module, we create a representative sample of household enterprises in a country in that given year. It is important to note that this sample of enterprises may not be representative of the entire non-agricultural private economy for a particular country.

Each dataset has to be individually applied for or downloaded from the World Bank Data catalogue (note that for most of these you also need an account and to fill out a small application): https://microdata.worldbank.org/index.php/catalog/?page=1&collection%5B%5D=lsms&ps=15

All the country and years that need to be downloaded from this source for this project can be found in Column 4 of Table 1 in the Appendix.

5. Ghana Census

In 2014 the Ghana Statistical Service conducted the Integrated Business Establishment Survey (IBES), which is a non-household economic census covering all sectors of the economy). The census identified 638,000 establishments across all sectors, that had a physical enterprise structure and any household-based enterprise with a sign indicating its presence within a household. The census data excludes mobile businesses, trades in open spaces, trades in home if the shop not visible, retail shops selling on small tables under shades (e.g., market shades, stall without permanent occupants), and shrines without structure.

This dataset is also publicly accessible in the following website: https://www2.statsghana.gov.gh/statistics.html

3. Computation Requirements

The software used for all the code is STATA. All code was last run on Stata 17.0 SE. All code was run on 2 MacBook Air 2021, and a Lenovo Legion computer running on windows 11. We recommend to install the following packages to make sure all the code runs: mmerge, winsor2, labutil, and estout. Code to download all these packages is provided.

4. List of datasets used in the analysis

Name of Dataset	Source	Description	Notes
global_wbes_foranalysis	All WBES regular data combined into one dataset	Complete dataset since 2006	Might need extra cleaning that is later performed after combining
global_wbes_informal_foranalysis	All WBES informal data combined into one dataset	Complete dataset since 2006	Might need extra cleaning that is later performed after combining
global_wbes_micro_foranalysis	All WBES micro data	Complete dataset since 2006	Might need extra cleaning that is

	combined into one dataset		later performed after combining
global_lsms_foranalysis	All HHS data combined into one dataset	Complete dataset since 2006, only includes datasets with nonfarm enterprise module	Might need extra cleaning that is later performed after combining
global_combined_foranalysis	All data sources available on this project combined into one final dataset	Complete dataset that includes the WBES regular, WBES micro, WBES informal, and HHS	

5. Folder Structure

The following diagram shows the project folder structure. All the folders under Data Gaps Replication Package/Data/ should be empty after downloading

Data Gaps Replication Package/
|__Code/
| |__01_Create/
| | |__Country Specific Code/
| | |__Global/
| |__02_Analysis/
|__Data/
|__01_Raw/
|__02_Country/
|__03_Global/
|__04_Analysis/

6. Instructions for replicators

- The folder "Code" includes all relevant code necessary for replication.
- The "01_Create" folder under the Code folder includes files that clean and prepare all relevant datasets.
- The "Country Specific Code" folder under "01_Create" includes all .do files that prepare and compile the data source level datasets i.e., the country survey year level dataset.
- The "Global" folder under "01_Create" folder includes all .do files that merge and prepare the data source level datasets into one dataset that is utilized for analysis. This folder also includes the do file that compiles all exchange rate information for all 168 country year level data in our datasets.

To compile the final analysis dataset, the replicator needs to follow the following steps:

1. Download all WBES, WBES Micro, WBES Informal and HHS datasets from their appropriate sources specified in section 2 and save them in the folder Data/01_Raw.
    a. HHS datasets: make sure to keep the folder structure that the dataset is downloaded in. Often, when the HHS datasets are downloaded, they are contained in a folder with a specific name. Within each folder, there may be additional folders for different datasets. Keep the folder structure as the code used to import these datasets relies on the original folder structure.
2. Change the working directory
    a. Code that sets the working directory for this replication package is named "gpg_working_directory.do" and it is found in the folder "Code". Edit this working directory to reflect the path for the downloaded replication package on your computer.
3. Install all relevant STATA packages specified under the Computation Requirements Section.
4. To run all relevant code needed to produce the final analysis dataset, simply run the file "run_all.do" (Code/01_Create/Global). This file will run all relevant datasets including country year survey level .do files and global .do files that merge all datasets together. Namely the "run_all.do" calls two other .do files, namely "run_all_country_code.do" and "run_all_global_code.do" found in the same folder (Code/01_Create/Global).
    a. "run_all_country_code.do" file calls and runs all the data source level codes found in the folder "Country Specific Code" (Code/01_Create) that clean and prepare country survey year level datasets. The .do files in this dataset clean raw files from the Data/01_Raw folder, prepare them for merging and save the cleaned and prepared dataset in the data folder Data/02_Country.
    b. "run_all_global_code.do" file calls and runs all .do files in the "Global" folder (Code/01_Create). The .do files in the "Global" folder merge the country year survey datasets together and prepare them for analysis. All datasets, except the one, from this .do files are saved in the folder Data/03_Global. The exception to this is that the "prepare_global_combined_10_06_2022 .do" file that saves the final dataset called "global_combined_foranalysis.dta" in the folder "Data/04_Analysis".

    As mentioned earlier, running the code will take approximately 15 minutes.

Appendix

Table 1: List of Survey Years for Sub-Saharan African Countries

	WBES			HHS
Country	**Regular**	**Micro**	**Informal**	
Angola	2006, 2010	2006	2010	
Benin	2009, 2016			2018
Botswana	2006, 2010	2006	2010	
Burkina Faso	2009	2009	2009	2018
Burundi	2006, 2014	2006		
Cameroon	2006, 2009, 2016	2009	2006, 2009	
CAR	2011			
Chad	2009, 2018			
DCR	2006, 2010, 2013	2006, 2013	2010, 2013	
Republic of Congo	2009			
Cote d'Ivoire	2009, 2016	2009	2009	2018
Djibouti	2013			
Eritrea	2009			
Eswatini	2006, 2016	2006		
Ethiopia	2006, 2011, 2015	2011		2013, 2015, 2018
Gabon	2009			
Gambia	2006, 2018	2006		
Ghana	2007, 2013		2013	1992, 1998, 2005, 2013, 2017
Guinea	2006, 2016	2006		
Guinea-Bissau	2006	2006	2018	
Kenya	2007, 2013, 2018	2007, 2013	2013	
Lesotho	2009, 2016			
Liberia	2009, 2017			2018
Madagascar	2009, 2013	2009	2009	
Malawi	2009, 2014			2004, 2010, 2016, 2019
Mali	2007, 2010, 2016		2010	2018
Mauritania	2006, 2014	2006		
Mauritius	2009, 2020	2009	2009	
Mozambique	2007, 2018	2018	2018	
Namibia	2006, 2014	2006		
Niger	2009, 2017			2011, 2014, 2018
Nigeria	2007, 2010, 2014			2010, 2011, 2012, 2013, 2016, 2018, 2019
Rwanda	2006, 2011, 2019	2006, 2011	2011	
Senegal	2007, 2014			2018
Sierra-Leone	2009, 2017			
Somlia	2019	2019	2019	
South Africa	2007, 2020			
South Sudan	2014			
Tanzania	2006, 2013	2006		2008, 2010, 2012, 2014, 2019
Togo	2009, 2016	2009		2018
Uganda	2006, 2013	2006		2009, 2010, 2011, 2013, 2015, 2018, 2019
Zambia	2007, 2013, 2019	2019	2019	
Zimbabwe	2011, 2016	2016	2017	