Replication for "Disruptive Technologies and Finance: An Analysis of Digital Startups in Africa" by Marcio Cruz, Mariana Pereira-Lopez, and Edgar Salgado

Overview

The code in this replication package constructs the analysis file from three data sources (Crunchbase, 2023; Pitchbook, 2023; Bloom et al., 2021) using Stata and R (only for Figure 1). One Master file (Master.do) runs the different pieces of code (including the R script) to generate the data for the 6 main Figures and 5 tables included in the main paper and 6 Tables and 5 graphs from the Appendix. The replicator should expect the code to run for about 5 minutes.

Data Availability and Provenance Statements

Data Accessibility Statement

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. Appropriate permissions are documented in the *LICENSE.txt* file.

Data sources

There are four data sources used in this paper, two of them are used under license and therefore, cannot be made publicly available:

1) ***Crunchbase data:*** The use of Crunchbase data is based on Crunchbase's license agreement, which can be found at https://data.crunchbase.com/docs/license-agreement#enterprise-access-advanced
2) ***Pitchbook data***: The use of Pitchbook data is based on the terms of PitchBook's content license agreement, which can be found at https://pitchbook.com/content-license-agreement

   Important: Access to these two datasets can be purchased according to each company's license agreement.

The use of these two data sources is licensed according to the terms detailed in the file LICENSE.txt

3) Data from Bloom et al. (2021) listing all the bigrams identified as disruptive technologies.
4) ILO education mismatch indicators
5) World Development Indicators

Summary of Availability

■ **No data can be made** publicly available.

The data for this project under license agreements. Researchers interested in access to these data should refer to https://pitchbook.com/ and https://www.crunchbase.com/ for details regarding pricing, products, and solutions available.

Details on each Data Source

Table 1. Data sources

Data.Name	Data.Files	Location	Provided	Citation
"Crunchbase"	fintech_africa3.csv it_colombia.csv fintech_brazil_bottom.csv it_egypt.csv fintech_brazil_top.csv it_keghaug.csv fintech_london_bottom.csv it_london_bottom.csv ecommerce_africa1.csv fintech_london_top.csv it_london_top.csv ecommerce_africa2.csv fintech_mexico.csv it_mexico_bottom.csv ecommerce_africa3.csv fintech_paloalto.csv it_mexico_top.csv ecommerce_brazil_bottom.csv fintech_seattle.csv it_nigeria.csv ecommerce_brazil_top.csv fintech_southamerica1.csv it_paloalto.csv ecommerce_london_bottom.csv fintech_southamerica2.csv it_seattle_bottom.csv ecommerce_london_top.csv fintech_tokyo.csv it_seattle_top.csv	data/Crunchbase	FALSE	Crunchbase (2023)

	ecommerce_mexico.csv fintech_top1000.csv it_tokyo_bottom.csv ecommerce_paloalto.csv fintech_top1000rev.csv it_tokyo_top.csv ecommerce_seattle.csv it_top1000.csv ecommerce_southamerica1.csv it_top1000rev.csv ecommerce_southamerica2.csv it_africa3.csv it_zaf_bottom.csv ecommerce_tokyo.csv it_argentina.csv it_zaf_top.csv ecommerce_top1000.csv it_boecpapeur.csv people_ghana.csv ecommerce_top1000rev.csv it_brazil_bottom.csv wb.csv fintech_africa1.csv it_brazil_top.csv wdi_mobile.csv fintech_africa2.csv it_chile.csv			
“Pitchbook”	first_batch_mar_28_2023.csv second_batch_mar_30_2023.csv third_batch_may_18_2023.csv fourth_batch_may_30_2023.csv fifth_batch_may_30_2023.csv PitchBook_All_Columns_2023_04_28_08_03_33.csv PitchBook_All_Columns_2023_06_16_14_42_18.csv PitchBook_All_Columns_2023_06_20_10_04_46.csv	data/Pitchbook	FALSE	Pitchbook (2023)
“Bloom et al. (2021)”	DisruptiveTechnologies2.csv	data/Bloom	TRUE	Bloom et al. (2021)
“ILO (2023)	EMP_NORM_SEX_STE_EDU_NB_A.csv	data/other/ilo	TRUE	ILO (2023)

"Number of firms"	predicted.dta	data/other	TRUE	Castro e.t. al. (2023)
"World Development Indicators"	wdi.dta	Data/other	TRUE	World Bank (2023)

Table 2. Intermediate datafiles produced by code

Name	Data.Files	Location	Provided	Source code
"Roster"	roster.dta, rosterp.dta	working	FALSE	roster.do, rosterp.do
"Roster and Technology"	roster_technology.dta, roster_technology_list.dta, roster_technology_wide.dta, rosterp_technology.dta, rosterp_technology_list.dta, rosterp_technology_wide.dta,	working	FALSE	roster_technology.do, roster_technology_list.do, roster_technology_wide.do, rosterp_technology.do, rosterp_technology_list.do, rosterp_technology_wide.do, rosterp_technology_wide2005.do
"Mismatch"	mismatch_normative.dta	working	TRUE	mismatch_normative.do
"IV"	shareF.dta	working	FALSE	shareF.do
"29 technologies"	the29.dta, technology_year.dta	working	TRUE	the29.do
"Geocoded firms"	Geocoded_roster_for_geo_vf_cityccountry.dta	working	FALSE	Crunchbase (2023)

Description of folder contents

This folder structure of this document is as follows:

- **dofiles:** Holds the Stata code to produce all the different stages of the analysis

- **data:** Holds the raw datasets
- **working:** Will store the datasets produced for estimations and graphs.
- **tables:** Stores the tables produced by the code when we ran it. These match the tables in the paper.
- **figures:** Stores the figures produced by the code when we ran it. These match the figures in the paper.
- **tex:** Contains the LaTeX documents for compiling the results in the paper and annexes.

Computational requirements

Software requirements

Stata is used in the analysis of this project. This project was written using Stata 18 and has not been tested on older versions.

R is used on the background to create the map for Figure 1, but it is called from Stata. The version used to create the script was 4.2.1. (The R script to run the figure separately is geocoded_firms.R.

Memory and Runtime Requirements

Summary

Approximate time needed to reproduce the analyses on a standard (CURRENT YEAR) desktop machine: 5-10 minutes (if starting from the data provided. It could take around 60 minutes starting from the raw data-not provided for confidentiality).

Details

The code was last run on a **11th Gen Intel(R) Core(TM) i5-1145G7 @ 2.60GHz 1.50 GHz laptop with 16GB RAM and Windows 10 Enterprise. Computation took 5 minutes.**

Description of programs/code

-Program **roster.do** organizes the raw confidential data pulled from CrunchBase and prepares it to be used in regressions, tables and figures. At the moment of collecting data from CrunchBase we focused on 3 sectors: e-commerce, fintech and IT. We used CrunchBase filter for each sector in the keywords option. *Due to confidentiality, this do-file is included just as reference. The rest of code works based on its output.*

-Program **the29.do** prepares the bigrams associated to the 29 technologies described in Bloom et al (2021).

-Program **roster_technology.do** matches every firm from roster.do to any of the 29 technologies. The match is done using the text found in roster variables "industry", "description", and "fulldescription". The variable "match" counts how many of the 29 technologies were matched to the firm.

-Program **roster_technology_list.do, i**n addition to what roster_technology.do does, keeps the technology name as string, which will be input for other analysis.

-Program **roster_technology_list_wide.do** takes the output of roster_technology_list.do and reshapes the 29 technologies as columns.

-Program **rosterp.do** organizes the raw confidential data pulled from pitchbook and prepares it to be used in additional regressions. Like the download from crunchbase, we focused on 3 sectors: e-commerce, fintech and IT. We drop duplicates since Pitchbook regressions were not split by sector. *Due to confidentiality, this do-file is included just as reference. The rest of code works based on its output.*

-Programs **rosterp_technology.do, rosterp_technology_list.do, and rosterp_technology_list_wide.do,** match firms with the 29 technologies using the same approach detailed in the analogous do-files for Crunchbase data.

-Program **sharing.do** removes confidential information from roster.dta and rosterp.dta

-Program **shareF.do** calculates the share of disruptive technologies by industry in the frontier, that will be used as IV.

-Program **mismatch_normative.do** calculates the share of skill mismatch by country.

-Program **results.do** prepares figures and tables produced with Crunchbase data.

-Program **resultsp.do** prepares tables produced with pitchbook data.

-Program **results_iv.do** prepares tables for IV results produced with Crunchbase data.

-Program **results_iv_lac.do** prepares tables for IV results produced with Crunchbase data, for LAC.

The remaining dofiles are table and figure specific as shown in Table 2.

Instructions to Replicators

1. Open Stata
2. Open the Master do file (dofiles/Master.do)
3. Set the working directory (line 11 of code and the path of your Rscript.exe. The code runs even if you don't have R but does not generate the Map in Figure 1).
4. Run the Master do file in Stata (to create all the tables and figures). This is the only do-file you should need to run.

We install all the programs required to run the code at the beginning of the Master using ssc. If something pops up that you need to download, please download it.

Details

- dofiles/01_dataprep:
  - These programs were last run at various times between 2023 and 2024
- dofiles/02_analysis/Master.do.
  - If running programs individually, note that ORDER IS IMPORTANT.
  - The programs were last run top to bottom on February 27, 2024

List of tables 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.

Table 3. Index of code to reproduce tables and figures

Figure/Table # Object	Program	Line number	Output file	Notes
Table 1	results.do	79	Table1_funding_extensive.tex	
Table 2	results.do	110 140 188	Table2a_funding_extensiveIHS.tex Table2b_funding_extensiveIHS10.tex Table2c_funding_extensiveIHS11.tex	
Table 3	results.do	242 271 319	Table3a_funding_intensiveIHS.tex Table3b_funding_intensiveIHS10.tex Table3c_funding_intensiveIHS11.tex	
Table 4	results_iv.do	127 151 175	Table4a_funding_extensiveIV.tex Table4b_funding_extensiveIHSIV.tex Table4c_funding_intensiveIHSIV.tex	
Table 5	resultsp.do		Table5a_pitch_f.tex Table5b_pitch_a.tex	
Table A1				This table does not require computations. It comes from Bloom et al. (2021)
Table A2	results.do	115 164 212	TableA2a_funding_extensiveIHS.tex TableA2b_funding_extensiveIHS5.tex TableA2c_funding_extensiveIHS6.tex	

Figure/Table # Object	Program	Line number	Output file	Notes
Table A3	results.do	247 295	TableA3a_funding_intensiveIHS.tex TableA3b_funding_intensiveIHS5.tex TableA3c_funding_intensiveIHS6.tex	
Table A4	results_iv.do	276	TableA4_firststage.tex	
Table A5	results_iv.do	205 229 253	TableA5a_funding_extensiveIVi.tex TableA5b_funding_extensiveIHSIVi.tex TableA5c_funding_intensiveIHSIVi.tex	
Table A6	results_iv_lac.do	126 150 174	TableA6a_funding_extensiveIV_lac.tex TableA6b_funding_extensiveIHSIV_lac.tex TableA6c_funding_intensiveIHSIV_lac.tex	
Table A7	results_iv_lac.do	204 228 252	TableA7a_funding_extensiveIVi_lac.tex TableA7b_funding_extensiveIHSIVi_lac.tex TableA7c_funding_intensiveIHSIVi_lac.tex	
Figure 1	Rmap.do	6	Figure1_nfirm_geocoded_v2.png	In the R script (geocoded_firms) the map is generated in lines 66-90
Figure 2	figure2.do	34	Figure2_histogram.png	
Figure 3	results.do	402 441	Figure_3a_frontier_split.png Figure_3b_frontiernm_split.png	
Figure 4	results.do	463 480	Figure_4a_age_region_labor.png Figure_4b_age_region_labor_nm.png	
Figure 5	results.do	513 545	Figure_5a_age_region_sector.png Figure_5b_age_region_sector_nm.png	
Figure 6	results.do	621 675	Figure_6a_africa.png Figure_6b_africa_nm.png	
Figure B1	cities.do	36	FigureB1_Concentration_cities.png	
Figure B2	results.do	363	Figure_B2_frontier	
Figure B3	screenshot			Comes from external sources (Bloom et al., 2021, Crunchbase)
Figure B4	results_hetero.do		Figure B4_results_hetero.png	
Figure B5	it_detail.do		Figure_B5_textgap_ecommerce.png	
Figure B6	it_detail.do		Figure_B6_textgap_fintech.png	
Figure B7	it_detail.do		Figure_B7_textgap_it.png	

References

Bloom, N., Hassan, T. A., Kalyani, A., Lerner, J., & Tahoun, A. (2021). The diffusion of disruptive technologies (No. w28999). National Bureau of Economic Research.

Castro, L., M. Cruz, F. Molders, A. Volk, and E. Salgado. 2023. “Firm Demographics in Africa” (2023, unpublished manuscript).

Crunchbase (2023). Crunchbase. Retrieved from www.crunchbase.com

PitchBook (2023). PitchBook. Retrieved from https://pitchbook.com/

International Labour Organization (2023). Education and Mismatch Indicators database (EMI), https://www.ilo.org/ilostat-files/WEB_bulk_download/html/bulk_indicator.html look for indicator EMP_NORM_SEX_STE_EDU_NB_A.csv.gz