

Reproducibility package for “The Impact of Technology on Migration to the United States from Central America and the Dominican Republic”

Overview

This package creates the output of the document “The Impact of Technology on Migration to the United States from Central America and the Dominican Republic” using Stata. The package consists of a series of do files and databases that allow the user to reproduce all the figures, tables and estimates that are part of the manuscript. All the material is stored in a shared zipped folder. To replicate these outputs, just run the code “main_dofile.do” after modifying the user path.

Data Availability

The data folder is divided into a “raw” data folder and a “cleaned” data folder. The latter contains the following databases that do not require any modification to produce the outputs:

- sample_1970_cleaned.dta: 1970 1% Census sample
- sample_1980_cleaned.dta: 1980 5% Census sample
- sample_1990_cleaned.dta: 1990 5% Census sample
- sample_2000_census_cleaned.dta: 2000 5% Census sample
- sample_2000_cleaned.dta-sample_2021_cleaned.dta: American Community Survey (ACS) – 1 year sample from 2000 to 2021.

Note: all these databases were downloaded from the IPUMS USA site in June 2023 and contain harmonized variables related to individual’s demographics, education, job, and migration status.

- Eloundou et al. (2023) exposure to GPTs measures. This database was shared by the authors through a request.
- un_migration_data.dta: United Nations data on international migrant stock between 1990 and 2020 by country of origin and destination.
- un_net_mig_rate_cadr_data_2020.dta: United Nations data on net migration rate for CADR countries in 2020.
- onet_acs_2010_Crosswalk.dta: crosswalk between the ONET-SOC 2010 and ACS-SOC 2010 classifications. This file was created by merging the two classifications, assigning a code to occupations that didn’t match (based on an analysis of the occupation descriptions) and eliminating duplicate occupations.
- occ2010_soc2010.dta: crosswalk between the IPUMS 2010 and ACS-SOC 2010 classifications. This file was created by keeping both variables from the 2010 ACS

harmonized by IPUMS, eliminating duplicate occupations, and adding four categories to the IPUMS 2010 classification that didn't have some correspondence with ACS-SOC classification (based on an analysis of the occupation descriptions).

- `acemoglu_autor_2011_measures.dta`: this is an adapted version of the routinization measures created by Acemoglu and Autor (2011). As the classification is based on the O*NET-SOC 2000 classification, we use the crosswalk to incorporate the O*NET-SOC 2010 codes and reweight the measures when multiple 2000 codes correspond to the same 2010 code, according to their respective shares of employment covered by that code.
- `non_matched_frey_osborne.dta`: this database stores the reweighted probability of computerization measures from Frey and Osborne (2017) for occupations which did not have a direct correspondence with the 2010 ACS-SOC classification.
- `eloundou_2023.dta`: this database contains an adapted version of the exposure to GPTs measures created by Eloundou et al. (2023). These measures are reweighted when multiple O*NET-SOC 2010 codes correspond to the same ACS-SOC 2010 code, according to their respective shares of employment covered by that code.
- `sample_2000_cleaned_cz.dta`: 2000 5% Census sample, disaggregated at commuting zone level, based on David Dorn's crosswalk.¹
- `sample_2010_cleaned_cz`: 2010 ACS 1-year sample, disaggregated at commuting zone level, based on David Dorn's crosswalk.
- `sample_2019_cleaned_cz.dta`: 2019 ACS 1-year sample, disaggregated at commuting zone level, based on David Dorn's crosswalk.
- `cw_czone_state.dta`: this crosswalk associates each commuting zone with a unique state code and eliminate those that span multiple states. The goal is to incorporate state controls in regressions at the commuting zone level.
- `panel_cadr_cz.dta`: panel data on the characteristics of CADR migrants at the commuting zone level from 2000 to 2019.
 - This database is broken into three separate databases, one for each year: `panel_2000_1.dta`, `panel_2010_1.dta/panel_2010_2.dta` and `panel_2019_2.dta`.
- `panel_non_cadr_cz dta`: panel data on the characteristics of non-CADR migrants at the commuting zone level from 2000 to 2019.
 - This database is broken into three separate databases, one for each year: `panel_2000_noncadr_1.dta`, `panel_2010_noncadr_1.dta/panel_2010_noncadr_2.dta` and `panel_2019_noncadr_2.dta`.
- `tot_ind_2000.dta-tot_ind_2019.dta`: total employment by industry code for the years 2000, 2010 and 2019.

¹ A copy of this file is stored as "sample_2000_census_cleaned_cz.dta"

- `tot_ind_aux_2000.dta-tot_ind_aux_2019.dta`: total employment by industry code and commuting zone for the years 2000, 2010, and 2019.
- `shrind_aux_2000.dta-shrind_aux_2019.dta`: share of employment by industry code in each commuting zone for the years 2000, 2010, and 2019.
- `shrind_2000.dta-shrind_2010.dta`: share of employment by industry code in each commuting zone for the years 2000 and 2010. The only difference from the previous databases is that these ones add the industry identifier.
 - These databases are broken down into individual databases for each of the 16 industry codes. The databases corresponding to the year 2000, for example, are called `shrind_ind1_2000`, `shrind_ind2_2000`, and so on.
- `tot_occ_2000.dta-tot_occ_2010.dta`: total employment by occupation code for the years 2000 and 2010.
- `shrocc_aux_2000.dta- shrocc_aux_2010.dta`: share of employment by occupation code in each commuting zone for the years 2000 and 2010.
- `shrocc_2000.dta- shrocc_2010.dta`: share of employment by occupation code in each commuting zone for the years 2000 and 2010. The only difference from the previous databases is that these ones add occupation identifiers.
 - These databases are broken down into individual databases for each of the occupation codes. The databases corresponding to the year 2000, for example, are called `shrocc_occ1_2000`, `shrocc_occ2_2000`, and so on.
- `ifr_usa.dta`: information on sector-level robot adoption from the International Federation of Robotics (IFR) from 1993 to 2020 for the United States. This database is broken down into three separate databases for the years 2000, 2010 and 2019 (`ifr_usa_2000.dta`, `ifr_usa_2010.dta`, and `ifr_usa_2019.dta`).
- `roboasts_usa_0010.dta/roboasts_usa_1019.dta`: combines data on the change in the stock of robots between 2000 and 2010 and between 2010 and 2019 with employment data by industry in the year 2000/2010.
- `bartik_2000.dta-bartik_2010.dta`: Bartik productivity shock by commuting zone for the years 2000 and 2010. The shock is calculated by multiplying the change in employment by industry and commuting zone by the share of that industry in the employment of a particular commuting zone in the first year.
- `panel_0010_cz.dta`: panel on the characteristics of CADR migrants and robot adoption by commuting zone from 2000 to 2010.
- `panel_1019_cz.dta`: panel on the characteristics of CADR migrants and robot adoption by commuting zone from 2010 to 2019.
- `panel_0010_noncadr_cz.dta`: panel on the characteristics of non-CADR migrants and robot adoption by commuting zone from 2000 to 2010.
- `panel_1019_noncadr_cz.dta`: panel on the characteristics of non-CADR migrants and robot adoption by commuting zone from 2010 to 2019.
- `panel_0010_cz_regs.dta`: final version of the panel database on the characteristics of CADR migrants and robot adoption by commuting zone from 2000 to 2010, used as input for the estimates.

- The folder includes a restricted version of this database containing variables related to the entire population, not just CADR migrants (panel_0010_1_controls.dta).
- panel_1019_cz_regs.dta: final version of the panel database on the characteristics of CADR migrants and robot adoption by commuting zone from 2010 to 2019, used as input for the estimates.
 - The folder includes a restricted version of this database containing variables related to the entire population, not just CADR migrants (panel_1019_1_controls.dta).
- panel_0010_1019_cz_regs.dta: appended version of the previous databases.
- panel_0010_noncadr_cz_regs.dta: final version of the panel database on the characteristics of non-CADR migrants and robot adoption by commuting zone from 2000 to 2010, used as input for the estimates.
- panel_1019_noncadr_cz_regs.dta: final version of the panel database on the characteristics of non-CADR migrants and robot adoption by commuting zone from 2010 to 2019, used as input for the estimates.
 - panel_0010_1019_noncadr_cz_regs.dta: appended version of the previous databases.

The “raw” data folder contains the following sub-folders:

- ACS: stores the raw versions of the ACS 1-year samples for 2000-2021 and the 1970, 1980, 1990 and 2000 census samples.
 - These databases were downloaded from the IPUMS USA website in June 2023. To download the data, please follow these steps:
 - **Apply for access:** submit your application for access to restricted microdata through this [link](#).
 - **Log in:** log in to the site using this [link](#).
 - **Select samples and data:** navigate to the GET DATA window. There, you can select harmonized variables and samples.
 - The samples to be selected for replication are the following: 1970 1% Census sample (form 1 state), 1980 5% Census sample (state), 1990 5% Census sample (state), 2000 5% Census sample, 2000 ACS, 2001 ACS, 2002 ACS, 2003 ACS, 2004 ACS, 2005 ACS, 2006 ACS, 2007 ACS, 2008 ACS, 2009 ACS, 2010 ACS, 2011 ACS, 2012 ACS, 2013 ACS, 2014 ACS, 2015 ACS, 2016 ACS, 2017 ACS, 2018 ACS, 2019 ACS, 2020 ACS, 2021 ACS.
 - The variables to be selected are the following: YEAR, SAMPLE, SERIAL, CBSERIAL, STATEICP, STATEFIP, COUNTYICP, PUMA, CITY, HHWT, CLUSTER, STRATA, GQ, PERNUM, CBPERNUM, PERWT, EXPWTP, SLWT, FAMUNIT, FAMSIZE, SUBFAM, SFTYPE, SFRELATE, CBSUBFAM, CBSFTYPE, CBSFRELATE, MOMLOC, MOMRULE, POPLOC, POPRULE, SPLOC, SPRULE, MOMLOC2, MOM2RULE, POPLOC2, POP2RULE, NCHILD, NCHLT5, NSIBS, ELDCH, YNGCH, RELATE, RELATED, SEX, AGE, BIRTHQTR, MARST, BIRTHYR, MARRNO, MARRINYR, YRMARR, DIVINYR, WIDINYR, FERTYR, RACE, RACED, HISPAN, HISPAND, BPL, BPLD, ANCESTR1, ANCESTR1D, ANCESTR2, ANCESTR2D,

CITIZEN, YRNATUR, YRIMMIG, YRSUSA1, YRSUSA2, LANGUAGE, LANGUED, SPEAKENG, TRIBE, TRIBED, RACHSING, PREDAL, PREDAPI, PREDBLK, PREDWHT, PREDHISP, RACAMIND, RACASIAN, RACBLK, RACPACIS, RACWHT, RACOTHER, RACNUM, HCOVANY, HCOVPRIV, HINSEMP, HINSPUR, HINSTRI, HCOVPUB, HINSCAID, HINSCARE, HINSVA, HINSIHS, SCHOOL, EDUC, EDUCD, GRADEATT, GRADEATTD, SCHLTYPE, EMPSTAT, EMPSTATD, LABFORCE, CLASSWKR, CLASSWKRD, OCC, OCC1950, OCC1990, OCC2010, OCCSOC, IND, IND1950, IND1990, INDNAICS, WKSWORK1, WKSWORK2, UHRSWORK, WRKLSTWK, ABSENT, LOOKING, AVAILBLE, WRKRECAL, WORKEDYR, INCTOT, FTOTINC, INCWAGE, INCBUS00, INCSS, INCWELFR, INCINVST, INCRETIR, INCSUPP, INCOTHER, INCEARN, POVERTY, MIGRATE1, MIGRATE1D, MIGPLAC1, MIGCOUNTY1, MIGMET1, MIGMET131, MIGMET13ERR, MIGTYPE1, MOVEDIN

- **Create a data extract:** to create a data extract, first choose the "View cart" option to see the selected variables and samples. Then, choose "Create data extract" to create a dataset. You can specify the *.dta* format to obtain a Stata database. IPUMS USA will notify you when the extract is ready for download.
- **Download the data:** navigate to the MY DATA window to check the status of the extract. If it is ready, download it.
- AA: stores the routinization measures by occupation created by Acemoglu and Autor (2011). The raw database is called "onet-soc.dta".
- BLS: stores the "national_M2010_dl.xls" file, which contains total employment information by occupation code for May 2010.
- DD: stores the crosswalk files provided by David Dorn on his website to match Census and Public Use Micro Areas (PUMA) to 1990 Commuting Zones.
- FO: stores the XLS files "frey_osborne_appendix", which contains the probability of computerization measure by occupation created by Frey and Osborne (2017). This measure was published in Appendix A of the paper, which is accessible via a ScienceDirect subscription (see the Data Sources section). This file contains two sheets:
 - Appendix A: Columns A-J store the raw information downloaded from Appendix A of Frey and Osborne (2017). Columns H-J store the manually added variables. The most important of these is the "Match" variable, which links each occupation listed in Appendix A to its corresponding code in the ACS-SOC 2010 classification. Since 55% of the occupations do not match any code, we manually assign a code based on the occupation's description.
 - non_matched_raw: store the same information as the first sheet but restricts the list to non-matched occupations and adds a variable called "2010 employment" which stores total employment by occupation code, taken from the "national_M2010_dl.xls" file. This information is used to reweight the probability measures when more than one occupation in Appendix A matches the same ACS-SOC code.
- IFR: stores the "table_industry.dta" database, which contains information on sector-level robot adoption from the International Federation of Robotics (IFR) from 1993 to

2020. The team used the World Bank's institutional subscription to the International Federation of Robotics *World Robotics 2021* to obtain this data.

- ONET: stores the crosswalks between different versions of O*NET occupation taxonomy based on SOC 2010. These files were combined in a single database called "2000_to_2010_Crosswalk.dta".
- TE: stores the following files related to the exposure to GPTs measures by Eloundou et al. (2023):
 - occ_level_abc (1).csv: this file contains exposure measures to GPTs by occupation code, as described by Eloundou et al. (2023)². This database is not publicly available and was shared by the authors upon request.
 - eloundou_soc_acs.xlsx: this file consists of three sheets:
 - eloundou_raw: this sheet contains the same information as the raw data provided by the authors, plus the ACS-SOC 2010 code and total employment by occupation in May 2010. See the Data Sources section for access to these two files (2010soc_acs.dta and national_M2010_dl.xls). This sheet was generated by the cleaning file.
 - eloundou_cleaned: this sheet contains the same information as the previous one, but it assigns a value of total employment to occupations that did not match any SOC code. This sheet was generated manually based on the analysis of the occupation description.
 - employment: this sheet contains total employment by occupation code, as presented in the "national_M2010_dl.xls" file.
- WB: stores the World Bank country classification by income level. The folder contains the following files:
 - OGHIST.xlsx: historical World Bank country classification by income level.
 - country_income_wb.dta: this database was manually compiled by combining IPUMS birthplace codes³ reported in censuses and surveys with World Bank Group country classifications by income level for 2023 (see the Data Sources section).
- UN: stores the public files downloaded from the site of the United Nations Population Division:
 - undesa_pd_2020_ims_stock_by_sex_destination_and_origin.xls: data on international migrant stock between 1990 and 2020 by country of origin and destination.
 - "WPP2022_GEN_F01_DEMOGRAPHIC_INDICATORS.xls": data on general demographic indicators by country from 1952 to 2021.

In addition to these subfolders, the raw data includes the followings files:

² Eloundou, T., Manning, S., Mishkin, P., and Rock, D. 2023. "GPTs Are GPTs: An Early Look at the Labor Market Impact Potential of Large Language Models," *arXiv preprint arXiv:2303.10130*.

³ The detailed codes can be found in this website: https://usa.ipums.org/usa-action/variables/BPL#codes_section

- 2010soc_acs.dta: this database stores the SOC classification used in the ACS, including the six-digit code, major group, and description. This database is also available as an XLS file (2010soc_acs.xlsx).
- onet_soc_2010_employment.dta: this database records total employment by occupation in May 2010 according to O*NET-SOC 2010 and ACS-SOC 2010 classifications.

All these data are publicly available except for those databases that contain information from the International Federation of Robotics and Eloundou et al. (2023) exposure to GPTs measures.

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.

Data Sources

File name	Source	URL	Access year
sample_1970_raw.dta, sample_1980_raw.dta, sample_1990_raw.dta, sample_2000_census_raw.dta, sample_2000_raw.dta- sample_2021_raw.dta	IPUMS USA	https://usa.ipums.org/usa-action/samples	2023
undesd_pd_2020_ims_stock_by_ sex_destination_and_origin.xls	United Nations Population Division	https://www.un.org/development/desa/pd/content/international-migrant-stock	2023
WPP2022_GEN_F01_ DEMOGRAPHIC_INDICATORS.xls	United Nations Population Division	https://population.un.org/wpp/assets/Excel%20Files/5_Archive/WPP2022-Excel-files.zip	2023
OGHIST.xlsx	World Bank Group Data*	https://datahelpdesk.worldbank.org/knowledgebase/articles/906519-world-bank-country-and-lending-groups	2023
cw_puma2000_czone.dta, cw_puma2010_czone.dta	David Dorn website**	https://www.ddorn.net/data.htm	2023
2000_to_2006_Crosswalk.xlsx		https://www.onetcenter.org/taxonomy/2006/walk.html	
2006_to_2009_Crosswalk.xlsx		https://www.onetcenter.org/taxonomy/2009/walk.html	
2009_to_2010_Crosswalk.xlsx	O*NET	https://www.onetcenter.org/taxonomy/2010/walk.html	2023
2010soc_acs.dta	US Census Bureau***	https://www2.census.gov/programs-surveys/acs/tech_docs/pums/data_dict/PUMSDict10.pdf	2023
onet-soc.dta	MIT Economics Data Archive	https://www.dropbox.com/scl/fi/86paxjkauqtg72o73gp2h/task-construct-1.zip?rlkey=o21usoigscnrki3r77oo610pg&e=1&dl=0	2023
frey_osborne_appendix.xlsx	ScienceDirect'	https://www.sciencedirect.com/science/article/abs/pii/S0040162516302244	2023
national_M2010_dl.xls	Bureau of Labor Statistics	https://www.bls.gov/oes/tables.htm	2023

*The World Bank: Dataset name: historical classification by income in XLSX format. License: <https://creativecommons.org/licenses/by/4.0/>

**Crosswalk files E4 and E5

***The database was built by downloading the categories of the variable SOCP (SOC Occupation code 2010) from the ACS 2010 dictionary.

Instructions for Replicators

- Update the main_dofile with your directory paths.
- Run the main_dofile in Stata.
- Check the outputs folder.
 - The migration10_1.xls and migration10_noncadr_1.xls files store the raw data produced by the regressions. The final figures are stored in fig22.xlsx and tab4.xlsx files, located in the same folders. To check the final outputs, open the raw and the cleaned files simultaneously.

List of Exhibits

The provided code reproduces all tables and figures in the paper.

Exhibit name	Output filename	Sheet	Script	Note
Figure 1	figures.xlsx	fig1	mainresults.do (line 6)	Found in outputs/main/figures
Figure 2	figures.xlsx	fig2	mainresults.do (line 30)	Found in outputs/main/figures
Figure 3	figures.xlsx	fig3	mainresults.do (line 43)	Found in outputs/main/figures
Figure 4	figures.xlsx	fig4	mainresults.do (line 93)	Found in outputs/main/figures
Figure 5	figures.xlsx	fig5	mainresults.do (line 93)	Found in outputs/main/figures
Figure 6	figures.xlsx	fig6	mainresults.do (line 140)	Found in outputs/main/figures
Figure 7	figures.xlsx	fig7	mainresults.do (line 174)	Found in outputs/main/figures
Figure 8	figures.xlsx	fig8	mainresults.do (line 215)	Found in outputs/main/figures
Figure 9	figures.xlsx	fig9	mainresults.do (line 215)	Found in outputs/main/figures
Figure 10	figures.xlsx	fig10	mainresults.do (line 269)	Found in outputs/main/figures
Figure 11	figures.xlsx	fig11	mainresults.do (line 269)	Found in outputs/main/figures
Figure 12	figures.xlsx	fig12	mainresults.do (line 341)	Found in outputs/main/figures
Figure 13	figures.xlsx	fig13	mainresults.do (line 404)	Found in outputs/main/figures
Figure 14	figures.xlsx	fig14	mainresults.do (line 404)	Found in outputs/main/figures
Figure 15	figures.xlsx	fig15	mainresults.do (line 452)	Found in outputs/main/figures
Figure 16	figures.xlsx	fig16	mainresults.do (line 452)	Found in outputs/main/figures
Figure 17	figures.xlsx	fig17	mainresults.do (line 642)	Found in outputs/main/figures
Figure 18	figures.xlsx	fig18	mainresults.do (line 726)	Found in outputs/main/figures
Figure 19	figures.xlsx	fig19	mainresults.do (line 808)	Found in outputs/main/figures
Figure 20	figures.xlsx	fig20	mainresults.do (line 892)	Found in outputs/main/figures
Figure 21	figures.xlsx	fig21	mainresults.do (line 949)	Found in outputs/main/figures
Figure 22	migration10_1.xls	Sheet1	regs_migration.do	Found in outputs/main/regs
Figure 23	figures.xlsx	fig23	mainresults.do (line 995)	Found in outputs/main/figures
Appendix Figure 1	figures.xlsx	fig1	appendix.do (line 6)	Found in outputs/annex/figures
Appendix Figure 2	figures.xlsx	fig2	appendix.do (line 97)	Found in outputs/annex/figures
Appendix Figure 3	figures.xlsx	fig3_4	appendix.do (line 175)	Found in outputs/annex/figures
Appendix Figure 4	figures.xlsx	fig3_4	appendix.do (line 175)	Found in outputs/annex/figures
Appendix Table 1	tables.xlsx	tab1	appendix.do (line 281)	Found in outputs/annex/tables
Appendix Table 2	tables.xlsx	tab2	appendix.do (line 281)	Found in outputs/annex/tables
Appendix Table 3	tables.xlsx	tab3	appendix.do (line 281)	Found in outputs/annex/tables
Appendix Table 4	migration10_noncadr_1	Sheet1	regs_migration_noncadr.do	Found in outputs/annex/regs

Note: figures and tables were exported from Stata and finalized manually in Excel. To preserve the format and labels of the figures and tables, do not erase the files located in the Output folder.

Requirements

The code was reproduced on a computer with the following specifications:

- OS: Windows 11 Pro
- Processor: 12th Gen Intel(R) Core(TM) i7-12700 2.10 GHz
- Memory available: 63.8 GB
- Software version: Stata version 17

Memory and Runtime and Storage Requirements

Implementing the reproducibility package does not require additional storage beyond the 89 GB used by the shared materials in the zipped folder. Running the code takes approximately one and a half hours.