

Reproducibility packages of Policy Research Working Paper-10605:

'How Have Gender Gaps in the Colombian Labor Market Changed during the Economic Recovery?' by Dávalos, María E; Londoño, Diana; Medina-Gaspar, Daniel.

Free access: [<http://hdl.handle.net/10986/40639>]

The code in this replication package constructs the analysis file from the Great Integrated Household Survey (GEIH, Spanish acronyms) using Stata 15. The do-files run all of the code to generate the data for figures 1 to 34 and tables 2 to 5 in the paper. Approximate time needed to reproduce the analyses on a standard 2021 desktop machine: 3-5 hours.

All data used in this work are publicly available through the web pages of the entities that hold the information. Our main source is the Departamento Administrativo Nacional de Estadística-DANE: [www.dane.gov.co]. Terms and conditions of use: [<https://www.dane.gov.co/index.php/servicios-al-ciudadano/tramites/ayudas-de-navegabilidad/politica-de-uso>]

Because our work requires comparing results over time from 2015 to 2021 and using information from the Mission for the Splicing of the Employment, Poverty, and Inequality Series (MESEP) as a source for labor income variables, we use the 2005 frame GEIH data between 2015 and 2021.

The full dataset and documentation of GEIH can be downloaded from [<https://microdatos.dane.gov.co/index.php/catalog/MERCLAB-Microdatos>]. We obtain the GEIH microdata for each year from the 'Mercado Laboral' collection. For example, to download the year 2015, the label with which the information is identified is 'Gran Encuesta Integrada de Hogares - GEIH - 2015'. Within the same 'Mercado Laboral' collection, you will also find files that update the sector's CIU code to revision 4 for the years 2015-2019. The file is named 'Gran Encuesta Integrada de Hogares - GEIH - CIU 4 a.c. - 2015 - 2019'. It's also available at the following link [<https://microdatos.dane.gov.co/index.php/catalog/661>].

In the same repository, you can find the full dataset and documentation of 'Medición de Pobreza Monetaria y Desigualdad' from the Mission for the Splicing of the Employment, Poverty, and Inequality Series (MESEP). It can be downloaded from [<https://microdatos.dane.gov.co/index.php/catalog/POBCONVID>]. We obtain the microdata for each year from the 'Pobreza y Condiciones de Vida' collection. For example, to download the year 2015, the label with which the information is identified is 'Medición de Pobreza Monetaria y Desigualdad - 2015'.

The consumer price index used to deflate income variables is downloaded from the Banco de la República website [<https://www.banrep.gov.co/es/estadisticas/indice-precios-consumidor-ipc>].

The information on vacancies corresponds to the 'ofertas laborales' reported by the Public Employment Service (SPE, Spanish acronyms) in the Labor Demand Statistical Annex [<https://www.serviciodeempleo.gov.co/dataempleo-spe/demanda-laboral/2022/prueba>].

Finally, the names and codes of departments and municipalities are obtained from the 'Datos Abiertos' page of the Ministry of ICTs [https://www.datos.gov.co/Mapas-Nacionales/Departamentos-y-municipios-de-Colombia/xdk5-pm3f/about_data].

Folder setting

To process the raw data, you must ensure to properly organize them into folders. First, the GEIH files should be placed in a folder named 'Raw_data' per year; there, they should be uncompressed and ensure that all filenames maintain the same structure year after year. In the files used for this code, each month's folder starts with the month number followed by a period, for example: '1.Enero', similar to those downloaded from the DANE website. In the files for each month (except for 2020, which are already stacked), ensure that each of the 'Cabecera' and 'Resto' files always contain these words, to ensure proper aggregation. An example for the January 2015 folder would be 'Data_data\2015\1.Enero'.

On the other hand, data from 2015-2019 should be matched with sector information using revision 4 of the CIIU. For this, the files with CIIU information for each year downloaded from DANE should be uncompressed and placed in the folder for each year, i.e., for the year 2015, it would be in the folder 'Data_data\2015'. The file names contain the same structure, which must be maintained. For the year 2015 the file is named 'CIIU 15'.

Regarding the MESEP data, these are annualized and should only be uncompressed into a folder named MESEP, with subfolders for each year; the 'Personas' and 'Hogares' files should go in the same folder. Ensure that the filenames of these two files do not change each year. An example for the January 2015 folder would be 'Data_data\2015'.

Once all uncompressed files are in folders, run the do-file: 1_run_dofiles.do, which runs all the codes to obtain the results of the article.

Do-files description

We used nine codes to obtain the results of the paper:

- 1_run_dofiles: this is the master code that runs all do-files to generate the panel, descriptive plots, and estimates in both tables and graphs.
- 2_create_annual_files: transforms the raw files downloaded from DANE into annualized files for panel construction. Also, this builds the repeated cross-sectional panel by appending the annual data and matching the MESEP data.
- 3_cod_general_variables_WB: this is a code to create standard variables used by the World Bank team for employment issues in Colombia.
- 4_key_variables: constructs key variables for estimation and descriptive exercises.

- 5_graphs_1_20: Generates descriptive graphs 1 to 20 of the paper.
- 6_table_2: Constructs table 2 for the descriptive statistics in the paper.
- 7_set_data_est: Creates some key variables and labels for the income gap estimation models.
- 8_tables_3_5: Generates the estimates of the decomposition methods used and exports data to tables 3-5 and figures 33-34. These are then organized in Excel.
- 9_graphs_21_32: Generates the estimates of the probabilistic model and generates result graphs 21-32.

Steps to Create figures 33-34 and tables 3-5

- Each of the values in Figures 33 and 34 come from the coefficients generated in the 'Table_results' table, in the 'OB_heckman_hourly' and 'OB_heckman_monthly' sheets, respectively. Specifically, from the 'difference' coefficient we obtain the 'Estimated gap', from the 'endowments' coefficient we obtain the 'Explained' part, and from the 'coefficients' coefficient we obtain the 'Unexplained' part. To obtain the percentage difference they are transformed by subtracting 1 from each of them and multiplying by 100.
- Tables 3 and 5 come from the coefficients generated in the 'Table_results' table, in the 'OB_heckman_hourly' and 'OB_heckman_monthly' sheets, respectively. Only the coefficients of the 'endowments' section of the table are taken because we are linking the coefficients related to the gap explained by individuals' endowments. Again, to obtain the percentage difference they are transformed by subtracting 1 from each of them and multiplying by 100; For the area and sector coefficients, the percentages obtained from the table are summed to reduce the size of the table. In this way, the total explained presented in the table corresponds to the sum of the percentages explained by each variable.
- The total and the unexplained total (bottom lines) are those of the "overall" section, 'difference' and 'coefficients', respectively, using the same transformation. There is a difference between the result of the explained gap and that obtained by the sum, we kept the second one because the interest of this table is to disaggregate the explained part.

Disclaimer

We can not guarantee that the results are 100% reproducible. We have high confidence in the replicability of 2015-2019 and 2021. However, 2020 was a complex year in terms of data collection, which led to constant changes in the information reported. In addition, through 2022, DANE changed the framework of the survey expansion factors, which also generated constant changes in data reporting.

The information we used for most years was downloaded in the first half of 2022. In the case of 2020, the last information was downloaded on February 15, 2023.