Introduction

This document lays out the process for replicating the paper Beraja, Kao, Yang, and Yuchtman (2023) "AI-tocracy", 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 1 and Figure A.5 map, which was produced in R.

Code has been tested on Stata 17. 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, xtevent.

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:

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/
- Unrest data comes from the GDELT project. The data is hosted at https://www.gdeltproject.org/
- Weather data comes from the World Meteorological Organization (WMO). The data is hosted at https://www.ncei.noaa.gov/data/global-summary-of-the-day/access/
- Police hiring data comes from OffCN Education Technology. The company appears to now be defunct. Data originally scraped from http://sd.offcn.com/
- Export data is based on "The Global Expansion of AI Surveillance" (Feldstein, 2019) and additionally included web-scraping and manual validation of numerous corporate websites and news reports. The original bibliography is accessible at https://www.zotero.org/groups/2347403/global_ai_surveillance/library.
- Politician career incentives come from "Career incentives of city leaders and urban spatial expansion in China." (Wang et al., 2020). Data is available at https://dataverse.harvard.edu/dataverse/restat

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. **china_weather_panel.dta**: data at the weather station-day level on daily weather conditions. Original data collected from the World Meteorological Organization.
4. **cityPS_history_allcities_by2015_finalrestat_temp.dta**: data at the prefecture-year level on local politician incentive
5. **cn_eng_pref_crosswalk.dta**: prefecture-level crosswalk between English/Chinese names
6. **contracts_gdp_pop_admin-unit.dta**: contract level data, containing data on facial AI contracts. Original data scraped from the Chinese Government Procurement Database
7. **export_regression.dta**: data at the company level, containing data on whether the company has begun exporting facial recognition AI abroad, merged with data on contracts and other company-level controls. Export data is an extension of Feldstein 2019
8. **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.
9. **firm_data.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, demographic data from the CEIC, unrest data from GDELT, and weather data from the WMO

10. **firms_matched_prefecture.dta**: prefecture-level crosswalk between English/Chinese names for prefectures with good covariate data
11. **GDELT_China_072920.dta**: data at the (unrest) event level, containing merged demographic data from the CEIC. Original data from GDELT
12. **GDELT_China_contemp_distance_111820.dta**: data at the prefecture-prefecture-quarter level, containing merged demographic data from the CEIC, unrest data from GDELT, and contract information comes from the Chinese Government Procurement Database
13. **Hardware_Capacities_20201020.dta**: prefecture-month level data on total # of surveillance cameras procured. Original data scraped from the Chinese Government Procurement Database
14. **map_prefecture_2015**: folder contains shapefiles for prefectures in China
15. **nonpolice_def.dta**: data at the contract level, containing information on non-public security contracts. Original data scraped from the Chinese Government Procurement Database
16. **pd_meta_screening_by_keywords.dta**: data at the software level, containing information on surveillance software. Original data scraped from the Chinese Government Procurement Database
17. **police_new_recruit.dta**: data at the prefecture-year level, containing information on new police hires. Data from OffCN Education
18. **police_office_share.dta**: data at the prefecture-year level, containing information on the composition of police hires. Data from OffCN Education
19. **predict_customer_10_32_32.dta, predict_customer_20_16_32.dta, predict_customer_20_32_16.dta**: software level data containing LSTM model predictions. Different files contain different model configuration parameters. Original software data from the Chinese Ministry of Industry and Information Technology
20. **prefec_station.dta, prefectureDist.dta**: crosswalk mapping prefectures in China to their closest weather station
21. **province_fiscal_expenditure.dta, province_fiscal_revenue.dta** data at the province-year level containing information about local government. Data from the CEIC
22. **software_version_X.0.dta**: version numbers for software releases. Data from the Chinese Ministry of Industry and Information Technology
23. **time_series_contracts.dta**: number of contracts in China over time. Data from the Chinese Government Procurement Database
24. **unrest_data.dta**: data at the prefecture-quarter level, containing information on number of unrest events (some instrumented by weather). Data from the GDELT project.

Data citations

- Feldstein, Steven. "The Global Expansion of AI Surveillance." Vol. 17. Washington, DC: Carnegie Endowment for International Peace, 2019.
- 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).
- Leetaru, Kalev, and Philip A. Schrodt. "GDELT: Global data on events, location, and tone, 1979–2012." ISA annual convention. Vol. 2. No. 4. Citeseer, 2013.
- Ministry of Industry and Information Technology. "Company software registry." Retrieved January 1, 2020.

- Ministry of Finance. "Chinese Government Procurement Database." Retrieved January 1, 2020.
- OffCN Education Technology. "Job postings." Retrieved January 1, 2020.
- Pitchbook. "Global companies." Retrieved January 1, 2020.
- Tianyancha. "Chinese corporate data." Retrieved January 1, 2020.
- Wang, Zhi, Qinghua Zhang, and Li-An Zhou. "Career incentives of city leaders and urban spatial expansion in China." Review of Economics and Statistics 102.5 (2020): 897-911.
- World Meteorological Organization. "Global summary of the day." Retrieved January 1, 2020.