

Fusing data sources to measure multidimensional poverty

Overview

This repository contains code to reproduce the results in **Fusing data sources to measure multidimensional poverty** by Ben J. Brunckhorst, Minh C. Nguyen, Nishant Yonzan, Hai-Anh Dang & Christoph Lakner.

To replicate, follow the Instructions for Replicators below.

Data Availability

Some data cannot be made publicly available.

Data Sources

- **Filename:** `Survey_mpm_groups_true.dta`
- **Source:** Produced by the authors for this project from the World Bank Global Monitoring Database (GMD).
- **URL:** https://datacatalog.worldbank.org/int/data/dataset/0067094/multidimensional_poverty_group_statistics (restricted access)
- **Production date:** 27 August 2025
- **Access instructions:** The dataset is archived in the World Bank Development Data Hub (DDH) to facilitate reproducibility and available by request due to licensing restrictions. The Stata code used to generate the data is also archived.
- **Name:** Population data - accessed programmatically using the `pipr` R package (no data file included)
- **Source:** World Bank, Poverty and Inequality Platform
- **URL:** <https://pip.worldbank.org>, <https://github.com/worldbank/pipr>
- **Access date:** 13 May 2026
- **Citation:** World Bank (2026). Poverty and Inequality Platform (version 20260324_2021_01_02_PROD) [data set]. pip.worldbank.org. Accessed on 2026-05-13.
- **License:** Creative Commons Attribution 4.0 International license (CC-BY 4.0)
- **Access instructions:** The data is accessed programmatically using the `pipr` R package from the analysis code. No data file is necessary.

The authors of the manuscript have legitimate access to and permission to use the data used in this manuscript.

Instructions for Replicators

1. Add the input data file `Survey_mpm_groups_true.dta` to the `data/` folder if it is not included in the package.

2. Run `main.R`

- `renv::restore()` will install all required R packages at the exact versions used.
- If needed, update the `root` path to the package directory (by default it uses `here::here()` which auto-detects the project root).

List of Exhibits

The code reproduces all tables and figures in the paper presenting results from the analysis. In order of appearance:

Exhibit	Output	Script
Table 4	output/tables/mpm_summary_tables.xlsx	code/analysis/03_mpm_fusion_results.R
Figure 8	output/figures/MPM/heatplot_mpm_mae.png	code/analysis/03_mpm_fusion_results.R
Figure 9	output/figures/MPM/lineplot_mae_samplebias_national.png	code/analysis/03_mpm_fusion_results.R
Figure 10	output/figures/MPM/scatter_all_pop_sh.png	code/analysis/03_mpm_fusion_results.R
Figure 11	output/figures/MPM/scatter_all_p50_local.png	code/analysis/03_mpm_fusion_results.R
Figure 12	output/figures/MPM/errorplot_At_least_1_p_c_r_e_w_s_National.png	code/analysis/03_mpm_fusion_results.R
Figure 13	output/figures/MPM/errorplot_At_least_1_pcew_rews_National.png	code/analysis/03_mpm_fusion_results.R
Table A1	output/tables/validation_data.xlsx	code/analysis/07_validation_data.R
Table A2	output/tables/rank_correlations.xlsx	code/analysis/06_compare_rank_correlations.R
Table A3	output/tables/inf_summary_tables.xlsx	code/analysis/05_inf_fusion_results.R
Figure A2	output/figures/INF/heatplot_inf_mae.png	code/analysis/05_inf_fusion_results.R
Figure A3	output/figures/INF/lineplot_mae_samplebias_national.png	code/analysis/05_inf_fusion_results.R
Figure A4	output/figures/INF/scatter_all_pop_sh.png	code/analysis/05_inf_fusion_results.R
Figure A5	output/figures/INF/scatter_all_p50_local.png	code/analysis/05_inf_fusion_results.R
Figure A6	output/figures/INF/errorplot_At_least_1_ew_ws_Subnational.png	code/analysis/05_inf_fusion_results.R
Figure A7	output/figures/MPM/mpiw_comparison_combined_MPM.png	code/analysis/03_mpm_fusion_results.R

Figure A8	output/figures/MPM/lineplot_mae_scenarios.png	code/analysis/03_mpm_fusion_results.R
Table A4	output/tables/inf_extended_tables.xlsx sheet = "At_least_1"	code/analysis/05_inf_fusion_results.R
Table A5	output/tables/inf_extended_tables.xlsx sheet = "At_least_2"	code/analysis/05_inf_fusion_results.R
Table A6	output/tables/inf_extended_tables.xlsx sheet = "All_3"	code/analysis/05_inf_fusion_results.R
Table A7	output/tables/mpm_extended_tables.xlsx sheet = "MPM"	code/analysis/03_mpm_fusion_results.R
Table A8	output/tables/mpm_extended_tables.xlsx sheet = "MPM_AF"	code/analysis/03_mpm_fusion_results.R
Table A9	output/tables/mpm_extended_tables.xlsx sheet = "At_least_1"	code/analysis/03_mpm_fusion_results.R
Table A10	output/tables/mpm_extended_tables.xlsx sheet = "At_least_3"	code/analysis/03_mpm_fusion_results.R

Note: Unlisted tables and figures in the manuscript do not present results from the analysis.

Requirements

Software

- **R version 4.5.3 (2026-03-11)**, available at <https://cran.r-project.org/>
- Platform: aarch64-apple-darwin20
- Earlier versions of R can be used
- All package versions are managed via `renv`.
- **Key R packages:** `data.table`, `ggplot2`, `haven`, `here`, `highs`, `Matrix`, `nanoparquet`, `openxlsx`, `pipr`, `Hmisc`, `renv`, `scales`, `stringr`

Operating System

- Run on macOS 26.4.1

Runtime

- 141 minutes (macOS 26.4.1, M2 chip, 16GB memory)

Storage

- The total size of all inputs, outputs, and code is ~ 500MB
- Recommended disk space: at least 5GB

Code Description

code/analysis/ — Pipeline scripts (executed in order by main.R)

File	Purpose
01_prep_validation_data.R	Loads GMD survey data, merges population data, computes weights
02_mpm_fusion.R	Runs MPM fusion across 7 scenarios x 11 bias factors; saves fused_mpm.parquet; computes MAE, MAPE, RMSE, and coverage metrics
03_mpm_fusion_results.R	Generates tables and figures summarizing MPM fusion validation results
04_inf_fusion.R	Mirrors 02_mpm_fusion.R for infrastructure indicators (electricity, water, sanitation); includes validation metrics
05_inf_fusion_results.R	Generates tables and figures summarizing infrastructure fusion validation results
06_compare_rank_correlations.R	Generates table comparing MPM rank correlations using fusion method vs monetary poverty
07_validation_data.R	Generates table summarizing validation data (surveys) by region and decade

code/functions/ — Pure function definitions

File	Purpose
mpm_fns.R	Core poverty metric functions: <code>atleast_k_deps()</code> , <code>atleast_k_dims()</code> , <code>weighted_headcount()</code> , <code>adjusted_headcount()</code>
compute_mpm.R	Computes MPM point estimates with validity flags
fuse_indicators.R	Fuses probability distributions under conditional independence; computes Frechet-Hoeffding (lower/upper) bounds
define_scenarios.R	Parses scenario strings (e.g., "p-c-r-e-w-s") into indicator subsets
mpm_bounds.R	Solves LP problems (HiGHS) to get theoretical min/max bounds on poverty metrics
conformal_intervals_loo.R	Leave-one-out conformal prediction intervals (global and localized)
validation_metrics.R	<code>compute_val_metrics()</code> : bias, MAE, MAPE, RMSE, correlation, MPIW
run_fusion_scenarios.R	Orchestrates the loop over fusion scenarios and bias factors
run_validation.R	Shared validation pipeline: computes level and national metrics, saves outputs

results_tables.R	Shared Excel workbook generation (create_summary_workbook, create_extended_workbook)
------------------	--

Folder Structure

```

mpm-fusion/
├── main.R                                # run this to replicate all results
├── README.md
├── LICENSE
├── data/
│   └── Survey_mpm_groups_true.dta      # GMD survey microdata
├── code/
│   ├── functions/                      # Pure function definitions
│   │   ├── mpm_fns.R
│   │   ├── compute_mpm.R
│   │   ├── fuse_indicators.R
│   │   ├── define_scenarios.R
│   │   ├── mpm_bounds.R
│   │   ├── conformal_intervals_loo.R
│   │   ├── validation_metrics.R
│   │   ├── run_fusion_scenarios.R
│   │   ├── run_validation.R
│   │   └── results_tables.R
│   └── analysis/                      # Pipeline scripts
│       ├── 01_prep_validation_data.R
│       ├── 02_mpm_fusion.R
│       ├── 03_mpm_fusion_results.R
│       ├── 04_inf_fusion.R
│       ├── 05_inf_fusion_results.R
│       ├── 06_compare_rank_correlations.R
│       └── 07_validation_data.R
├── output/
│   ├── tables/                        # Excel output tables
│   └── figures/
│       ├── MPM/                      # MPM fusion figures
│       └── INF/                      # Infrastructure fusion figures

```

References

Barrett T, Dowle M, Srinivasan A, Gorecki J, Chirico M, Hocking T, Schwendinger B (2025). *data.table: Extension of `data.frame`*. R package version 1.18.2.1. URL <https://r-datatable.com>.

Fujs T, Eilertsen A, Shah R, Castañeda RA (2022). *pipr: Client for the PIP API*. R package version 0.0.3. URL <https://github.com/worldbank/pipr>.

Hahsler M, Piekenbrock M, Doran D (2019). dbscan: Fast Density-Based Clustering with R. *Journal of Statistical Software*, 91(1), 1–30. doi:10.18637/jss.v091.i01.

Huangfu Q, Hall JAJ (2018). Parallelizing the dual revised simplex method. *Mathematical Programming Computation*, 10(1), 119–142. doi:10.1007/s12532-017-0130-5.

R Core Team (2026). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL <https://www.R-project.org/>.

Schumacher D (2025). *highs: R Interface to HiGHS*. R package version 1.12.0-3. URL <https://cran.r-project.org/package=highs>.

Wickham H (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. ISBN 978-3-319-24277-4. URL <https://ggplot2.tidyverse.org>.