README Replication Package — Where You Are Born Matters: Inequality of Opportunities and Intergenerational Mobility across Colombia's Territory (Policy Research Working Paper 11126)

Abstract. The circumstances into which individuals are born are beyond their control, yet they play a significant role in shaping people's economic opportunities and are thus key drivers of inequality and its persistence over time. Understanding the role of place of birth is essential to understanding inequality of opportunities and social mobility, both of which directly affect overall inequality. This paper uses machine learning techniques and data from Colombia, one of the most unequal countries in Latin America and the Caribbean, to estimate inequality of opportunity and intergenerational education mobility indexes. The analysis incorporate place of birth and a more granular geographic lens to capture the extent of regional disparities. The findings show that 49 percent of the Gini income inequality is explained by circumstances at birth, and place of birth accounts for up to half of these inequalities. Intergenerational mobility measures at the department (province) level also reveal striking disparities in opportunities across the country. These findings underscore the critical role that place of birth plays in perpetuating inequality, providing important insights for policies aimed at promoting social mobility and reducing territorial disparities.

Overview

- This reproducibility package replicates the results of PRWP 11126: "Where You Are Born Matters: Inequality of Opportunities and Intergenerational Mobility across Colombia's Territory" by María Eugenia Dávalos and Juan Manuel Monroy.
- It produces:
    - Intergenerational mobility (IM) measures (national and department levels) using GDIM methods.
    - Inequality of Opportunity (IOp) estimates, including Shapley value decompositions and top-income extensions.
    - Machine learning robustness checks (random trees and random forests).

Citation and Links

- Preferred citation:
    - Davalos, M., & Monroy Barragan, J. M. (2025). Where You Are Born Matters: Inequality of Opportunities and Intergenerational Mobility across Colombia's Territory (Policy Research Working Paper No. WPS11126). World Bank. https://documentos.bancomundial.org/es/publication/documents-reports/documentdetail/099721005192589316

Data Accessibility and Confidentiality

- ECV 2019, 2021, 2022: Download from DANE. Respect licensing and terms of use.
    - ECV documentation (DANE): https://www.dane.gov.co/.php/estadisticas-por-tema/pobreza-y-condiciones-de-vida/calidad-de-vida-ecv
- GDIM V3: Download from the GDIM page.
    - GDIM. 2023. Global Database on Intergenerational Mobility. Development Research Group, World Bank. Washington, D.C.: World Bank Group GDIM (V3, March 2023): https://www.worldbank.org/en/topic/poverty/brief/what-is-the-global-database-on-intergenerational-mobility-gdim
    - Ramasubbaiah, Rakesh; van der Weide, Roy; Lakner, Christoph; Mahler, Daniel Gerszon; Narayan, Ambar. 2021. Intergenerational Mobility around the World. Policy Research Working Paper;No. 9707. © World Bank. http://hdl.handle.net/10986/35827 License: CC BY 3.0 IGO
- VPN folder: Contains outputs computed within DANE's secure environment from restricted microdata at municipality level. Under DANE rules, underlying microdata cannot be shared. This package includes only permitted derivative files. Replication without DANE secure access may require skipping steps that depend on restricted intermediates.
- Colombia_tax_2022.xlsx: See the reference link above for documentation and licensing conditions.
    - Colombia tax records from Baquero, Juan Pablo; Dávalos, María Eugenia; Monroy, Juan Manuel. 2023. Revisiting the Distributive Impacts of Fiscal Policy in Colombia. Policy Research Working Papers; 10520. © World Bank. http://hdl.handle.net/10986/40035 License: CC BY 3.0 IGO : https://openknowledge.worldbank.org/entities/publication/7ccb3d1d-3816-4ed8-83a1-edd75b9dad3b
- No personal data are included in this package. Follow World Bank information classification and privacy policies when handling data.

- Mapping tutorial (Naqvi (2022). Making visually appealing maps in Stata. https://blogs.worldbank.org/en/impactevaluations/making-visually-appealing-maps-stata-guest-post-asjad-naqvi

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.
- I certify that the author(s) of the manuscript have documented permission to redistribute/publish the data contained within this replication package. Appropriate permission are documented in the LICENSE.txt file.

Repository Structure

data/

- in/
    - World.dta, world_shp/ — World shapefiles for global mapping (Stata format)

      Source https://datacatalog.worldbank.org/int/search/dataset/0038272
    - GDIM_2023_03/ — Global Database on Intergenerational Mobility (V3, Mar 2023)
    - depto.dta, depto_shp/ — Colombia department-level shapefiles

      Source https://bart.ideam.gov.co/ospa/shape/Colombia/DEPARTAMENTOS/
    - Colombia_tax_2022.xlsx — Interpolated tax records input (2022). Note: forthcoming data in DDH.
    - ECV/ — ECV household surveys (2019, 2021, 2022) from DANE. (Encuesta Nacional de Calidad de Vida, ENCV, 2019, 2021, 2022, DANE), this files are for public distribution. The year of the processing is 2025. It can be consulted and downloaded directly from the source:

      Source https://www.dane.gov.co/index.php/estadisticas-por-tema/pobreza-y-condiciones-de-vida/calidad-de-vida-ecv

      Processing steps. It consists of creating harmonized variables following GDIM, 2023, Global Database on Intergenerational Mobility v3 guidelines.
    - VPN/ — .dta files computed within DANE secure environment (municipality-level; restricted).
        - **Raw Input Source:** The primary source is the Encuesta Nacional de Calidad de Vida (ECV/ENCV), conducted by DANE for the years 2019,

2021, and 2022. While the ECV is not publicly representative at the municipality level, DANE granted us access to the restricted variables through a secure virtual environment (VPN) starting in July 2024. **Derivative Steps Taken in the Secure Environment:** Within DANE's secure environment, we constructed a dummy variable identifying individuals born in a municipality that experienced armed conflict in approximately a five-year window around their year of birth. Critically, the municipality identifier itself is not included in the output files — only the derived dummy variable is retained. The files were shared with and reviewed by DANE before extraction from the secure environment. **Publication Status:** Because the output files do not contain municipality-level identifiers and were cleared by DANE following their internal review process, they do not violate DANE's data publication policies. These files can be included in the replication package.

    - pobreza_monetaria_21_dep.xls. This file can be downloaded directly from the DANE´s webpage in Anexo pobreza monetaria departamental. Published 26/04/2022. By DANE https://www.dane.gov.co/index.php/estadisticas-por-tema/pobreza-y-condiciones-de-vida/pobreza-monetaria/pobreza-monetaria-2021

    - IM_index.xlsx : it contains the list of OECD countries using the 3-digit identification for each country. An additional column of values 1 indicating each raw is an OECD country is included.

- out/
    - depto_pool_meaures_ext2.dta — Department-level IM outputs
    - natio_pool_measures.dta — National-level IM outputs
    - COL_192122_ENCV_v01_M_v01_A_IGM_2.dta — Harmonized pooled ECV dataset (2019, 2021, 2022)

dofiles/ (Stata)

- im_io 1.0 Master_out.do
- im 2.0 Harmonize from global codes 2019_21_22.do
- im 3.0 Measures from global codes. 2019_21_22.do
- im 4.0 Measures from global codes dpto. 2019_21_22.do
- im 5.0 Global comparisson.do

- im 6.0 Maps.do
- io 2.0 Data Preparation.do
- io 3.0 IO estimation.do
- io 4.0 Shapley value.do
- io 5.1 Top incomes.do
- io 5.2 Full data top.do
- io 6.0 IO estimation top.do
- io 8.0 IO Aditional outputs.do

R/

- io 7.1 randomtree pcincome.R
- io 7.2 randomtree laborincome.R
- io 7.3 randomtree topincomes.R
- io 7.4 randomforest pcincome.R
- io 7.5 randomforest laborincome.R
- io 7.6 Oportunity tree.R

Data cleaning ECV

Data Assembly. The script begins by loading and merging five separate module files for each survey year — dwelling, household conditions, household services, education, and labor force — along with the household roster. These are joined using dwelling and household sequence identifiers (directorio, secuencia_encuesta, secuencia_p). Once all modules are merged within each year, the three annual files are appended into one pooled dataset. A composite household identifier (llavehogar) is constructed by concatenating the two key identifiers.

Technical and Identifier Variables. Standard identifiers are generated: country code (COL), survey name (ENCV), household and individual IDs, and survey weights. Several design variables — strata and PSU — are left missing as they could not be recovered from the available files, which is transparently flagged in the code.

Sociodemographic Variables. The script constructs sex, age, birth year, urban/rural status, region, marital status, household size, number of children, relationship to household head, ethnicity, and co-residence with each parent. Birth year is derived by subtracting age from the survey year. A

basic plausibility filter is applied to parental birth years, setting them to missing when the implied age gap between parent and child is less than 13 years. Ethnicity is coded into two alternative groupings that differ in how they handle Afro-descendant subgroups.

Labor and Asset Variables. Labor force status follows SEDLAC harmonization conventions, distinguishing employed, unemployed, and out-of-the-labor-force individuals based on a combination of activity and job search questions. Employment type is classified into paid employee, self-employed, employer, and unpaid worker. Labor income is drawn from a single wage variable, with implausible codes (98, 99) set to missing and non-wage workers excluded. The script also constructs supplementary income components — bonuses and in-kind payments — and documents the reference period, hours worked, and currency units for each. On the asset side, household-level ownership of phones, TV, computer, washing machine, car, and motorcycle is derived by taking the household maximum across individual responses, after recoding the original binary scale (1/2) to (1/0).

Final Output. The dataset is trimmed to a predefined list of harmonized variables, reordered for consistency, compressed, and saved.

Education Variable Construction

The education section is the most elaborate part and builds several variables aligned with the **ISCED 2011** classification framework.

Respondent's education (educcatIGM) is built from two complementary survey questions: one capturing the level currently being attended (for those enrolled, variable p1088) and one capturing the highest level completed (for those no longer studying, variables p8587 and p8587s1). The mapping follows this logic:

- No schooling or preschool → ISCED 0
- Primary (up to grade 5) → ISCED 1
- Lower secondary (up to grade 9) → ISCED 2
- Upper secondary (grade 11+) or technical/vocational → ISCED 3
- Short-cycle tertiary (technical/technological) → ISCED 5
- University (bachelor's) → ISCED 6
- Postgraduate (master's or doctorate) → ISCED 7–8

Children under age 5 are assigned ISCED 0 by default.

Parents' education (educcatIGM_dad, educcatIGM_mom) is constructed through a two-step process that handles a key complexity in household surveys: whether the parent lives in the same household or not. For **absent parents**, a coarser categorical variable from the survey (p6087 for fathers, p6088 for mothers) is used to assign ISCED levels. For **co-resident parents**, the script uses a preserve/restore block to extract the parent's own education record (already coded as educcatIGM) and merges it back to the child using the household member link identifiers (p6081s1 for father, p6083s1 for mother). The merge uses update to fill in only the missing values, so the absent-parent estimates are not overwritten.

Derived education variables are then generated in three levels of aggregation:

- **8-category** (educcat8IGM): divides educcatIGM by 10,000 and collapses ISCED 7 and 8 into a single "7–8" group.
- **5-category** (educcat5IGM): further recodes by grouping ISCED 3–4 and ISCED 5–8 into two broad bands, useful for cross-country comparisons.
- **Continuous years of schooling** (educyIGM): maps grade/level information to estimated years of schooling. For the currently enrolled, it uses the grade within the level (p1088s1); for those no longer studying, it uses p8587s1. Tertiary levels add 11 or 16 baseline years depending on the degree type.

The same continuous years-of-schooling logic is applied to parents, though for absent parents a simplified fixed-value imputation is used (e.g., 5 years for primary, 8 for lower secondary, 16 for university) given the coarser categorical information available.

Software and Dependencies

- Stata 16 or later (Stata/MP recommended).
- Stata user-written packages used in the dofiles (install in the master script as needed):
    - Iop, shp2dta, spmap
    - For IOP command see: Florian Chavez Juarez & Isidro Soloaga, 2012. "**IOP: Stata module to to compute different measures of inequality of opportunity for dichotomous, ordered and continuous outcome variables**," Statistical Software Components S457442, Boston College Department of Economics, revised 21 Dec 2015.
- R 4.2+; recommended packages:
    - ranger, randomForest, rpart, rpart.plot, data.table, dplyr, haven or readstata13, ggplot2, sf (optional for mapping)

How to Run (One-Click Pipeline)

- Clone the repository and keep the folder structure unchanged.
- Place inputs in data/in/ according to “Repository Structure.”
- Open Stata and edit dofiles/im_io 1.0 Master_out.do to set root and path globals (e.g., $root, $data, $dofile).
- Optional: add/install blocks for any user-written Stata commands (e.g., ssc install shp2dta, ssc install spmap).
- Run im_io 1.0 Master_out.do. It will execute:
    - Harmonize pooled ECV (2019, 2021, 2022): im 2.0 Harmonize from global codes 2019_21_22.do
    - Estimate IM measures (national): im 3.0 Measures from global codes. 2019_21_22.do
    - Estimate IM measures (department): im 4.0 Measures from global codes dpto. 2019_21_22.do
    - Generate national outcomes and global comparisons: im 5.0 Global comparisson.do
    - Generate department outcomes and maps: im 6.0 Maps.do
- IOp estimation:
    - Prepare data and assumptions: io 2.0 Data Preparation.do
    - Specifications:
        - i) without municipality of birth
        - ii) with municipality of birth
        - iii) with municipality of birth and conflict variable
    - Run: io 3.0 IO estimation.do
- Shapley and top-income extensions:
    - io 4.0 Shapley value.do
    - io 5.1 Top incomes.do
    - io 5.2 Full data top.do

        - io 6.0 IO estimation top.do
        - io 8.0 IO Aditional outputs.do
- Machine learning (R):
    - Ensure R paths match your repository.
    - Run:
        - Main.R
        - io 7.1 randomtree pcincome.R
        - io 7.2 randomtree laborincome.R
        - io 7.3 randomtree topincomes.R
        - io 7.4 randomforest pcincome.R
        - io 7.5 randomforest laborincome.R
        - io 7.6 Oportunity tree.R

Example Path Setup (Stata) Add to the top of im_io 1.0 Master_out.do:

```
* ---- Master globals ----
global root   "<your-local-path-to-repo>"
global data   "$root/data"
global in     "$data/in"
global out    "$data/out"
global dofile "$root/dofiles"
cd "$dofile"
* ---- Reproducibility ----
set seed 1234567
* ---- User-written packages (install if missing) ----
cap which shp2dta
if _rc ssc install shp2dta
cap which spmap
```

if _rc ssc install spmap

Example R Setup

- Optional renv snapshot:

 # In R

install.packages("renv")

renv::init()

Install your packages, then:

install.packages("party") # To calculate trees and forests

install.packages("haven") # To load .dta files

install.packages("dineq") # Weighted MLD and Gini

renv::snapshot()

- Reading Stata data (haven):

library(haven); library(dplyr)

df <- read_dta(file.path("data","out","COL_192122_ENCV_v01_M_v01_A_IGM_2.dta"))

set.seed(1234567)

Mapping and CRS Notes

- Coordinate Reference System (CRS): EPSG:4326 (WGS84) for world_shp and depto_shp.
- Ensure shp2dta/spmap usage references EPSG:4326 consistently to avoid projection issues.

Expected Outputs

- data/out/depto_pool_meaures_ext2.dta (department-level IM)
- data/out/natio_pool_measures.dta (national IM)
- data/out/COL_192122_ENCV_v01_M_v01_A_IGM_2.dta (harmonized pooled ECV 2019, 2021, 2022)
- Figures, maps, and tables saved under output/ or subfolders defined in scripts; image format: PNG.

Reproducibility Notes

- Paths: Use relative paths and set them in the master dofile and Main.R only.
- Randomness: Seed set to 1234567 for both Stata and R.
- Environment capture:
    - Stata: consider logging a session summary (about) and package list.
    - R: include renv.lock after renv::snapshot to lock package versions.
- Restricted steps: Clearly annotate any steps that require DANE-secure access; provide dummy-data instructions if feasible.

Troubleshooting

- Path errors: Verify globals in im_io 1.0 Master_out.do.
- Missing inputs: Confirm data are in data/in/ and named as in this README.
- Shapefiles: Confirm EPSG:4326 and file references before rendering maps.
- Memory: Stata/MP and R randomforest steps can be memory intensive—ensure sufficient RAM and consider limiting trees or sample sizes if necessary.

Contact

- For replication questions: Maria Davalos (World Bank Group), mdavalos@worldbank.org , and Juan Manuel Monroy (World Bank Group) jmonroybarragan@gmail.com

Acknowledgments

- DANE for secure processing support.
- GDIM team for documentation and code guidance, Daniel Gerszon Mahler - World Bank.

License

- No explicit license is granted for the code in this repository. Data remain subject to the original providers' terms (DANE, World Bank GDIM, etc.). Review each source's license before redistribution.

Appendix A: Folder Tree (Illustrative)

repo_root/

├– data/

| ├– in/

```
|  |  ├– World.dta
|  |  ├– world_shp/
|  |  ├– GDIM_2023_03/
|  |  ├– depto.dta
|  |  ├– depto_shp/
|  |  ├– Colombia_tax_2022.xlsx
|  |  ├– ECV/
|  |  └– VPN/
|  └– out/
|     ├– depto_pool_meaures_ext2.dta
|     ├– natio_pool_measures.dta
|     └– COL_192122_ENCV_v01_M_v01_A_IGM_2.dta
├– dofiles/
└– R/
```