

Reproducibility Package for Prosperity Insights Paper:

People, Jobs, and Places: A Spatial Analysis of Welfare Disparities in Low- and Middle-Income Countries

Contents

1. Overview
2. Data Availability
3. Instructions for Replicators
4. List of Exhibits
5. Requirements
6. Code Description
7. Folder Structure

Overview

Begin by offering a concise overview for the replicator regarding the materials included in the package and provide a brief guide on how to proceed from start to finish. Ensure to include any crucial information that replicators should be aware of.

Data Availability

- ☐ All data are publicly available.
- ☒ Some data cannot be made publicly available.
- ☐ No data can be made publicly available.

Data Sources

Analysis presented in this reproducibility package includes two country-specific deep dives, one for Mali and another one for Ecuador. While the data used is different, the common denominator for the microdata used in this project is the Data Lib Web (DLW) <https://datalibweb2.worldbank.org/home> (<https://datalibweb2.worldbank.org/home>). This platform disseminates harmonized data from the global monitoring database (GMD).

DLW is an internal World Bank Group platform. Access requires a valid World Bank network account (staff or consultant credentials) and is **not available to external, non-World-Bank users**. Replicators without World Bank access will not be able to authenticate to DLW and therefore cannot execute the steps of this pipeline that download data from it (see the `Private - API` access mode below). See <https://github.com/worldbank/datalibweb> (<https://github.com/worldbank/datalibweb>) for details on who can access DLW and how.

One needs to have access to the MLI survey on the data lib web to be able to reproduce this analysis as well as specify DLW API token in R and Stata to be able to load this data in. Detailed instructions on access to the DLW data are available here: <https://datalibweb2.worldbank.org/assets/images/DLW%20manual%20-%20general%20user.pdf> (<https://datalibweb2.worldbank.org/assets/images/DLW%20manual%20-%20general%20user.pdf>). Stata package to access DLW data is available here: <https://github.com/worldbank/datalibweb> (<https://github.com/worldbank/datalibweb>) R code relies on the `d1w` R package <https://github.com/worldbank/dlw> (<https://github.com/worldbank/dlw>) for data access. More detailed instructions are present in the section Instructions for Replicators.

DLW offers a range of data sets “modules” available for each country. We use:

- Full harmonized country specific data: module “ALL” (GMD)
- Household locations geocodes: module “H3”
- Spatial data for the locations: module “SPAT”

- Location identifiers: module “LOC”
- CPI/ICP deflators: module “CPIICP”
- Global Monitoring Indicators poverty data: module “GMI”

All documentation, metadata, and details on the modules are available on the DLW website.

In addition, to further augment other data sources with the spatial data, we use zonal statistics precomputed at the global Hexagonal (grid H3) at the resolution level ‘6’. This data set is used for aggregating zonal statistics for DLW SPAT module, therefore, metadata that describes ‘SPAT’ module is also relevant to it.

Below we provide additional details on the specific data sets used for each specific country analysis. Note that some data is private and requires appropriate access permissions. We use three modes of access permissions:

1. **Private - OneDrive** : private data that cannot be publicly disseminated is shared through the OneDrive.
2. **Private - API** : Some private data is accessible through the DLW web API after proper authentication. DLW (Data Lib Web) is an **internal World Bank Group platform** — it is restricted to users with World Bank staff or consultant network credentials and is not accessible to external users. See <https://github.com/worldbank/datalibweb> (<https://github.com/worldbank/datalibweb>) for access details.
3. **Public - API** : Some geospatial data is publicly available and does not require special access permissions. If API for it is available, it is used.

For the reproducibility purposes private data is being stored in a single place outlined with the R variable `fldr_dta_root` in the vignettes/01-setup.R script. User must set this variable in their own session to be able to execute this analysis.

Note: if you have access to the data folder `spat-disp-insight-data` on the OneDrive you do not need to download any additional data or save it anywhere. If you’d like to reproduce this analysis from scratch without having access to the data, feel free to organize files as you need.

Mali

Data used in the analysis for Mali is coming from multiple sources, which include

- HH survey, location, CPI and poverty data for 2021 (loaded within the code from the DLW)
- Raw household survey data (private),
- Enterprise census data (private),
- publicly available geospatial data,
- Agro-ecological zones shapefile (public),
- World Bank CLASS data (public)

Raw household survey data (private - OneDrive)

The data was confidentially shared by the Poverty Economist of Mali, and cannot be made available. Please contact Johannes Hoogeveen (jhoogeveen@worldbank.org (<mailto:jhoogeveen@worldbank.org>)) for access to the datasets. To access the data from outside of the World Bank, please refer to the Mali Statistics Office (<https://www.instat-mali.org/fr> (<https://www.instat-mali.org/fr>), direction@instat.ml (<mailto:direction@instat.ml>)).

The data is shared only for the reproducibility verification.

- **Filename 1:** `<fldr_dta_root>/MLI/private/ehcvm_menage_MLI2021.dta`
- **Filename 2:** `<fldr_dta_root>/MLI/private/s19_me_MLI2021.dta`
- **Filename 3:** `<fldr_dta_root>/MLI/private/s05_me_MLI2021.dta`
- **Filename 4:** `<fldr_dta_root>/MLI/private/s13_2_me_MLI2021.dta`
- **Filename 5:** `<fldr_dta_root>/MLI/private/s15_me_MLI2021.dta`
- **Source:** Mali Statistical Office
- **URL:** NA
- **Access date:** 11/14/2025

Enterprise census data (private - OneDrive)

The data was confidentially shared by the Poverty Economist of Mali, and cannot be made available. It is shared only for the reproducibility verification.

- **Filename 1:** `<fldr_dta_root>/MLI/private/Enterprise_census/QG_M2a_Global.sav`

- **Source:** Mali Statistical Office
- **URL:** NA
- **Access date:** 02/10/2026

World Population global rasters (public - API)

This data is fetched from the WorldPop open data portal. The reproducibility package downloads the files automatically — no API keys are needed. Files are saved locally and reused on subsequent runs.

- **Filename:** 20 .tif files following the pattern `mli_{sex}_{age}_2020_constrained_UNadj.tif` (10 age bands 15–60, male and female), saved to `data/raw/MLI/worldpop/`
- **Source:** WorldPop — Open Spatial Demographic Data
- **URL:** https://data.worldpop.org/GIS/AgeSex_structures/Global_2000_2020_Constrained_UNadj (https://data.worldpop.org/GIS/AgeSex_structures/Global_2000_2020_Constrained_UNadj)
- **Access date:** 06/10/2026

JRC: Global Human Settlement Layer (public)

Accessed in-memory via GDAL virtual file system (`/vsizip//vsicur1/`). The raster is streamed directly — no file is downloaded or saved to disk. No API key is needed.

- **Filename:** `GHS_WUP_BUILT_S_E2020_GLOBE_R2025A_4326_30ss_V1-0.tif` (single raster, built-up surface fraction, year 2020, WGS84, ~1 km)
- **Source:** JRC — Global Human Settlement Layer (GHSL R2025A)
- **URL:** https://jeodpp.jrc.ec.europa.eu/ftp/jrc-opendata/GHSL/GHS_WUP_BUILT_S_GLOBE_R2025A/ (https://jeodpp.jrc.ec.europa.eu/ftp/jrc-opendata/GHSL/GHS_WUP_BUILT_S_GLOBE_R2025A/)
- **Access date:** 06/10/2026

Census-specific spatial boundaries at different administrative levels (public)

This data is used in mapping.

- **Filename 1:** `<fldr_dta_root>/MLI/mli_adm_1m_dnct_2021_shp/mli_admbnda_adm3_1m_gov_20211110b.shp`
- **Source:** HDX
- **URL:** https://data.humdata.org/dataset/3feaf6d7-8b21-4db1-a097-fa8a2b680a89/resource/303f57d8-965f-4c70-a081-3e5e0fcb2bf/download/mli_adm_1m_dnct_2021_shp.zip (https://data.humdata.org/dataset/3feaf6d7-8b21-4db1-a097-fa8a2b680a89/resource/303f57d8-965f-4c70-a081-3e5e0fcb2bf/download/mli_adm_1m_dnct_2021_shp.zip)
- **Access year:** 2026

Agro-ecological zones shapefile (public)

Used to assign households to agro-ecological zones for wage imputation. Publicly available from the FAO Global Agro-Ecological Zones (GAEZ) portal.

- **Filename 1:** `<fldr_dta_root>/MLI/agrizones/_Zones_Agroclimatiques_WGS84.shp`
- **Source:** FAO GAEZ — Global Agro-Ecological Zones
- **URL:** <https://gaez.fao.org/datasets/-zones-agroclimatiques-de-mali> (<https://gaez.fao.org/datasets/-zones-agroclimatiques-de-mali>)
- **Access date:** 02/10/2026

World Bank CLASS data (public)

This data contains World Bank country classifications.

- **Filename 1:** `CLASS.dta`
- **Source:** Github GPID-WB
- **URL:** <https://raw.githubusercontent.com/GPID-WB/Class/master/OutputData/CLASS.dta> (<https://raw.githubusercontent.com/GPID-WB/Class/master/OutputData/CLASS.dta>)
- **Access year:** 2026

Ecuador

Data used in the analysis for Ecuador is coming from multiple sources, and has a three-step processing pipeline:

1. Raw data re-saving in `parquet` or other fast-readable formats.
2. Cleaning the data and transforming it to the format used in the analysis.

3. Calculating indicators and assembling the data for analysis.

We only develop the reproducibility pipeline from the second step, while the code that converts step one data into the step two data is present for reference.

`<raw-data-folder>` below refers to the location, where private data is stored in the raw format and from where it is resaved in parquet format. Since this is not part of a reproducibility pipeline, we do not provide details on its structure. All relevant files used from these locations are referenced explicitly. Note that although `<raw-data-folder>` is not shared as a part of the reproducibility package the data from the raw is resaved in exact the same structure but in a fast- readable format. Essentially same raw data is the part of the reproducibility pipeline.

The data was confidentially shared by the Poverty Economist of Ecuador, and cannot be made available. Please contact Gustavo Javier Canavire Bacarreza gcanavire@worldbank.org (mailto:gcanavire@worldbank.org) for access to the datasets. It was shared with us in a pre-processed and pre-harmonized way that combined survey data for different periods. To access the data from outside of the World Bank, please refer to the data-set-specific instructions below.

- Annual household survey data for 2016–2024 (ENEMDU)
- Administrative subnational data on production, consumption, and taxes for 2017–2023 (private - OneDrive)
- Census microdata from the National Institute of Statistics and Census (INEC) for 2001, 2010, and 2022 (private - OneDrive)
- Publicly available geospatial data

Annual Household Survey — ENEMDU (private - OneDrive)

Annual National Survey of Employment, Unemployment and Underemployment (Encuesta Nacional de Empleo, Desempleo y Subempleo — ENEMDU), covering the period 2016–2024. Data is a product of harmonization made by the country team and is not publicly available. Reproduction of the harmonization code is beyond the scope of this analysis. To access this data outside the World Bank, please refer to National Institute of Statistics and Censuses (Ecuador) <https://www.ecuadorencifras.gob.ec/enemdu-anual/> (<https://www.ecuadorencifras.gob.ec/enemdu-anual/>). This is a data set underlying the analysis.

- **Filename (raw):** `<raw-data-folder>/ECU/Household Surveys/0_enemdu2007_24.dta`
- **Filename (resaved):** `<fldr_dta_root>/ECU/hh-dta/part-0.parquet`
- **Source:** WB Ecuador Poverty Economist
- **URL:** <https://www.ecuadorencifras.gob.ec/enemdu-anual/> (<https://www.ecuadorencifras.gob.ec/enemdu-anual/>)
- **Access year:** December 2025

Administrative Subnational Data on Production, Consumption and Taxes

(private - OneDrive)

Canton-level administrative data on production, consumption, and tax revenues by industry and year, covering 2017–2023. Data is a product of harmonization made by the country team and is not publicly available. It is based on the country-specific administrative data on taxes and does not have a publicly available dissemination.

- **Filename 1 raw (production / consumption):**
`<raw-data-folder>/ECU/Tax data parroquias/Anexo_F104_AGR_CIIU_PARROQUIA 2.csv`
- **Filename 2 raw (taxes, SRI):** `<raw-data-folder>/ECU/Tax data parroquias/SRI_2017_2024/*.csv`
- **Filename 1 resaved data:** `<fldr_dta_root>/ECU/prod-cons-dta/part-0.parquet`
- **Filename 2 resaved data:** `<fldr_dta_root>/ECU/tax-dta/part-0.parquet`
- **Source:** WB Ecuador Poverty Economist
- **URL:** NA
- **Access year:** December 2025

Census Microdata — INEC (private - OneDrive)

Microdata from the national population and housing censuses conducted by the National Institute of Statistics and Census (Instituto Nacional de Estadística y Censos — INEC) for the years 2001, 2010, and 2022. Used to construct canton- and parish-level socio-demographic covariates. Shared through the OneDrive shared folder. This data is publicly available at <https://www.ecuadorencifras.gob.ec/> (<https://www.ecuadorencifras.gob.ec/>)

- **Filename 1 (2001):**
<raw-data-folder>/ECU/Censos/BBD_CPV_2001_NACIONAL_CON_CORRESP_2022_SPSS/CPV_2001_poblacion_Nacional_corresp2022.sav
- **Filename 2 (2010):**
<raw-data-folder>/ECU/Censos/BBD_CPV_2010_NACIONAL_CON_CORRESP_2022_SPSS/CPV_2010_poblacion_Nacional_corresp2022.sav
- **Filename 3 (2022):**
<raw-data-folder>/ECU/Censos/BBD_CPV2022_SECTOR_SAV/BBD_CPV2022_SECTOR_SAV/CPV_Población_2022_Nacional_Sector.sav
- **Source:** INEC — Instituto Nacional de Estadística y Censos
- **URL:** <https://www.ecuadorencifras.gob.ec/> (<https://www.ecuadorencifras.gob.ec/>)
- **Access year:** February 2026

Census-specific spatial boundaries at different administrative levels (private - OneDrive)

Administrative boundary shapefiles at ADM0–ADM3 level (Parroquia), drawn from the CONALI 2022 delimitation and sector census sources. Used for spatial matching census data with the survey and administrative data (Filename 1 below). Filename 2 below was used for mapping and analysis as it corresponded to the survey data locations. This file was provided by the poverty economist as part of the survey and is not available publicly.

- **Filename 1 raw:** <fldr_dta_root>/ECU/LIMITE_PARROQUIAL_CONALI_CNE_2022/LIMITE_PARROQUIAL_CONALI_CNE_2022.shp
- **Filename 2 raw:** <fldr_dta_root>/ECU/Shapefile CNE - Ecuador/sector_censal_parr/sector_censal_parr.shp
- **Source for file 1:** Consejo Nacional Electoral (CNE)
- **URL for file 1:** <https://www.arcgis.com/home/item.html?id=48b3a76be08e4eb8b2bbae542fb04da9>
(<https://www.arcgis.com/home/item.html?id=48b3a76be08e4eb8b2bbae542fb04da9>)
- **Access year:** February 2026

H3 Hexagon-level zonal statistics for Ecuador

This data is used as the underlying data for Spatial data module in the GMD data set. It is not publicly available. GMD only disseminates this data on the per-country basis with the hexagonal data aggregated at the location level defined in the GMD. We use original H3 level zonal statistics computed for the GMD data but aggregate them in the way as our analysis-specific spatial boundaries and individual coordinates permit. For example, we use H3 statistics for the locations where individuals reside, with a buffer around individuals and different polygon-specific aggregates.

- **Filename 1:** <fldr_dta_root>/ECU/h3-ecu-dta/part-0.parquet
- **Source:** GMD
- **URL:** <https://datalibweb2.worldbank.org/home> (<https://datalibweb2.worldbank.org/home>)
- **Access year:** February 2026

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.

Instructions for Replicators

New users should follow these steps to run the package successfully:

1. Clone this repository from the Github to your local computer.
2. Restore R `renv` environment for the correct package versions management by running in the fresh R session opened in this project `renv::restore()`
3. Synchronize private data storage through the OneDrive. This is the read-only data it is required for the replication. To request access to this data reach out to Nga Thi Viet Nguyen nnguyen3@worldbank.org (<mailto:nnguyen3@worldbank.org>)
4. Specify path to this synchronized folder by defining the `fldr_dta_root` in the line 110 of `vignettes/01-setup.R` as the string where the data is located on your computer. If this path is misspecified, the script will fail with an error message.
5. Get access to the Data Lib Web <https://datalibweb2.worldbank.org/home> (<https://datalibweb2.worldbank.org/home>).

DLW access is restricted to World Bank Group staff and consultants with a valid WB network account. This step cannot be completed by external (non-World-Bank) replicators — see <https://github.com/worldbank/datalibweb> (<https://github.com/worldbank/datalibweb>) for details. Without DLW access, the Mali GMD download step (`MLI-01-data-load.qmd`) and the Ecuador H3 zonal statistics input cannot be reproduced directly; contact the data-access contacts listed in Data Availability for alternative arrangements.

1. Authenticate on the <https://datalibweb2.worldbank.org/home> (<https://datalibweb2.worldbank.org/home>). On your name on the top right, click renew token and copy it.
2. All Mali household surveys from GMD should be accessible from within the World Bank network. If you cannot access the data, request access to Mali HH surveys. Go to Directory, and select MLI under Select Country. Select GMD_MLI and Add to Cart.
3. In R: Install R package `dlw` by running `remotes::install_github("worldbank/dlw")`
4. Import the copied access token into the R environment by running: `dlw::dlw_set_token("<your_token_here>")`
5. Once authenticated and access is granted, the DLW modules (ALL, SPAT, LOC, H3, CPIICP, GMI) for Mali (or global) are downloaded automatically when running the master script with the correct vintage. No manual file download is required — the code in `vignettes/MLI-01-data-load.qmd` handles this via the `dlw_get_gmd()` function and saves the data to `data/interm/gmd_*/`.
6. Run in R `vignettes/00-master.R` file in full by sourcing it.

List of Exhibits

After successful code execution, folder `output` will contain all figures and tables published in the paper. Files names in Output correspond to the figures names in the paper.

The provided code reproduces:

- ☐ All numbers provided in text in the paper
- ☒ All tables and figures in the paper
- ☐ Selected tables and figures in the paper, as explained and justified below

Requirements

Computational Requirements

This analysis does not require any sophisticated hardware and could be executed on a regular laptop with 32 GB of RAM.

Software Requirements

It requires R version 4.5.3. All packages used in running code are defined in the respective scripts and version controlled through `renv`.

Memory and Runtime and Storage Requirements

Memory: 32gb Runtime: ~30 min. Storage: ~20 GB

Code Description

This reproducibility package is structured as a formal **R package** (`spatdispinsight`), following the standard R package conventions with a `DESCRIPTION`, `NAMESPACE`, and `man/` directory for documentation. All analytical functions are defined in the `R/` folder and are automatically installed and made available when the master script is executed.

Package dependency management — ensuring that all supporting R packages are available at the correct versions — is handled by `renv` (<https://rstudio.github.io/renv/>). The `renv.lock` file records the exact state of the R library at the time of the analysis. Running `renv::restore()` in a fresh R session reproduces that environment exactly before any analysis is executed.

The single entry point for the full reproducible pipeline is `vignettes/00-master.R`. When run, it:

1. Sources `vignettes/01-setup.R` to configure shared folder paths and settings. (Ecuador-specific paths in `vignettes/01-setup-ECU.R` are sourced separately, at the top of `ECU-03-data-reg.qmd` and `ECU-05-reproducibility.qmd` themselves.)
2. Optionally clears all previously generated intermediate and output files (controlled by the `FORCE` flag at the top of the script) to ensure a clean re-run from scratch.
3. Renders each analytical Quarto document in sequence via `quarto::quarto_render()`, passing execution parameters where needed.

The pipeline is strictly sequential: each script depends on outputs produced by the preceding one. All intermediate and final outputs are written to git-ignored folders (`data/raw/`, `data/interm/`, `data/clean/`, `output/`), so that generated files never contaminate the git history. Only source code, configuration files, and the `renv.lock` are tracked by version control.

Mali

The Mali data pipeline is implemented across eight scripts (`vignettes/MLI-01` through `MLI-06`, including the split `MLI-03a` / `MLI-03b`, plus one shared geometry script). All scripts are executed as part of the master `00-master.R` file.

The overall analysis is built as an R package with an appropriate `DESCRIPTION` and `NAMESPACE` files. All R functions are developed in `R/`.

`vignettes/MLI-01-data-load.qmd` — GMD Data Download

Downloads all required GMD modules for Mali from the Data Lib Web API (`dlw_get_gmd()`). Queries the DLW server catalogue, identifies available modules, and writes raw module data (ALL, SPAT, LOC, H3, CPIICP, GMI) to partitioned Parquet datasets. Also downloads World Bank CLASS country classification data.

- **Inputs:** DLW API (authenticated); World Bank CLASS data (public GitHub)
- **Outputs** (written to `data/interm/`):
 - `dlw_server_catalog.parquet` — DLW module catalogue
 - `gmd_all/` — raw GMD ALL module (Parquet)
 - `gmd_h3/` — household geocodes H3 module (Parquet)
 - `gmd_spat/` — spatial zonal statistics SPAT module (Parquet)
 - `gmd_loc/` — location LOC module (Parquet)
 - `gmd_cpiicp/` — CPI/ICP support module (Parquet)
 - `gmd_gmi/` — GMI poverty module (Parquet)

`vignettes/11-geom.qmd` — Spatial Data Preparation

Prepares and processes geospatial data used across both country analyses. Downloads and processes World Population rasters (WorldPop API) and Global Human Settlement Layer (JRC GHSL API). Processes administrative boundary shapefiles and computes H3 hexagon geometries at resolution 6.

- **Inputs:**
 - Administrative boundary shapefiles (private - OneDrive): `mli_admbnda_adm3_1m_gov_20211110b.shp`
- **Outputs** (written to `data/raw/` and `data/interm/`):
 - Administrative boundaries at ADM0–ADM3 (Parquet/GeoJSON)
 - H3 hexagon grid geometries

`vignettes/MLI-02-data-clean.qmd` — GMD Data Cleaning

Cleans and standardises the raw GMD ALL module data using the shared `f_clean_gmd()` function. Harmonises variable names and types across survey years (2018, 2021) and writes the cleaned dataset for downstream use.

- **Inputs:** `data/interm/gmd_all/` (from MLI-01)
- **Outputs:** `data/interm/gmd/` — cleaned GMD data (Parquet)

`vignettes/MLI-03a-imputation-prep.R` — Wage Imputation Data Preparation

Prepares the input dataset for wage imputation by merging the cleaned GMD data with private EHCVM survey modules and assigning households to agro-ecological zones (AEZ) via spatial join.

- **Inputs:**
 - data/interm/gmd/ (from MLI-02)
 - Private EHCVM modules (OneDrive): ehcv_menage_MLI2021.dta (livestock and agricultural land), s19_me_MLI2021.dta (agricultural assets), s05_me_MLI2021.dta (non-labour income), s13_2_me_MLI2021.dta (remittances), s15_me_MLI2021.dta (social transfers)
 - AEZ shapefile (public - FAO GAEZ): _Zones_Agroclimatiques_WGS84.shp
- **Outputs** (written to data/interm/):
 - mli_nonlabor.rds — household non-labour income variables
 - mli_testing.rds — merged dataset ready for imputation

vignettes/MLI-03b-imp-data-annex.R — Wage Imputation and Annex Table

Runs within-cell wage imputation regressions (hours slope + cell fixed effects) by sector using `fixest`. Exports the regression coefficient table as Appendix 5 and saves the AEZ-augmented dataset for downstream use.

- **Inputs:** data/interm/mli_testing.rds, data/interm/mli_nonlabor.rds (from MLI-03a)
- **Outputs:**
 - output/tables/Appendix_4.xlsx — imputation regression table
 - data/interm/mli_aez_2021.rds — imputed dataset with AEZ assignment

vignettes/MLI-04-census-prep.R — Enterprise Census Data Preparation

Loads and cleans the private enterprise census microdata. Selects and renames variables, geocodes establishments, and saves the cleaned census for use in the final data preparation step.

- **Inputs:** MLI/private/Enterprise census/QG_M2a_Global.sav (OneDrive)
- **Outputs:** data/raw/MLI/mli_econ_census.rds

vignettes/MLI-05-data-prep.qmd — Final Data Preparation

Constructs the analysis-ready dataset by merging cleaned GMD data with imputed wages, enterprise census aggregates, and spatial covariates. Applies survey-year-specific variable harmonisation (2018 vs. 2021), calculates occupation and industry variables, and writes the final dataset.

- **Inputs:**
 - data/interm/gmd/ (from MLI-02)
 - data/interm/mli_aez_2021.rds (from MLI-03b)
 - data/raw/MLI/mli_econ_census.rds (from MLI-04)
 - Spatial data from data/interm/ (from 11-geom)
 - WorldPop API (public): working-age population rasters (15–64)
https://data.worldpop.org/GIS/AgeSex_structures/Global_2000_2020_Constrained_UNadj
- **Outputs:**
 - data/clean/MLI/ — regression-ready dataset (Parquet/RDS)
 - data/raw/MLI/worldpop/ — downloaded WorldPop rasters

vignettes/MLI-06-analysis.qmd — Regression Analysis and Output

Runs the full set of wage regressions, variance decompositions, and produces all tables and figures for the Mali section of the paper.

- **Inputs:** - data/clean/MLI/ (from MLI-05)
 - JRC GHSL API (public): built-up surface rasters
https://jeodpp.jrc.ec.europa.eu/ftp/jrc-opendata/GHSL/GHS_WUP_BUILT_S_GLOBE_R2025A/
- **Outputs** (written to output/figures/ and output/tables/): all figures and tables for the Mali section of the paper

Ecuador

The Ecuador data pipeline is implemented across four Quarto documents (vignettes/ECU-01 through ECU-05). The first two scripts are exploratory in nature — they reset and compare raw data sources, resave original raw data in the data pipeline-friendly and compact way and prepare administrative boundaries for geographical analysis. They are included in the reproducibility package for reference and documentation purposes only, and are **not executed** as part of the master 00-master.R file. Scripts 3 and 5 are the ones directly relevant to the reproducibility package and the paper.

The overall analysis is built as an R package with an appropriate `DESCRIPTION` and `NAMESPACE` files. All R functions are developed in the `R/`.

`vignettes/ECU-01-data-prep.qmd` — Raw Data Preparation

This script pre-processes the raw Ecuador household (HH) survey data and compiles it into a more compact, pre-processed format. It reads the raw ENEMDU longitudinal survey file, selects and cleans relevant variables, and geocodes respondents by mapping their coordinates to hexagons (H3) and administrative units via spatial joins. It also processes three rounds of census microdata (2001, 2010, 2022). This script is exploratory and **not part of the reproducible pipeline**.

- **Inputs:** Raw HH survey (`Household Surveys/0_enemdu2007_24.dta` , from OneDrive); variable dictionary (`data/ECU-dic.yaml`); census microdata files (from OneDrive).
- **Outputs** (written to `fldr_raw/` and `fldr_interim/`):
 - `fldr_raw/hh-survey/` — cleaned HH survey (Parquet)
 - `fldr_raw/cens-2001/` , `fldr_raw/cens-2010/` , `fldr_raw/cens-2022/` — census data (Parquet)
 - `fldr_interim/hh-dta/` , `fldr_interim/all-hh-points/` — geocoded HH data
 - `fldr_interim/point-hex-mapping/` , `point-1km-hex-mapping/` , `point-5km-hex-mapping/` — HH-to-hexagon spatial joins
 - `fldr_interim/point-adm3-v3/` , `point-adm4-v4/` — HH-to-administrative unit spatial joins

`vignettes/ECU-02-data-geom.qmd` — Administrative Boundaries and Geometries

This script processes and compares multiple versions of Ecuadorian shapefile sources (sector census boundaries, CNE shapefiles, CONALI 2022 boundaries). It constructs clean administrative geometries at levels ADM0–ADM3, comparing GMD shapefiles against survey-provided parish-level boundaries. This script is exploratory and **not part of the reproducible pipeline**.

- **Inputs:** Raw shapefiles from OneDrive (`CNE` , `sector_censal_parr` , `LIMITE_PARROQUIAL_CONALI_CNE_2022`).
- **Outputs** (written to `fldr_raw/geom/` and `fldr_interim/geom/`):
 - `fldr_raw/geom/ECU-seccens-01/{raw,adm0,adm1,adm2,adm3}/part-0.parquet` — dissolved administrative boundaries (version 1)
 - `fldr_raw/geom/ECU-seccens-02/` (and further geometry versions) — analogous outputs for additional shapefile versions
 - `fldr_interim/geom/ECU-seccens-01/` — dissolved boundaries with WKB geometry for spatial joins

`vignettes/ECU-03-data-reg.qmd` — Regression Data Assembly (*part of the reproducible pipeline*)

This script assembles the final regression-ready datasets by merging the household survey, tax/production, census, and spatial data prepared in the exploratory stage. It builds canton-level covariates (industry mix, population density, spatial characteristics) and the administrative boundary geometries used for mapping.

- **Inputs:** Pre-processed data from the exploratory stage (`ECU-01` , `ECU-02`): geocoded HH survey, census, and administrative boundary datasets; administrative subnational tax/production data.
- **Outputs** (written to `fldr_clean/`):
 - `fldr_clean/reg-dta/` — regression-ready HH-level dataset (Parquet)
 - `fldr_clean/dta-location-level-adm2/` — canton-level data (Parquet)
 - `fldr_clean/dta-adm1.geojson` , `fldr_clean/dta-adm2.geojson` — spatial boundaries for mapping

`vignettes/ECU-05-reproducibility.qmd` — Tables and Figures for the Paper (*part of the reproducible pipeline*)

This is the final output-production script. It loads all pre-compiled analytical data, runs the full set of regressions and variance decompositions using the package functions, and renders every table and figure appearing in the paper's main text and appendices.

- **Inputs** (from `fldr_clean/` , produced by `ECU-03`):
 - `fldr_clean/reg-dta/` — regression-ready HH-level dataset (Parquet)
 - `fldr_clean/dta-location-level-adm2/` — canton-level data (Parquet)
 - `fldr_clean/dta-adm1.geojson` , `fldr_clean/dta-adm2.geojson` — spatial boundaries for mapping
 - `fldr_clean/reg-fe-obs.rds` — cached regression results (written and subsequently read back within the same script)
- **Outputs** (written to `fldr_out_tbl/` and `fldr_out_fig/`):
 - **Main text figures:**
 - `fig-1` — Spatial distribution of informal workers and low-skilled workers across cantons (3-panel patchwork map: % of population, % of informal workers, % of low-skilled workers)
 - `fig-2` — Monthly wage distributions across provinces by formality status (box-plot + line, formal vs. informal)

- fig-3 — Spatial distribution of national workforce by sector (2-panel map: % construction workers, % trade workers)
- fig-4a through fig-4d — Scatter plots of location fixed effects and decomposition components (standardised) vs. mean log wage by location: (a) Location FE, (b) Skills, (c) Regional industry, d. Local observables
- fig-5 — Variance decomposition of provincial wage premiums for formal and informal workers across sub-periods (2016–2024, 2016–2019, 2020–2021, 2022–2024), without covariance terms
- fig-6 — Variance decomposition for informal workers by demographic subgroup (gender, skill level, age group), without covariance terms
- **Appendix figures:**
 - fig-a2-8 — Scatter of individual-characteristic component (standardised) vs. mean log wage by location
 - fig-a2-9 — Variance decomposition for informal workers by demographic subgroup (gender, skill level, age group), including covariance terms
 - fig-a2-10 — Variance decomposition by formality and sub-period, including covariance terms
 - fig-a2_1_rate_agr — Correlation between canton wage premiums and share of agricultural workers within region
 - fig-a2_2_rate_con — Correlation between canton wage premiums and share of construction workers within region
 - fig-a2_3_rate_edu_sec — Correlation between canton wage premiums and share of workers with secondary education
 - fig-a2_4_rate_edu_hi — Correlation between canton wage premiums and share of workers with higher education
 - fig-a2_5_indadm2_hhi — Correlation between canton wage premiums and industry concentration (HHI)
 - fig-a2_6_x21..spat_elev__1 — Correlation between canton wage premiums and elevation (log metres)
 - fig-a2_7_x21..spat_slope__1 — Correlation between canton wage premiums and slope (log degrees)
- **Main text table** (saved as .docx via `save_as_docx()`):
 - Table 3.docx — Ratio between province-averaged wages before (raw) and after controlling for worker composition, skills, industry concentration, and local endowments
- **Appendix tables** (saved as .docx via `save_as_docx()`):
 - Table A2-1.docx — Share of informal workers in the local workforce by province, urban/rural, and sector
 - Table A2-2.docx — Share of national informal workers living in each province, by urban/rural, gender, and sector
 - Table A2-3.docx — Monthly wages by industry (all workers)
 - Table A2-4.docx — Monthly wages by industry (informal workers)
 - Table A2-5.docx — Monthly wages by industry (formal workers)
 - Table A2-6.docx — Provincial wage dispersion ratios (log) by informality, sub-period, gender, age group, and skill level
 - Table A2-7-Full-Adm2.docx — Full regression coefficient table, full sample, FE Adm2, 2016–2024
 - Table A2-8-Formal-Adm2.docx — Full regression coefficient table, formal workers, FE Adm2, 2016–2024
 - Table A2-9-Informal-Adm2.docx — Full regression coefficient table, informal workers, FE Adm2, 2016–2024
- **Combined data export:**
 - all-data-ecu.xlsx — single workbook (via `openxlsx`) bundling the underlying data for every figure and table above, one worksheet per figure/table

Folder Structure

```

spat-disp-insight/
├── DESCRIPTION
├── NAMESPACE
├── LICENSE
├── CITATION.cff
├── README.md
├── WB-IGO-RIDER.md
├── renv.lock
├── .Rprofile
├── .Rbuildignore
├── .gitignore
├── .here
├── .pre-commit-config.yaml
├── R/                                ← Package functions
│   ├── f-calc-summary-stats.R
│   ├── f-calc-vars.R
│   ├── f-cleaners.R
│   ├── f-data-helpers.R
│   ├── f-data-load.R
│   ├── f-feols.R
│   ├── f-feols-format.R
│   ├── f-feols-grid.R
│   ├── f-feols-grid-helpers.R
│   ├── f-feols-helpers.R
│   ├── f-feols-summary.R
│   ├── f-feols-validate.R
│   ├── f-feols-var-decomp-plot.R
│   ├── f-fltr-frmt-reg.R
│   ├── f-format.R
│   ├── f-format-sum-stats-ft.R
│   ├── f-ft.R
│   ├── f-ft-helpers.R
│   ├── f-geom.R
│   ├── f-gg.R
│   ├── f-plot-dist.R
│   ├── f-qmd.R
│   ├── f-reg-composition-table.R
│   ├── f-regsampl.R
│   ├── f-resolve-cols.R
│   ├── f-rt2.R
│   ├── f-rt-helpers.R
│   ├── f-select-reg-vars.R
│   ├── f-variables-definitions.R
│   └── f-variables-helpers.R
├── vignettes/                        ← Analysis scripts & notebooks
│   ├── 00-master.R                  ← Master script (run first)
│   ├── 01-setup.R
│   ├── 01-setup-ECU.R
│   ├── 11-geom.qmd
│   ├── ECU-01-data-prep.qmd         ← Exploratory, not in pipeline
│   ├── ECU-02-data-geom.qmd         ← Exploratory, not in pipeline
│   ├── ECU-03-data-reg.qmd          ← Regression data assembly (part of pipeline)
│   ├── ECU-05-reproducibility.qmd   ← Final output production
│   ├── MLI-01-data-load.qmd         ← Download GMD modules via DLW API
│   ├── MLI-02-data-clean.qmd        ← Clean and standardise GMD data
│   └── MLI-03a-imputation-prep.R    ← Wage imputation data preparation

```

		MLI-03b-imp-data-annex.R	← Wage imputation & Appendix 5 table
		MLI-04-census-prep.R	← Enterprise census data preparation
		MLI-05-data-prep.qmd	← Final data preparation
		MLI-06-analysis.qmd	← Regression analysis and output
		data/	← Data (gitignored content)
		raw/	
		interm/	
		clean/	
		inst/	← Package metadata/config
		vars.yaml	
		vars-groups.yaml	
		man/	← Auto-generated documentation
		*.Rd	
		renv/	← renv environment lockfile
		activate.R	
		settings.json	
		.github/	
		dependabot.yml	
		workflows/	
		gh-pages.yml	
		release.yml	
		ISSUE_TEMPLATE/	
		bug_report.md	
		feature_request.md	
		peer_review.md	

License

The project is licensed under the **MIT License** (<https://opensource.org/licenses/MIT>) together with the World Bank IGO Rider (<https://github.com/worldbank/template/blob/main/WB-IGO-RIDER.md>). The Rider is purely procedural: it reserves all privileges and immunities enjoyed by the World Bank, without adding restrictions to the MIT permissions. Please review both files before using, distributing or contributing.