Replication Package

Helping Women Start Up: Understanding the Constraints Facing Women-Led Startups in India

Authors: Girija Borker, Luiza Cardoso, Ruchita Maghnani

Overview

This package reproduces the tables and figures in *Helping Women Start Up: Understanding the Constraints Facing Women-Led Startups in India*. The paper uses the Tracxn India Startup Database (downloaded July 2021) to document gender gaps in firm entry, founding-team composition, sector choice, networks, and access to funding among Indian high-growth entrepreneurs founded between 2000 and 2020.

The package contains the full Stata and R codebase used in the paper, together with intermediate ("precoded") datasets sufficient to reproduce all analysis outputs without access to the proprietary raw data. The cleaning, tidying, and construction pipeline is included for transparency but is disabled by default; replicators with Tracxn access can re-enable it via the switches at the top of MAIN.do.

Data Availability and Provenance

Summary

The analysis uses one underlying data source: the **Tracxn India Startup Database** (Tracxn Technologies Ltd.). Tracxn data are proprietary and **cannot be redistributed**. The package ships with intermediate datasets derived from Tracxn (under data/precoded/ and data/coded/) that are sufficient to reproduce all tables and figures in the paper. The raw Tracxn extract is not included.

Primary data source: Tracxn India Startup Database

Field	Value
Provider	Tracxn Technologies Ltd.
URL	https://tracxn.com/

Field	Value
Date downloaded	July 2021
Geographic coverage	India
Time coverage of analysis sample	Firms founded between 2000 and 2020
Initial extract	115,486 firms
Analysis sample after date filter	110,361 firms (firms founded 2000–2020; 3,852 firms with founding year 1533–1999 and 1,098 firms founded in 2021 are excluded)
Final analysis sample (gender identified)	55,074 firms (49.9 percent of the analysis sample)
License	Proprietary; redistribution prohibited
Access for replicators	Researchers wishing to reproduce the package from raw data must obtain a Tracxn license directly from the provider. Pre-processed intermediate files used by this package are available from the corresponding author on request (see *Contact*).

Tracxn collects firm information by scraping the internet to identify firms that are active online. Once a firm is identified, Tracxn analysts compile information from the company website, LinkedIn profile, and news media articles. The platform records, for each firm, attributes including sector, founding year, location (state and city), funding stage, total funding raised, valuation, employee count, revenue, and exit status, together with information on individuals associated with the firm (founders, C-suite, board members, angel investors).

Sample construction

The construction pipeline (code/clean/, code/tidy/, code/construct/) produces the analysis-ready datasets in data/precoded/ from the raw Tracxn extracts. The principal steps are:

1. **Date filter.** Firms with founding year before 2000 or after 2020 are dropped. The 2000 cutoff removes apparent data-entry outliers (some firms list founding years between 1533 and 1999); the 2020 cutoff ensures comparability across the most recent cohort, since newer firms have shorter online histories and are systematically under-represented in Tracxn.
2. **Cohort assignment.** Firms are assigned to one of four cohorts based on founding year:

Cohort	Years	Firms in analysis sample	Firms in final (gender-identified) sample
1	2000–2004	3,124 (2.8%)	1,703 (3.09%)
2	2005–2009	7,832 (7.1%)	4,376 (7.95%)
3	2010–2014	27,194 (24.6%)	17,360 (31.52%)
4	2015–2020	72,211 (65.4%)	31,635 (57.44%)

3. **Sector assignment.** Tracxn assigns each firm to one of 26 high-level sector categories (full list below). For Figure 3, these are further consolidated into the practice-area / sector mappings shown in data/precoded/sector-classification.dta.
4. **People-to-firm linking.** Tracxn records on people (founders, C-suite executives, staff, board members, angel investors) are matched to firms via company name and domain. The processed file data/coded/Intermediate/people-constructed.dta contains 183,954 person-firm-role records covering individuals associated with firms in the analysis sample.
5. **Gender identification.** Gender is assigned to individuals on the basis of first name, using a four-step procedure (see below). Of the 20,081 unique first names in the people data, gender was identified for the names corresponding to 97 percent of

person-records. Firms are classified as woman-founded if at least one founder is identified as a woman.

6. **Network construction.** Co-membership in firms (board members, angel investors, executives) is used to build a bipartite firm-investor / firm-board network, from which firm-level network statistics (d_connection_*, n_*_connections) are constructed in code/construct/construct-network.do. These statistics are stored in data/precoded/firms-connection.dta and data/precoded/network-contructed.dta.
7. **Funding panel.** Investment events are reshaped into a firm × round panel (data/coded/Intermediate/funding-rounds.dta) and merged with firm characteristics into data/coded/Intermediate/firms-panel-constructed.dta for the survival and exit analyses.

Gender identification procedure

The gender of each person in the data is identified by first name using the following four-step waterfall:

1. **Indian baby names dictionary.** The website https://indianhindunames.com/ was scraped to produce a dictionary of Indian first names with associated genders. Exact-match against this dictionary identifies 2,856 unique names, covering 64 percent of person-records. Code used code/precoded/scraping.R. It was scraped on Apr 29, 2020.
2. **Fuzzy matching.** A Python fuzzy-matching package is used to match additional spelling variants of names already in the dictionary, identifying 784 additional unique names, covering 4 percent of person-records.
3. **Manual tagging.** Remaining unmatched names are tagged manually by someone with knowledge of the context, identifying 1,180 additional unique names, covering 12 percent of person-records.
4. **Gender API.** Names still unmatched are passed through the gender-identification API at https://gender-api.com/, identifying 1,729 additional unique names, covering 17 percent of person-records. The API was used on December 2021.
5. **DTA format.** The final CSV file is converted to Stata (.dta) format using the code/precoded/construct.do do-file.

The resulting first-name → gender dictionary is shipped as data/precoded/names-gender.dta (20,081 unique names) and data/precoded/names-python.dta (824 fuzzy-matched names).

Note: The website https://indianhindunames.com/ is no longer available. However, the two scripts that rely on it are included in the reproducibility package.

Precoded dataset list

The following files in data/precoded/ are the inputs to the analysis pipeline (read directly by scripts in code/analysis/ and Main.R):

File	Observations	Variables	Size	Unit of analysis	Description
missing-gender.xlsx	16,614	3	327 KB	names	Names without gender classification.
network-contructed.dta	34,734	7	6.2 MB	firm-pair	Firm-firm connection edges used by network analysis scripts (note: filename is misspelled in source; preserved as-is to match script references).
names-gender.dta	20,081	2	0.4 MB	first name	First name → gender dictionary produced by the four-step identification procedure.
names-python.dta	824	2	0.0 MB	first name	Subset of names matched via Python fuzzy matching (step 2 of the

File	Observations	Variables	Size	Unit of analysis	Description
					identification procedure).

The full Tracxn variable codebooks (with variable types and value labels) are shipped under data/raw/codebooks/.

Dataset list

Data file	Source	Notes	Provided
data/raw/Excel/*	Tracxn India Startup Database	Tracxn export batches with company records, acquisitions, leaderboard metrics, and people sheets.	No
data/raw/codebooks/*	Tracxn India Startup Database	Codebooks for datasets.	No
data/raw/acquisition.dta	Tracxn India Startup Database	Acquisition-level dataset containing acquired companies, acquirers, sellers, transaction dates, acquisition types, stakes acquired, and transaction prices.	No
data/raw/angel.dta	Tracxn India Startup Database	Angel-investor dataset containing names, roles, associated companies, locations, contact information, profiles, and biographies.	No

Data file	Source	Notes	Provided
data/raw/board.dta	Tracxn India Startup Database	Board-member dataset containing names, roles, represented companies, board affiliations, director types, appointment status, and tenure dates.	No
data/raw/companies.dta	Tracxn India Startup Database	Master company-level dataset containing firm characteristics, location, founding year, sectors, funding, investors, founders, financial indicators, acquisitions, IPOs, employment, and Tracxn metrics.	No
data/raw/founders.dta	Tracxn India Startup Database	Founder-level dataset containing names, roles, companies, locations, contact details, LinkedIn profiles, biographies, and photographs.	No
data/raw/gender-api.xlsx	Created by authors	Gender API results used to assign gender to first names based on predicted gender and accuracy scores.	Yes
data/raw/investment.dta	Tracxn India Startup Database	Funding-round dataset containing round dates, types, amounts, recipient companies, investors, lead investors, facilitators, and related news.	No
data/raw/leaderboard.dta	Tracxn India Startup	Company ranking and performance dataset containing Tracxn scores, ratings, funding, social-media activity, downloads, reviews, news	No

Data file	Source	Notes	Provided
	Database	coverage, employment, revenue, and growth metrics.	
data/raw/other.dta	Tracxn India Startup Database	Dataset of other company team members containing names, positions, companies, locations, contact details, profiles, and biographies.	No
data/raw/positions-hierarchy.csv	Created by authors	Standardized list of job titles with their corresponding organizational hierarchy level and frequency.	Yes
data/raw/corrections/remove-people-duplicates.csv	Created by authors	Manual corrections used to remove duplicate people and replace incorrect biographies.	Yes
data/raw/corrections/standardize-location.csv	Created by authors	Crosswalk used to match nonstandard city and state names with official administrative areas.	Yes
data/raw/corrections/standardize-positions.csv	Created by authors	Manual rules used to harmonize spelling and variants of job titles.	Yes
data/raw/maps/india-osm.geojson	India Geodata	India Country Boundary. Accessed on July 2021. Source: https://yashveeeeeeer.github.io/india-geodata/	Yes

Variable definitions: cohorts, stages, exit, sectors

The categorical variables used throughout the analysis take the following values:

Cohort (cohort): 1 = 2000–2004 | 2 = 2005–2009 | 3 = 2010–2014 | 4 = 2015–2020.

Tracxn funding stage (stage): 1 Unfunded | 2 Seed | 3 Early Stage | 4 Funding Raised | 5 Late Stage | 6 Series A | 7 Series B | 8 Series C | 9 Series D | 10 Series E | 11 Series F | 12 Series G | 13 Series H | 14 Series J | 15 Public | 16 Acquired | 17 Acqui-Hired | 18 Deadpooled.

Simplified stage (stage_simp): 0 Pre-seed | 1 Seed | 2 Early traction | 3 Scaling stage | 4 Matured | 5 Deadpooled.

Exit status (exit): 0 Deadpooled | 1 IPO | 2 Acquired.

Sector (sector, 26 categories): 0 Other | 1 Aerospace, Maritime and Defense Tech | 2 Auto Tech | 3 Business Services | 4 Consumer | 5 Consumer Goods | 6 EdTech | 7 Energy Tech | 8 Enterprise Applications | 9 Enterprise Infrastructure | 10 Environment Tech | 11 FinTech | 12 Food and Agriculture | 13 Food and Agriculture Tech | 14 Gaming | 15 HealthTech | 16 Healthcare | 17 Industrial Goods and Manufacturing | 18 Life Sciences | 19 Media & Entertainment | 20 Real Estate and Construction Tech | 21 Retail | 22 Semiconductors | 23 Technology | 24 Telecom | 25 Travel and Hospitality Tech.

Confidentiality

The shipped files do not contain personal contact information (no email, phone number, or LinkedIn URL is included in precoded/). The first-name dictionaries (names-gender.dta, names-python.dta) contain isolated first names with no link to specific individuals. Firm-level identifiers (company name, Tracxn URL, web domain) are retained, as these refer to publicly listed business entities rather than to natural persons.

Folder layout

data/

├── raw/ # Placeholder folder. Raw Tracxn extracts go here if licensed.

│ # Includes GADM India shapefiles (gadm36_IND_*.rds)

│ # Codebook spreadsheets are at raw/codebooks/.

├── precoded/ # Analysis-ready intermediate datasets shipped with the package.

│ # Used directly by the analysis scripts.

└── coded/

├── Clean/ # Output of code/clean/ (regenerated only if cleaning=1)

├── Tidy/ # Output of code/tidy/ (regenerated only if tidying=1)

└── Intermediate/ # Output of code/construct/ (regenerated only if construction=1)

The precoded/ folder contains the minimum set of files required to run the analysis section of MAIN.do and Main.R end-to-end. See *Description of programs/code* for the role of each script.

Computational Requirements

Software

Software	Version	Notes
Stata	17 or higher (verified on 19.5)	The ado/ folder bundles **ietoolkit v6.4**, which is required because summary_table.do calls iebaltab with the pttest option (deprecated in ietoolkit v7+). MAIN.do automatically prepends ado/ to the adopath.
R	4.2.3	Verified on macOS via Homebrew GDAL.

Stata packages

The following SSC packages are required and can be installed by uncommenting the ssc install block at the top of MAIN.do:

reghdfe, ftools, estout, dataout, require, ietoolkit

R packages

```
install.packages(c(
  "here", "tidyverse", "sf", "scales",
  "readstata13", "raster", "rgdal",
  "ggraph", "tidygraph", "wordcloud2",
  "ggrepel", "ggalluvial", "pacman",
  "geodata"
))
```

rgdal and sf require a working GDAL/PROJ installation. On macOS, install via brew install gdal proj before installing the R packages.

Memory, runtime, and storage

Metric	Value
Approximate disk footprint (code + precoded data)	~360 MB
Approximate disk footprint (with raw Tracxn data)	~8.1 GB
Approximate runtime, MAIN.do analysis section only	~1 hour
Approximate runtime, Main.R end-to-end	5–8 minutes
Peak RAM observed	~6 GB

Verification environment

Component	Specification
Operating system	macOS 15.7.3 (Sequoia)
Processor	Apple M4
Memory	16 GB
Stata	19.5
R	4.2.3

The package is expected to run on Linux and Windows with the same Stata/R versions. Stata path separators in MAIN.do use forward slashes and are POSIX-compatible.

Description of Programs/Code

Top-level entry points

File	Role
MAIN.do	Master Stata script. Sets globals ($root, $code, $coded, $precoded, $raw, $output, $ado), prepends ado/ to the adopath, and runs the analysis pipeline. Contains four switches at the top: cleaning, tidying, construction, analysis.
Main.R	Master R script. Defines data_dir and out_dir, then calls construct-map.R, maps.R, and overleaf.R via here::here().

Folder layout

```
repo_package/
├── MAIN.do            # Stata master
├── Main.R             # R master
├── README.md          # This file
├── ado/               # Bundled ietoolkit v6.4
├── code/
│   ├── import/        # Reads raw Tracxn .csv into Stata (raw access required)
│   ├── clean/         # Variable typing, deduplication, basic recodes
│   ├── tidy/          # Reshapes and entity normalization
│   ├── construct/     # Builds firm/founder/network panels and the spatial dataset
│   └── analysis/      # Produces every .tex table and .png figure in outputs/
├── data/
│   ├── raw/           # Tracxn extracts and GADM shapefiles
│   ├── precoded/      # Analysis-ready datasets shipped with the package
│   └── coded/         # Output of clean/tidy/construct (regenerated on demand)
└── outputs/
    ├── tables/        # .tex output, one per regression table
    └── figures/       # .png output, one per figure
```

Pipeline switches in MAIN.do

local cleaning 0 *// Set to 1 only with raw Tracxn data*

local tidying 0 *// Set to 1 only with raw Tracxn data*

local construction 0 *// Set to 1 only with raw Tracxn data*

local analysis 1 *// Reproduces all tables and figures from precoded data*

By default, only the analysis section runs. The cleaning, tidying, and construction sections are included in the package for transparency and review but are not executed; they require the raw Tracxn extract.

Instructions to Replicators

1. **Obtain the package.** Clone or download this repository. The default location used during verification is ~/Downloads/repo_package/.
2. **Obtain the data.** Email the corresponding author (see *Contact* below) to request a Dropbox link to the data/ folder. Place its contents at repo_package/data/, preserving the subfolder structure documented above.
3. **Set the project root.** Open MAIN.do and update line 37 to the absolute path of your local copy of repo_package/:
4. global root "/your/path/to/repo_package"
5. **Install Stata packages.** Uncomment the ssc install block in MAIN.do (lines 50–55) on the first run, then re-comment it. Do *not* install ietoolkit from SSC if you intend to use the bundled v6.4 — the bundled version takes precedence via the adopath.
6. **Run MAIN.do** in Stata. With the default switches (analysis = 1), this produces every .tex file under outputs/tables/. A log is written to MAIN_log.log in the project root.
7. **Run Main.R** in RStudio. Update lines 24–25 to point to your local data and output folders:
8. data_dir <- "/your/path/to/repo_package/data/"
9. out_dir <- "/your/path/to/repo_package/outputs/"

Then run the script. This produces every .png file under outputs/figures/.

10. **Compare against the paper.** Tables in outputs/tables/ and figures in outputs/figures/ correspond one-to-one with the exhibits listed below.

List of Tables and Programs

All table outputs are written to outputs/tables/. The script paths below are relative to code/analysis/.

Exhibit	Title	Output file	Script	Status
Table 1	Definitions of women-owned enterprises	—	—	Definitional table; not generated by code
Table 2	Summary of differences between men-led and women-led firms	baltab.tex	summary_table.do	Reproduced
Table 3	Woman-founded firm entry by sector	entry_of_women_sector_level.tex	entry_of_women_sector_level.do	Reproduced
Table 4	Woman-founded firm entry	entry_of_women_with_growth.tex	entry_of_women_with_growth.do	Reproduced

Exhibit	Title	Output file	Script	Status
Table 5	Repeat founders have more access to funding	funding_serial.tex	funding_serial.do	Reproduced
Table 6	Women are not repeat founders	serial_founder.tex	serial_founder.do	Reproduced
Table 7	Women are more likely to have connections with angel investors active in their sector	network_descriptives.tex	network_descriptives.do	Reproduced
Table 8	Connections within sector increase probability of getting funding	network.tex	network.do	Reproduced

Exhibit	Title	Output file	Script	Status
Table 9	Women are less likely to be connected to board members and angel investors in their own firms	network_women.tex	network_women.do	Reproduced
Table 10	Total funding amount and network connections	network_funding_amount.tex	network_funding_amount.do	Reproduced
Table 11	Funding amount by round and network connections	network_funding_round.tex	network_funding_round.do	Reproduced
Table 12	Women-led startups and women's	team_any_women.tex	team_any_women.do	Reproduced

Exhibit	Title	Output file	Script	Status
	employment			
Table 13	Access to funding by gender of founder	funding.tex	funding.do	Reproduced
Table 14	Attending elite institutions and access to funding	funding_bio.tex	funding_bio.do	Reproduced
Table 15	Relationship to angel investors	funding_angel.tex	funding_angel.do	Reproduced
Table 16	Time to event	hazard.tex	survival-analysis.do	Reproduced
Table 17	Firm exit by gender	exit.tex	exit.do	Reproduced

List of Figures and Programs

All figure outputs are written to outputs/figures/. The script paths below are relative to code/analysis/ unless otherwise noted.

Exhibit	Title	Output file	Script	Status
Figure 1	Year of firm entry into the market and cohort classification	year_cohort.png	overleaf.R	Reproduced
Figure 2	Number of firms founded and gender of leadership by year	gender_by_year.png	overleaf.R	Reproduced
Figure 3	Sector consolidation (alluvial)	sector_classification.png	overleaf.R (data prepared by alluvial-plot-data.do)	Reproduced with manual label fix (see *Adjustments made* below)
Figure 4	Number of firms by sector and gender of leadership	gender_by_sector.png	overleaf.R	Reproduced
Figure 5	Percent of new firms entering main activity	sector_main.png	overleaf.R	Reproduced

Exhibit	Title	Output file	Script	Status
	centers by cohort			
Figure 6	Firm location	map.png	maps.R (data prepared by code/construct/construct-map.R)	Reproduced
Figure 7	Firm stage by gender of leadership	stages_v_gender.png	overleaf.R	Reproduced

References

- Tracxn Technologies Ltd. *Tracxn India Startup Database*. Bangalore, India. https://tracxn.com/
- Global Administrative Areas. *GADM Database of Global Administrative Areas, version 3.6: India (gadm36_IND_*.rds)*. https://gadm.org/
- DIME Analytics. *Reproducible Research: Best Practices for Data and Code Management*. World Bank, Washington, DC.

Contact

For data access requests, questions about the package, or to report reproduction issues, please contact:

Girija Borker (corresponding author) Development Impact Evaluation (DIME), World Bank
Email: gborker@worldbank.org