Introduction

This document lays out the process for replicating the paper Beraja, Yang, and Yuchtman (2022) "Data-intensive Innovation and the State: Evidence from AI Firms in China", including data and code requirements.

Structure of replication package

There are three top-level directories:

1. **Analysis:** contains code used to produce results
2. **Data:** contains raw and cleaned data
    - **Intermediate:** contains intermediate data produced by code
3. **Output:** the final tables and figures

Statistical software, packages, and code

All data cleaning and analysis were conducted in Stata, except for the Figure 2 map, which was produced in R.

Code has been tested on Stata 16. The following Stata packages need to be installed. Each can be installed by executing the command in Stata: “ssc install `x’”, where x is: carryforward, binscatter, estout, reghdfe, balancetable.

Code has been tested on R 4.0.0. The following R packages need to be installed. Each can be installed by executing the command in R: “install.packages('x')”, where x is: dplyr, ggplot2, sp, rgeos, scatterpie, maptools. Prior to running the code on R, GEOS and GDAL must also be installed. On OSX, this can be accomplished by executing `brew install geos` and `brew install gdal` through terminal (tested on OSX 12.3.1). On Linux, the command `sudo apt-get install libgeos-dev libgdal1-dev gdal-bin libproj-dev proj-data proj-bin` should work.

Running the replication files

To recreate the output files:

1. Modify the directory on line 15 of `Analysis/Analysis.do` to the replication folder and then run the file `Analysis/Analysis.do` on Stata
2. Modify the directory on line 8 of `Analysis/make_map.R` and then run the file `make_map.R` on R

Data

Data availability statement:

A summary of key variables and their sources are provided in Appendix Table A.1 in the manuscript. The following lists the types of data and their sources. All data below are available (in deidentified form) as part of this replication package:

- Software data (including counts by company and text data on software customer/function) comes from the Chinese Ministry of Industry and Information Technology. This data can be purchased from https://www.qcc.com/

- AI firm data comes from Tianyancha and Pitchbook. The data can be purchased from https://www.tianyancha.com/ and https://pitchbook.com/
- Firm capitalization and investment comes from Tianyancha. The data can be purchased from https://www.tianyancha.com/
- Demographic data (prefecture population, GDP, etc.) comes from Global Economic Data, Indicators, Charts & Forecasts (CEIC). The data can be purchased from https://www.ceicdata.com/en
- Contract information (including whether it is public security facial recognition, monetary value, surveillance camera capacity, and contract bidders) come from the Chinese Government Procurement Database. Data originally scraped from http://www.ccgp.gov.cn/
- Country level AI productivity data (investment, publications) come from Stanford HAI's AI Index Report. Data also available at https://hai.stanford.edu/research/ai-index-2022

Data file breakdown:

Data files include:

1. **ambiguous_public_security_agencies_firm_list.dta**: a list of companies that have first contracts that are ambiguously public security contracts
2. **baseline_data_04292020.dta**: this data is at the software/patent level and accordingly includes (much of the) firm-level and prefecture-level data. This dataset combines data from many sources. Contract information comes from the Chinese Government Procurement Database, firm data from Tianyancha and Pitchbook, software data from Chinese Ministry of Industry and Information Technology, and demographic data from the CEIC
3. **bidder_data.dta**: data at the bidder level for competitive contracts, containing information on the value of bids. Original data scraped from the Chinese Government Procurement Database
4. **cn_eng_pref_crosswalk.dta**: prefecture-level crosswalk between English/Chinese names
5. **company_basic_info.dta**: amount of firm capitalization by company. Data from Tianyancha.
6. **company_contract_dummy_20200202.dta**: contract status by company. Original data scraped from the Chinese Government Procurement Database
7. **contract_with_demographics.dta**: demographic information at the contract level. Data from the Chinese Government Procurement Database and CEI
8. **contracts_gdp_pop_admin-unit.dta**: contract level data, containing data on facial AI contracts. Original data scraped from the Chinese Government Procurement Database
9. **Elsevier - 2021 AI Index Reprot.xlsx**: country-year level data on AI publications. Data from Elsevier (collected by Stanford HAI)
10. **firm_characteristics.dta**: firm level data, containing information about 1st contracts earned by firms and firm age. Contract information from Chinese Government Procurement Database, firm data from Tianyancha and Pitchbook.
11. **Hardware_Capacities_20201020.dta**: prefecture-month level data on total # of surveillance cameras procured. Original data scraped from the Chinese Government Procurement Database
12. **map_prefecture_2015**: folder contains shapefiles for prefectures in China
13. **NetBase Quid - 2021 AI Index Report.xlsx**: data at the country level on amount of AI funding per year. Data from NetBase Quid (collected by Stanford HAI)
14. **pd_contracts_results_bert_sim_only.csv**: contract level data on text similarity. Original data scraped from the Chinese Government Procurement Database
15. **predict_customer_10_32_32.dta, predict_customer_20_16_32.dta, predict_customer_20_32_16.dta, predict_customer_thresh0.6.dta,**

predict_customer_thresh0.7.dta: software level data containing LSTM model predictions. Different files contain different model configuration/threshold parameters. Original software data from the Chinese Ministry of Industry and Information Technology

16. **prefecture_with_demographics.dta**: demographic information at the prefecture level. Data from the CEIC
17. **second_contract_data.dta**: a version of the baseline data that uses second contracts instead of first contracts. This dataset combines data from the same sources
18. **software_version_X.0.dta**: version numbers for software releases. Data from the Chinese Ministry of Industry and Information Technology
19. **time_series_contracts.dta**: number of contracts in China over time. Data from the Chinese Government Procurement Database

Data citations

- Daniel Zhang, Saurabh Mishra, Erik Brynjolfsson, John Etchemendy, Deep Ganguli, Barbara Grosz, Terah Lyons, James Manyika, Juan Carlos Niebles, Michael Sellitto, Yoav Shoham, Jack Clark, and Raymond Perrault, “The AI Index 2021 Annual Report,” AI Index Steering Committee, Human-Centered AI Institute, Stanford University, Stanford, CA, March 2021.
- Global Economic Data, Indicators, Charts & Forecasts (CEIC). "CEIC China Premium Database." Retrieved January 1, 2020.
- Hijmans, Robert, Nell Garcia, and John Wieczorek. "GADM: database of global administrative areas." Version 3.6 (released May 6, 2018).[Online] Retrieved March 12, 2019 from (2010).
- Ministry of Industry and Information Technology. "Company software registry." Retrieved January 1, 2020.
- Ministry of Finance. "Chinese Government Procurement Database." Retrieved January 1, 2020.
- Pitchbook. "Global companies." Retrieved January 1, 2020.
- Tianyancha. "Chinese corporate data." Retrieved January 1, 2020.