README

Data and Code for "Buffer or Bottleneck? Generative AI, employment exposure and the digital divide in Latin America"

by Pawel Gmyrek, Hernán Winkler and Santiago Garganta (July 2024)

Overview

This document contains the necessary information to replicate the results in "Buffer or Bottleneck? Generative AI, employment exposure and the digital divide in Latin America".

The main folders included in the package, which must be located in the same folder for a correct replication of the results, are as follows:

- bases
- PIAAC
- dofiles
- xls
- Results
- Figures

Computational requirements

Software requirements include STATA (version 14 or higher), R and Microsoft Excel.

Auxiliar dofiles needed to run the STATA codes of this paper: "cuantiles.do" which is included in the "ado" subfolder, inside folder "dofiles" ("dofiles/ado").

Memory and runtime requirements

This replication can be run in a 2024 standard desktop computer, and the runtime to replicate all results is approximately 1 hour and 30 minutes.

General Instructions

To replicate results from the paper "Buffer or Bottleneck? Generative AI, employment exposure and the digital divide in Latin America", three main tasks must be carried out.

Task 0 concerns running the R script titled “Task0_Transform_ILO_raw_data.rmd”, which converts the raw data from three standard microdata files produced by ILO STATISTICS (stored it in folder “/bases/ILO/Microdata_ILO”) into one file “LATAM_FINAL_ILO_data.csv”, (stored in the same folder), that becomes the input to Task 1. These three raw ILOSTAT files, as well as the new file “LATAM_FINAL_ILO_data.csv” are provided for reproducibility verification. However, these raw files cannot be made public according to the ILO micro data policy and the bilateral agreements with countries providing the micro data to the ILO

In case external researchers need these raw files, they can contact the ILO author (gmyrek@ilo.org) and the ILO will provide the latest update of these files on a case-by-case basis. The R script for Task 0 also produces a file “LATAM_FINAL_ILO_totals_small.csv”, stored in the same folder, which can be made public, and which is used in parts of Task 2.

Task 1 concerns running Stata .do files to produce the aggregated results based on SEDLAC and other input data files. These results are saved in different sheets of one main Excel file, which will be stored in the folder “Results”. Please note that running the .do files will generate a new results file named “Results_AI_LAC.xlsx”, which overwrites the existing file. The results generated by the authors through Task 1 are also stored in the same folder “Results” in a file named “Results_AI_LAC_FINAL.xlsx”. This means that Task 2 can be executed directly from this final file, without the need of running Task 1. This is important, since some of the SEDLAC micro data files required for Task 1 are not publicly available, so not all users will be able to run Task 1 without obtaining the necessary access permissions. (Executing Task 2 from file “Results_AI_LAC_FINAL.xlsx” requires replacing all references to file “Results_AI_LAC.xlsx” in R Markdown script (Task2.Buffer_or_Bottlenect_ALL_PLOTS.rmd) with “Results_AI_LAC_FINAL.xlsx”).

Task 2 uses the aggregate results from Task 1 to construct all figures and most of the tables presented in the paper. It relies on one R Markdown document titled “Task2.Buffer_or_Bottlenect_ALL_PLOTS.rmd” and the main R project file “REPRODUCIBILITY package.Rproj”. To run the code, the user should open the project and the .rmd file and ensure that all the R libraries listed at the beginning of that code are installed on their computer. The Markdown file can be executed in one go by knitting to html, which will produce a file named “Task2.Buffer_or_Bottlenect_ALL_PLOTS.html” in the same main folder. This file will contain all figures and most tables from the main paper. In addition, running the code will save all figures in high resolution to “Figures” folder. (Please note that running the entire code will overwrite the current version of that .html file and all the figures in the “Figures” folder that have been created by the authors of this

package). The R code in Markdown file can also be executed in sections. Each section has been marked in a way that corresponds to figure names in the main paper, and running a given code chunk will reproduce the figure to which the code is mapped. To execute in chunks, the user needs to first run the code chunk named "load basic project data", as it contains some crucial elements used by subsequent code chunks. Please note that running a code chunk corresponding to a given figure will also overwrite the previously saved version of that figure in "Figures" folder.

Finally, Tables A1-3 of the Appendix are produced by running only the Stata files, as marked in the detailed description of Task 1 below.

Folders and Contents

The content of the main folders included in this paper´s reproducibility package is detailed as follows:

1. "bases":

The "bases" folder contains most of the datasets used for this paper. It is organized by the following subfolders:

- Subfolder "sedlac": includes all sedlac databases used in this paper. We used additional subfolders to store each country database separately. The subfolders and the corresponding databases (.dta) included are:
    - \ARG\ARG_2021_EPHC-S2_v01_M_v02_A_SEDLAC-03_all
    - \BOL\BOL_2021_EH_v01_M_v02_A_SEDLAC-03_all
    - \BRA\BRA_2021_PNADC-E5_v01_M_v02_A_SEDLAC-03_all
    - \CHL\CHL_2022_CASEN_v01_M_v02_A_SEDLAC-03_all
    - \COL\COL_2019_GEIH_v01_M_v02_A_SEDLAC-03_all
    - \CRI\CRI_2019_ENAHO_v02_M_v02_A_SEDLAC-03_all
    - \DOM\DOM_2021_ECNFT_v01_M_v01_A_SEDLAC-03_all
    - \ECU\ECU_2021_ENEMDU_v01_M_v01_A_SEDLAC-03_all
    - \GTM\GTM_2014_ENCOVI_v02_M_v02_A_SEDLAC-03_all_ocupa
    - \HND\HND_2019_EPHPM_v01_M_v02_A_SEDLAC-03_all
    - \MEX\MEX_2018_ENIGHNS_v02_M_v02_A_SEDLAC-03_all
    - \NIC\NIC_2014_EMNV_v02_M_v02_A_SEDLAC-03_all
    - \PAN\PAN_2018_EH_v01_M_v05_A_SEDLAC-03_all
    - \PER\PER_2021_ENAHO_v01_M_v01_A_SEDLAC-03_all
    - \SLV\SLV_2021_EHPM_v01_M_v01_A_SEDLAC-03_all
    - \URY\URY_2021_ECH-S2_v01_M_v01_A_SEDLAC-03_all

The subfolder "sedlac" contains all restricted-use SEDLAC databases included in this paper and sourced from SEDLAC harmonization project (World Bank and CEDLAS). The original data source is the World Bank Stata API *datalibweb*, for internal use only. The SEDLAC databases listed above can be used for reproducibility verification at the WB but cannot be shared publicly on the portal. Please note that not all these SEDLAC databases are publicly available, and hence cannot be shared to external researchers without specific prior written permission.

- Subfolder "ILO": includes all data extracted from the databases of the International Labour Organization (ILO) used in this paper:
    - "Gmyrek_Berg_Bescond_scores_2023.csv": AI exposure scores by 4-digit ISCO08 occupations from "Gmyrek, P., Berg, J., Bescond, D., 2023. Generative AI and Jobs: A global analysis of potential effects on job quantity and quality (Working paper). ILO." The scores can be dynamically visualised and downloaded at this online portal: https://pgmyrek.shinyapps.io/AI_Data_Portal_Research/

    - Sub-folder "./bases/ILO/Microdata_ILO/" contains raw data extracted from ILO Microdata in 3 files by ILO STATISTICS, in which number of jobs in automation/augmentation/unknown has been tagged at 2-digit ISCO-08 level:
        - EMP_SEX_ISCO08_2D_Augmentation.csv
        - EMP_SEX_ISCO08_2D_Automation.csv
        - EMP_SEX_ISCO08_2D_Bigunknown.csv

      All three files were extracted by ILO STATISTICS on 23.12.2023.

      NOTE: This file cannot be made public according to the ILO micro data policy and the bilateral agreements with countries providing the micro data to the ILO.

    - Task 0 converts these three files into one file that becomes the input to Task 1:
        - LATAM_FINAL_ILO_data.csv

      NOTE: This file should not be made publicly available on the WB portal, because it contains data in a format that cannot be made public according to the ILO micro data policy and the bilateral agreements with countries providing the micro data to the ILO.

    - File "LATAM_FINAL_ILO_totals_small.csv", stored in the same folder, contains a small version of country-level total exposures.

NOTE: This file can be shared on the WB portal.

- The file "EAR_OCU2Digits_LAC.xlsx", provided as an extract from ILO micro data by ILO STATISTICS on 17.04.2024, contains data on mean and median income at 2-difgit ISCO-08 level. This file is the basis of Figure 14.
  NOTE: This file should not be made publicly available on the WB portal, because it contains data in a format that cannot be made public according to the ILO micro data policy and the bilateral agreements with countries providing the micro data to the ILO.

- The R code in markdown file (Task 2) under Figure 14 produces a file "EAR_OCU2Digits_LAC_Processed.csv", which filters out small observations and tags the ones with low reliability.
- **NOTE: This file can be shared on the WB portal**

- These raw ILO files listed above can be used for reproducibility verification at the WB but cannot be shared publicly on the portal. In case external researchers need these raw files, they can contact the ILO (gmyrek@ilo.org) and the ILO will provide the latest update of these files on a case-by-case basis.

- The main folder "./bases/ILO" contains a set of additional files, which are used in different parts of the code as per specific descriptions under each figure and in R Markdown code for task 2. These files can be shared on the WB portal.

- Subfolder "PPP_Poverty_SEDLAC": includes database "ipc_sedlac_wb.dta" from SEDLAC project. This database contains PPP conversion factors and CPIs (Consumer Price Indexes) for each LAC country, which enable to estimate poverty in SEDLAC database using international poverty lines.

  The database "ipc_sedlac_wb.dta" is also sourced from SEDLAC harmonization project (World Bank and CEDLAS), for WB internal use only. It can be used for reproducibility verification at the WB but cannot be shared publicly on the portal. Please note this database is not publicly available, and hence cannot be shared to external researchers without specific prior written permission.

- Subfolder “Processed”: contains all auxiliar databases generated using the databases described in the previous subfolders (subfolders “sedlac”, “ILO”, and “PPP_Poverty_SEDLAC”):
    - “matchISCO_ILOWB.dta”
    - “matchISCO_hnd_SEDLAC.dta”
    - “matchISCO_slv_SEDLAC.dta”
    - “matchISCO_ury_SEDLAC.dta”
    - “ILO_db.dta”
    - “AI_regress_LAC.dta”
    - “ipc_sedlac_wb.dta”

    See description of folder “dofiles” below for a more detailed explanation of the content of each database included in subfolder “Processed”.

2. “PIAAC”:

The “PIAAC” folder contains all information regarding PIAAC survey. It is organized by the following subfolders:

- Subfolder “bases”: contains PIAAC survey database named “piaac_all.dta”.
- Subfolder “xls”: includes an excel file called “cpu_int_work_regress.xls”, which contains the results of the estimated coefficients of computer use at work (and internet use at work) using PIAAC data. This excel file is generated by executing dofile “ai_sedlac_piaac.do” from the master dofile “master_AI_LAC.do” (see explanation of each dofile below).
- Subfolder “docs”: includes PIAAC survey documentation, such as questionnaires and reports (pdf files).

    PIAAC survey databases were accessed in August 2023 and are publicly available at https://www.oecd.org/en/about/programmes/piaac.html.

3. “dofiles”:
   The “dofiles” folder contains all the stata codes needed for this replication. This folder contains the following dofiles:
   - **master_AI_LAC.do**: this master dofile contains the execution of all dofiles regarding task 1 (see General Instructions section included above). The content of each dofile included and executed in this master dofile is explained as follows.

- **Internet_vs_gdp.do**: this dofile invokes WB data on GDP per capita (USD 2017 PPP) and internet users (ITU) to contruct Figure 3 of the paper (for more details of Figure 3 see the "Specific Instructions" section below).
- **prepara_base.do**: it contains all codes to get ready most of the databases used in this paper. Specifically:
    - prepare database with AI exposure scores at the 4-digit ISCO-08 occupations estimated by ILO (Gmyrek, Berg, Bescond, 2023). The input database is "Gmyrek_Berg_Bescond_scores_2023.csv" (in folder "bases\ILO") and the final processed database is named "ILO_db.dta" (saved in folder "bases\Processed").
    - prepare database to estimate poverty using international poverty lines (for SEDLAC databases). The input database is "ipc_sedlac_wb.dta", which is included in folder "bases\PPP_Poverty_SEDLAC", and the final processed database is saved with the same name "ipc_sedlac_wb.dta" in folder "bases\Processed".
    - prepare ILO & WB database to match AI exposure scores, by both 2-digit ISCO-08 occupations and gender, with some SEDLAC countries with 2-digit ISCO-08 information (BRA, COL, CRI and MEX). The input database is "LATAM_FINAL_ILO_data.csv" (in folder "bases\ILO\Microdata_ILO") and the final processed database is named "matchISCO_ILOWB.dta" (saved in folder "bases\Processed").
    - prepare SEDLAC databases with 4-digit ISCO-08 information (CHL, DOM, ECU, HND, PAN, PER, SLV and URY), and matched with "ILO_db.dta" to get AI exposure scores by both 2-digit ISCO-08 occupations and gender. This AI exposure scores imputation for SEDLAC countries with 4-digit ISCO-08 information is then used to estimate AI exposure in other similar SEDLAC countries with only 2-digit ISCO-08 information: ARG, BOL, GTM and NIC. In particular, the imputation for URY is used to estimate AI exposure in ARG, the imputation for SLV is used to estimate AI exposure in BOL, and the imputation for HND is used to estimate AI exposure in GTM and NIC. For this specific process, the input databases are the SEDLAC databases with 4-digit ISCO-08 information, particularly URY, SLV and HND (in folder "bases\sedlac"), and the final processed databases are named "matchISCO_URY_SEDLAC.dta", "matchISCO_SLV_SEDLAC.dta" and "matchISCO_HND_SEDLAC.dta", respectively (saved in folder "bases\Processed").

- **est_isco08_4d.do**: this dofile contains the imputation of ILO AI exposure scores (in proccessed database "ILO_db.dta") to each SEDLAC country database with 4-digit ISCO-08 information (CHL, DOM, ECU, HND, PAN, PER, SLV, URY). This imputation enables to estimate then the potential AI exposure in these countries: share of total workers exposed to Augmentation, Automation, The Big Unknown or Missing. It also enables to estimate AI exposure profiles in each country: by gender, geographic area (urban-rural), age groups, education, poverty (lp 6.85), income quintiles, legal and productive informality, labor relationship, economic activity sector. This dofile also construct part of the database used to estimate the pooled OLS regression with all individual observations (see Table A3 of the Appendix): "AI_regress_LAC.dta" (saved in folder "bases\Processed").
- **est_isco08_2d_mchSEDLAC.do**: this dofile contains the imputation of the AI exposure scores to some SEDLAC countries with only 2-digit ISCO-08 information (ARG, BOL, GTM, NIC). For the AI exposure imputation in these cases, we use the estimated AI exposure coefficients at the 2-digit ISCO-08 occupational level disaggregated by gender, coming from similar SEDLAC countries with 4-digit ISCO-08 information (for more details see the explanation above of the dofile "prepara_base.do"). This imputation enables to estimate the potential AI exposure in these countries (ARG, BOL, GTM, NIC): share of total workers exposed to Augmentation, Automation, The Big Unknown or Missing. It also enables to estimate AI exposure profiles in each country: by gender, geographic area (urban-rural), age groups, education, poverty (lp 6.85), income quintiles, legal and productive informality, labor relationship, economic activity sector. This dofile also construct part of the database used to estimate the pooled OLS regression with all individual observations (see Table A3 of the Appendix): "AI_regress_LAC.dta" (saved in folder "bases\Processed").
- **est_isco08_2d_mchILOWB.do**: this dofile contains the imputation of the AI exposure scores to some SEDLAC countries databases with 2-digit ISCO-08 information (BRA, COL, CRI, MEX). For the AI exposure imputation in these cases, we use for each country the corresponding ILO & WB AI exposure coefficients at the 2-digit ISCO-08 occupational level, disaggregated by gender (database "matchISCO_ILOWB.dta", for more details see the explanation above of the dofile "prepara_base.do"). This imputation enables to estimate the potential AI exposure in these countries (BRA, COL, CRI, MEX): share of total workers exposed to Augmentation, Automation, The Big Unknown or Missing. It also enables to estimate AI exposure profiles in each country: by gender, geographic area (urban-rural), age groups, education, poverty (lp 6.85), income quintiles, legal and productive informality, labor relationship, economic activity sector. This dofile also

construct part of the database used to estimate the pooled OLS regression with all individual observations (see Table A3 of the Appendix): "AI_regress_LAC.dta" (saved in folder "bases\Processed").

- **est_all.do:** this dofile merges all SEDLAC countries (16) AI exposure profiles (estimated by "est_isco08_4d.do", "est_isco08_2d_mchSEDLAC.do", and "est_isco08_2d_mchILOWB.do") and exports the results to an xls file called "isco_ai_all.xls" (saved in the "xls" folder). The content of this file must be pasted in the main Excel file of results called "Results_AI_LAC.xlsx", particularly in the worksheet named "AI_LAC_DATA". Once this is done, the final table of AI exposure profiles will be completed in the worksheet "AI_LAC", which is the input information to get "Figure 5", "Figure 8" and "Figure 9" of the paper. See the specific section of the R Markdown file (Task 2) for the details of how the figure is produced from this data.
- **ai_piaac_only.do**: this dofile contains the imputation of ILO GenAI exposure scores at the 4-digit ISCO-08 to the microdata from the Programme for the International Assessment of Adult Competencies (PIAAC) collected by the OECD. The PIAAC survey contains information for several countries on detailed tasks carried out by people at work, such as whether workers use a computer (and internet) at work. Thus, this dofile also considers this information to split each group of GenAI exposure into those who use a computer at work, and those who do not.
- **est_piaac.do:** this dofile merges all PIAAC countries AI exposure profiles (estimated by "ai_piaac_only.do") and exports the final results, which corresponds to an xls file called "ai_piaac_only_cpu.xls" (saved in the "xls" folder). The content of this file must be pasted in the main Excel file of results called "Results_AI_LAC.xlsx", particularly in the worksheet named "AI_PIAAC_ONLY_CPU_DATA". Once this is done, the final table of AI exposure profiles based only on PIAAC data will be completed in the worksheet "AI_PIAAC_ONLY_CPU", which is the input information to get "Figure 10" of the paper. See the specific section of the R Markdown file (Task 2) for the details of how the figure is produced from this data.
- **ai_sedlac_piaac.do**: it includes the imputation of the predicted probabilities (on the use of computer at work, and on the use of internet at work) based on PIAAC data to SEDLAC database and combine them with the imputation of the GenAI exposure scores from ILO (using the same methodology explained above: see description of dofiles "est_isco08_4d.do", "est_isco08_2d_mchSEDLAC.do" and "est_isco08_2d_mchILOWB.do"). This dofile also exports the results of the estimated coefficients of computer use at work (and internet use at work) from PIAAC to an xls file called "cpu_int_work_regress.xls" (saved in the

"xls" subfolder of "PIAAC" folder). See particularly lines 229-248 of this dofile for a more detailed presentation of the model. The content of the xls file "cpu_int_work_regress.xls" must be pasted in the main Excel file of results called "Results_AI_LAC.xlsx", particularly in the worksheet named "Regress_PIAAC_CPUINT_DATA". Once this is done, the Table A2 of the paper (Appendix) will be completed in the worksheet "Regress_PIAAC_CPUINT" of the Excel file of results "Results_AI_LAC.xlsx".

- **est_sedlac_piaac.do:** this dofile merges all SEDLAC countries estimations (processed in "ai_sedlac_piaac.do") and exports the final results to an xls file called "ai_sedlac_piaac_comp.xls" and "ai_sedlac_piaac_int.xls" (saved in the "xls" folder). The content of these files must be pasted in the main Excel file of results called "Results_AI_LAC.xlsx", particularly in the worksheets named "AI_SEDLAC_PIAAC_CPU_DATA" and "AI_SEDLAC_PIAAC_CPUINT_DATA", respectively. Once this is done, the final table of AI exposure combined with the use of computer and internet at work profiles (based on PIAAC and SEDLAC data) will be completed in the worksheets "AI_SEDLAC_PIAAC_CPU" and "AI_SEDLAC_PIAAC_CPUINT". The former worksheet is the input information to replicate "Figure 11" and "Figure 12" of the paper, while both worksheets are used to construct Figure A5 in the Appendix. See the specific section of the R Markdown file (Task 2) for the details of how these figures are produced from this data.
- **ai_sedlac_piaac_isco2d.do**: this dofile includes the imputation of the predicted probabilities (on the use of computer at work) based on PIAAC data to SEDLAC database and combine them with the imputation of the GenAI exposure scores from ILO (using the same methodology explained above: see description of dofiles "est_isco08_4d.do", "est_isco08_2d_mchSEDLAC.do" and "est_isco08_2d_mchILOWB.do"). Finally, it estimates for each country the combined measures of AI exposure and use of computer at work, by 2-digit ISCO-08 occupations.
- **est_sedlac_piaac_isco2d.do:** this dofile merges all SEDLAC countries estimations (processed in "ai_sedlac_piaac_isco2d.do") and exports the final results to an xls file called "ai_comp_byisco.xls" (saved in the "xls" folder). The content of this file must be pasted in the main Excel file of results called "Results_AI_LAC.xlsx", particularly in the worksheet named "AI_CPU_ISCO2d_DATA". Once this is done, the final table of AI exposure combined with the use of computer at work (based on PIAAC and SEDLAC data), by 2-digit ISCO-08, will be completed in the worksheet "AI_CPU_ISCO2d", which is the input information to get "Figure 13" and "Figure 14" of the paper. See the specific section of the R Markdown file (Task 2) for the details of how the figure is produced from this data.

- **ai_sedlac_compuHH.do**: this dofile imputes the ILO GenAI exposure scores to SEDLAC databases (using the same methodology explained above: see description of dofiles "est_isco08_4d.do", "est_isco08_2d_mchSEDLAC.do" and "est_isco08_2d_mchILOWB.do") and combine with SEDLAC information regarding computer at home.
- **est_sedlac_compuHH.do:** it merges all SEDLAC countries estimations (processed in "ai_sedlac_compuHH.do") and exports the final results to an xls file called "ai_sedlac_compHH.xls" (saved in the "xls" folder). The content of this file must be pasted in the main Excel file of results called "Results_AI_LAC.xlsx", particularly in the worksheet named "AI_COMPUHH_DATA". Once this is done, the final table of AI exposure combined with computer at home profiles (based on SEDLAC data) will be completed in the worksheet "AI_COMPUHH", which is one of the input data to get "Figure A5" in the Appendix of the paper (see also "est_sedlac_piaac.do"). See the specific section of the R Markdown file (Task 2) for the details of how the figure is produced from this data.
- **chk_obs.do:** this dofile counts total observations and observations used for AI exposure estimations for each SEDLAC database. Then, it exports the final result to an xls file called "ai_sedlac_obs.xls" (saved in the "xls" folder). The content of this file must be pasted in the main Excel file of results called "Results_AI_LAC.xlsx", particularly in the worksheet named "Obs_SEDLAC_AI_DATA". Once this is done, Table A1 of the paper (Appendix) will be completed in the worksheet "Obs_SEDLAC_AI".
- **reg_ai_LAC.do:** estimate results of the pooled OLS regression with all individual observations (AI exposure as the dependent variable), with country-level normalized population weights using the database "AI_regress_LAC.dta", which is created in dofiles "est_isco08_4d.do", "est_isco08_2d_mchSEDLAC.do", and "est_isco08_2d_mchILOWB.do" (see explanation of each dofile above). The results of the regressions for each AI exposure category are saved in an xls file called "regress_pool.xls" (saved in the "xls" folder). The content of this file must be pasted in the main Excel file of results called "Results_AI_LAC.xlsx", particularly in the worksheet named "REGRESS_AI_DATA" (from xls cell D4). The weighted mean values of the dependent variables are estimated in this dofile and saved in an xls file called "meandep.xls" (saved in the "xls" folder). The content of this file must be pasted in the same worksheet named "REGRESS_AI_DATA" (from xls cell I4). Once this is done, Table A3 of the paper (Appendix) will be completed in the worksheet "REGRESS_AI".

4. "xls":

This folder contains files that store intermediate results to construct some of the final figures and tables included in this paper. All files included in this folder are exported from running some of the STATA codes previously explained. For more details of the files included in this folder, see the description included above in folder "dofiles".

5. "Results":

This folder includes two main xlsx files storing intermediate and final results of the paper: "Results_AI_LAC.xlsx" and "Results_AI_LAC_FINAL.xlsx". For more details on the content and use of these files, see the description included above in section "General Instructions" and the explanations of folder "dofiles".

6. "Figures":

The "Figures" folder contains all Figures of the paper. For more specific details of each of the figures included in this paper, see the following section "Specific Instructions", considering also the description in section "General Instructions" and (in some cases) previous explanations in folder "dofiles".

Specific Instructions:

(NOTE: For each of the Figures, the user needs to first execute the chunk of Markdown with libraries and 'load basic project data')

Figure 1: The assigned part of the R script contacts the WB WDI database, downloads population and income data for 2022 and plots those. The data is extracted from WDI API in a dynamic way, but since it selects year 2022 as a reference, the results should be stable for other users. The plots used in the paper were produced with data accessed on 10 July 2024.

Figure 2: The assigned part of the R script relies on an extract of results from the already published paper of Gmyrek, Berg and Bescond (2023), available in the file "./bases/ILO/GBB_subregion_sex.csv", generated by Gmyrek (gmyrek@ilo.org) on 5 November 2023. It plots subregional results from that paper into two graphic objects that are then saved as one plot of augmentation and automation by subregion.

Figure 3: The assigned part of the R script contacts the WB WDI database, download the data on per capita income and internet use and produces two sub-plots: one for the world and one for LAC countries in the sample. These are subsequently joined into one graphic object. The data is extracted from WDI API in a dynamic way, but since it selects years 2019-2021 as a reference, the results should be stable for other users. The plots used in the paper were produced with data accessed on 10 July 2024.

Figure 4: The assigned part of the R script relies on the download from Rilostat library in the first chunk of the code titled "load basic project data". It uses ILO statistics on 1-digit occupations for LAC, calculates the means grouped by occupation, sex and income and plots them.

Figure 5: The dataset for this figure is constructed using input data from worksheet "AI_LAC" of the Excel file "Results_AI_LAC.xlsx" (using data from column "Total"). These results are obtained by executing dofiles "est_isco08_4d.do", "est_isco08_2d_mchSEDLAC.do" and "est_isco08_2d_mchILOWB.do" and "est_all.do" from the master dofile "master_AI_LAC.do" (see explanation of each dofile above). The comparison in Figure 5 is made to the calculations provided by the ILO based on its micro data in the file "./bases/ILO/Microdata_ILO/LATAM_FINAL_ILO_totals_small.csv" (see above for how the file is prepared). The assigned part of the R script loads both datasets, merges them, assigns WB labels and plots the comparison.

Figure 6: The figure is based on the file "bases/ILO/Clustering_data.xlsx", which contains data on ISCO-08 2-digit shares, populations and GDP, manually prepared for the LAC countries in horizonal format based on the columns B, C and M of the sheet "AI_CPU_ISCO2d" of the main results file "Results_AI_LAC.xlsx" (which is obtained by executing dofiles "**ai_sedlac_piaac_isco2d.do"** and "**est_sedlac_piaac_isco2d.do"** from the master dofile "**master_AI_LAC.do"** - see explanation of each dofile above) and based on manual extraction of population and income data from the WB data repository. The assigned part of the R script loads that dataset and conducts hierarchical clustering procedure (Ward), subsequently plotting and saving the result.

Table 1: Descriptive table, no data needed for replication.

Figure 7: The assigned part of the R script loads three datasets: "./bases/ILO/Microdata_ILO/LATAM_FINAL_ILO_totals_small.csv" (see above how the file is produced based on ILO micro data) and ./Results/Results_AI_LAC_FINAL.xlsx (produced by Task 1) and ./bases/ILO/GBB_2023_HIC_results_formatted.csv (which contains selected results for HICs from the already published paper by Gmyrek, Berg and Bescond (2023), generated for the purpose of this paper by Gmyrek (gmyrek@ilo.org) on 1

July 2024). The script formats, merges and plots the comparison on the basis of these three files.

Figure 8: The dataset for this figure is constructed with the same input data as Figure 5 (worksheet "AI_LAC" of the Excel file "Results_AI_LAC.xlsx"), but using the included heterogeneities by several socio-economic characteristics, and focusing only on Automation exposure. The assigned part of the R script (Task 2) provides details of how the variables are formatted to produce the final figure.

Figure 9: The dataset for this figure is constructed with the same input data as Figure 5 (worksheet "AI_LAC" of the Excel file "Results_AI_LAC.xlsx"), but using the included heterogeneities by several socio-economic characteristics, and focusing only on Augmentation exposure. The assigned part of the R script (Task 2) provides details of how the variables are formatted to produce the final figure.

Figure 10: The dataset for this figure is constructed using input data from worksheet "AI_PIAAC_ONLY_CPU" of the Excel file "Results_AI_LAC.xlsx" (only for the four countries in the LAC region - Chile, Ecuador, Mexico and Peru-, and 2 developed economies -Slovenia and New Zealand- included in the PIAAC dataset). These results are obtained by executing dofiles "**ai_piaac_only.do"** and "**est_piaac.do"** from the master dofile "**master_AI_LAC.do"** (see explanation of each dofile above). The assigned part of the R script (Task 2) adds a manual mapping of OECD member states, PIACC dates and countries, full names etc, which are added to the data frame to produce the plot.

Figure 11: The dataset for this figure is constructed using input data from worksheet "AI_SEDLAC_PIAAC_CPU" of the Excel file "Results_AI_LAC.xlsx" (using data from columns "female" and "male"). These results are obtained by executing dofiles "**ai_sedlac_piaac.do"** and "**est_sedlac_piaac.do"** from the master dofile "**master_AI_LAC.do"** (see explanation of each dofile above). See the specific section of the R Markdown file (Task 2) for the details of how the data is formatted to produce the figure from this data.

Figure 12: The dataset for this figure is constructed using input data from worksheet "AI_SEDLAC_PIAAC_CPU" of the Excel file "Results_AI_LAC.xlsx" (using data from Colombia and Costa Rica). These results are obtained by executing dofiles "**ai_sedlac_piaac.do"** and "**est_sedlac_piaac.do"** from the master dofile "**master_AI_LAC.do"** (see explanation of each dofile above). See the specific section of the R Markdown file (Task 2) for the details of how the figure is produced from this data, by focussing only on two selected countries.

Figure 13: The dataset for this figure is constructed using input data from worksheet "AI_CPU_ISCO2d" of the Excel file "Results_AI_LAC.xlsx". These results are obtained by executing dofiles "**ai_sedlac_piaac_isco2d.do"** and

"**est_sedlac_piaac_isco2d.do"** from the master dofile "**master_AI_LAC.do"** (see explanation of each dofile above). See the specific section of the R Markdown file (Task 2) for the details of how the data is formatted to produce the figure is produced from this data in a heatmap format.

Figure 14: The dataset for this figure is constructed using input data from three different files:

1. Raw data from ILO/STATS with employment related income for LAC countries at 2-digit level, local currency (./bases/ILO/Microdata_ILO/EAR_OCU2Digits_LAC.xlsx) – see description of ILO folder above for more details.

   NOTE: The raw microdata file "EAR_OCU2Digits_LAC.xlsx" can be used for reproducibility verification but should not be shared publicly. Instead, the R code in markdown file (Task 2) under Figure 14 produces a file "EAR_OCU2Digits_LAC_Processed.csv" and saves it in the same folder "./bases/ILO/Microdata_ILO/". This data can be shared with users. Please see the description of folders above for more information.

2. Data from the final results, with exposure type and computer use by ISCO 2-digit and 2-digit shares of employment by country ("./Results/Results_AI_LAC_FINAL.xlsx", sheet = "AI_CPU_ISCO2d").
3. Data with official mapping of ISCO labels at different digit levels (./bases/ILO/mapping.csv). **(This file can be shared on the WB portal)**

See the specific section of the R Markdown file (Task 2) for the details of calculations and how the figure is produced from this data.

APPENDIX:

Figure A1: This figure relies on the dataset of TechXPosure by Prytkova et al., 2024 (./bases/ILO/techXposure_isco4d_by_tech.csv) downloaded from the paper's GitHub repository on 7 February 2024 (https://github.com/FabienPetitEconomics/TechXposure/tree/main/_isco) and the results file of the already published paper by Gmyrek, Berg and Bescond (2023) (./bases/ILO/Gmyrek_Berg_Bescond_scores_2023.csv") (see description of Fig 2) . The assigned part of the R script (Task 2) provides details of how the variables are formatted to produce the final figure.

Figure A2: This figure relies on the dataset of IMF provided by Cazzaninga et al., 2024 ("./bases/ILO/IMF_AIOE_CAIOE_theta_for_sharing.xlsx) and the results file of the already published paper by Gmyrek, Berg and Bescond (2023)

(./bases/ILO/Gmyrek_Berg_Bescond_scores_2023.csv"). The assigned part of the R script (Task 2) provides details of how the variables are formatted to produce the final figure.

Figure A3: This figure is based on 2-digit calculations provided by the ILO based on its micro data files, available in "./bases/ILO/Microdata_ILO/ LATAM_FINAL_ILO_data.csv" (see in description of Task 0 above and the ILO folder for how this file is produced). **Please note that this file should not be made publicly available on the WB portal, because it contains data in a format that cannot be made public according to the ILO micro data policy and the bilateral agreements with countries providing the micro data to the ILO.** The assigned part of the R script (Task 2) provides details of how the variables are formatted to produce the final figure.

Figure A4: This Figure is based on a different presentation of data used already in Figure 7 above. See the specific section of the R Markdown file (Task 2) for the details of how the figure is produced from this data.

Figure A5: The dataset for this figure is constructed using input data from worksheets "AI_SEDLAC_PIAAC_CPU", "AI_SEDLAC_PIAAC_CPUINT" and "AI_COMPUHH" of the Excel file "Results_AI_LAC.xlsx" (using data from column "Total", in all worksheets). These results are obtained by executing dofiles "**ai_sedlac_piaac.do"**, "**est_sedlac_piaac.do", "ai_sedlac_compuHH.do"** and **"est_sedlac_compuHH.do"** from the master dofile "**master_AI_LAC.do"** (see explanation of each dofile above). See the specific section of the R Markdown file (Task 2) for the details of how the figure is produced from this data.

Figure A6: The dataset for this figure is constructed using input data from worksheets "AI_SEDLAC_PIAAC_CPU" of the Excel file "Results_AI_LAC.xlsx", already used in Figure 11 above. See the specific section of the R Markdown file (Task 2) for the details of how the figure is produced from this data.

Table A1: This table is included in worksheet "Obs_SEDLAC_AI" of the Excel file "Results_AI_LAC.xlsx". These results are obtained by executing dofile "**chk_obs.do"** from the master dofile "**master_AI_LAC.do"** (see explanation of each dofile above).

Table A2: This table is included in worksheet "Regress_PIAAC_CPUINT" of the Excel file "Results_AI_LAC.xlsx". These results are obtained by executing dofile "**ai_sedlac_piaac.do"** (see particularly lines 229-248 of the dofile) from the master dofile "**master_AI_LAC.do"** (see explanation of each dofile above).

Table A3: This table is included in worksheet "REGRESS_AI" of the Excel file "Results_AI_LAC.xlsx". These results are obtained by executing dofile

"**reg_ai_LAC.do"** from the master dofile "**master_AI_LAC.do"** (see explanation of each dofile above).

References

Cazzaniga, M., Jaumotte, F., Li, L., Melina, G., Augustus J., P., Carlo, P., Emma J., R., Mendes, M., 2024. Gen-AI: Artificial Intelligence and the Future of Work [WWW Document]. IMF. URL https://www.imf.org/en/Publications/Staff-Discussion-Notes/Issues/2024/01/14/Gen-AI-Artificial-Intelligence-and-the-Future-of-Work-542379

Gmyrek, P., Winkler, H., Garganta, S., 2024. "Buffer or Bottleneck? Generative AI, employment exposure and the digital divide in Latin America". World Bank Policy Research Working Paper

Gmyrek, P., Berg, J., Bescond, D., 2023. Generative AI and Jobs: A global analysis of potential effects on job quantity and quality (Working paper). https://www.ilo.org/sites/default/files/wcmsp5/groups/public/@dgreports/@inst/documents/publication/wcms_890761.pdf

Prytkova, E., Petit, F., Deyu, L., Sugat, C., Tommaso, C., 2024. The Employment Impact of Emerging Digital Technologies. (https://www.fabienpetit.com/wp/PPLCC_EmploymentImpactEmergingDigitalTechnologies_Feb2024.pdf)