

Reproducibility Package README

Closing Data Gaps in Fertilizer Subsidy Analysis in Bangladesh: A Survey-to-Survey Imputation Approach

Authors: FNU Jonaed, Ivan Gachet, and Leopoldo Tornarolli

Corresponding author: FNU Jonaed (fjonaed@worldbank.org)

Date: 2026-04-02

1. Overview

This package documents the data dependencies, code files, execution order, and raw outputs used to reproduce the empirical results in the manuscript `Fertilizers survey to survey imputation.docx`.

The workflow has three stages.

- Stage 1 (Stata): harmonizes HIES 2022 and BIHS 2018–2019 and produces the source outputs for Table 1.
- Stage 2 (R): generates Figure 1, imputes fertilizer-type shares from BIHS into HIES, and produces validation and quality-control outputs.
- Stage 3 (Stata): uses the imputed HIES file together with incidence-analysis inputs to estimate direct and indirect fertilizer subsidy incidence.

This package does not currently have a single end-to-end main file. Replication should therefore follow the stage-specific entry files listed in Section 3.

2. Data Availability

Summary of availability: some data are publicly available, and some data are restricted or confidential.

Data sources

1. Household Income and Expenditure Survey (HIES) 2022 raw modules

Filename(s): raw HIES 2022 modules stored in `01. data/01. raw/HIES\_2022`

Source: Bangladesh Bureau of Statistics

URL: not publicly linked in this package

Access year: 2023

Access: Limited-access

Access instructions: replication requires authorized access to the HIES 2022 files is required.

Note: To access raw HIES 2022 data, an external user may contact Kabir Uddin Ahmed, Director, Computer Wing, Bangladesh Bureau of Statistics (kabir.ddd@gmail.com).

Citation: Bangladesh Bureau of Statistics. 2023. "Household Income and Expenditure Survey (HIES) 2022" [dataset]. Accessed 2023.

2. HIES 2022 SARMD-derived files used in the workflow

Filenames: `BGD\_2022\_HIES\_v02\_M\_v03\_A\_SARMD\_INC.dta`,

`BGD\_2022\_HIES\_v02\_M\_v03\_A\_SARMD\_IND\_full.dta`,

`BGD\_2022\_HIES\_v02\_M\_v05\_A\_SARMD\_INC.dta`,

`BGD\_2022\_HIES\_v02\_M\_v05\_A\_SARMD\_IND\_full.dta`

Source: World Bank SARMD derivatives of HIES 2022

URL: internal/restricted

Access year: 2025

Access: restricted/internal

Access instructions: replication requires authorized access.

Note: For an authorized World Bank user, these SARMD files can be accessed through Datalibweb by subscribing to the relevant survey and module (for example, BGD_2022_HIES, SARMD, IND/INC). Access requires a World Bank computer or virtual machine, Stata, the Datalibweb package, and an approved token/subscription.

Citation: World Bank. 2025. "HIES 2022 SARMD-Derived Files" [dataset]. Accessed 2025.

3. Bangladesh Integrated Household Survey (BIHS) 2018–2019

Filename(s): BIHS 2018–2019 files stored in `01. data/01. raw/BIHS\_2019`

Source: International Food Policy Research Institute

URL:

<https://dataverse.harvard.edu/dataverse/IFPRI/?q=title%3A%22Bangladesh+Integrated+Household+Survey+%28BIHS%29%22>

Access year: 2025

Access: public

Access instructions: download the BIHS 2018–2019 files from the IFPRI Dataverse release and place them in `01. data/01. raw/BIHS\_2019`.

Citation: International Food Policy Research Institute. 2025. "Bangladesh Integrated Household Survey (BIHS) 2018–2019" [dataset]. Accessed 2025.

<https://dataverse.harvard.edu/dataverse/IFPRI/?q=title%3A%22Bangladesh+Integrated+Household+Survey+%28BIHS%29%22>

4. Bangladesh Integrated Household Survey (BIHS) 2015

Filename(s): BIHS 2015 files stored in `01. data/01. raw/BIHS\_2015`

Source: International Food Policy Research Institute

URL:

<https://dataverse.harvard.edu/dataverse/IFPRI/?q=title%3A%22Bangladesh+Integrated+Household+Survey+%28BIHS%29%22>

Access year: 2025

Access: public

Access instructions: download the BIHS 2015 files from the IFPRI Dataverse release and place them in `01. data/01. raw/BIHS\_2015`.

Citation: International Food Policy Research Institute. 2025. "Bangladesh Integrated Household Survey (BIHS) 2018–2019" [dataset]. Accessed 2025.

<https://dataverse.harvard.edu/dataverse/IFPRI/?q=title%3A%22Bangladesh+Integrated+Household+Survey+%28BIHS%29%22>

5. Input-output matrix used in Stage 3

Filename: `distributional\_assessment/data/I\_O\_matrix.xlsx`

Source: General Economics Division, Planning Commission

URL: not public in this package

Access year: 2025

Access: confidential/restricted.

Access instructions: this file should be shared only through the approved verification pathway.

Note: The input-output table used in Stage 3 was received directly from Dr. Bazlul H. Khondker, who compiled the table and provided us permission to use it. Within the World Bank, it can be

used for research and knowledge generation purposes. A potential replicator may contact him directly to request access: Bazlul H. Khondker (bazlul.khondker@gmail.com).

Citation: Khondker, Bazlul H. 2025. "Input-Output Matrix" [dataset]. General Economics Division, Planning Commission. Accessed 2025.

6. Additional Stage 3 supporting inputs

Filenames: `distributional\_assessment/data/input\_prices.xlsx`,
`distributional\_assessment/data/items\_expenditure2.dta`

Source: author-compiled / project-specific supporting files

URL: not applicable

Access year: not applicable

Access: package inputs

Access instructions: these files are required for Stage 3.

2.2. Derived files created by the code

The code creates, among others, the following derived files.

- `01. data/02. intermediate/HIES\_fertilizer\_harmonized.dta`
- `01. data/02. intermediate/BIHS\_fertilizer\_harmonized.dta`
- `01. data/03. final/HIES\_fertilizer\_imputed.dta`
- `01. data/03. final/HIES\_fertilizer\_imputed.rds`
- raw table and figure outputs under `04. output/`
- incidence-analysis outputs under `distributional\_assessment/results`

2.3. Rights statement

The authors had legitimate access to all data used in the manuscript. Public data may be redistributed subject to their original terms of use. Restricted and confidential data should be shared only through the access pathway approved for reproducibility verification.

3. Instructions for Replicators

Run the package in the following order.

1. Stage 1

Run `02. code/01. stata/00\_master\_stage1\_harmonization.do`

2. Stage 2

Run `02. code/02. r/03\_figure1\_kernel\_density.R`

Then run `02. code/02. r/05\_s2s\_imputation.R`

3. Stage 3

Run `distributional\_assessment/do/00\_master\_stage3.do`

Before running on a different machine, update the top-level path in the following files.

- `00\_master\_stage1\_harmonization.do`, line 4

- `03\_figure1\_kernel\_density.R`, line 8
- `05\_s2s\_imputation.R`, line 14
- `00\_master\_stage3.do`, line 18

4. List of Exhibits

The package reproduces all main manuscript exhibits except Figure 2, which is a manually prepared conceptual schematic. Figure 5 should be checked carefully because the current code produces the underlying simulation diagnostics but does not clearly export a manuscript-ready figure with the exact caption used in the paper.

Exhibit	Raw output file(s)	Script	Note
Table 1	`Table1_summary_stats_wide.xlsx`	`04_table1_hies_bihs.do`	Source outputs for manuscript table
Figure 1	`Figure1_kernel_density_fertilizer.png`	`03_figure1_kernel_density.R`	Automatically reproduced
Table 2	`mean_ci_comparison.csv`	`05_s2s_imputation.R`	Source output for manuscript table
Table 3	`adoption_rate_by_fertilizer.csv`, `number_of_fertilizers_used.csv`	`05_s2s_imputation.R`	Source outputs for manuscript table
Figure 3	`lasso_qc_train_test_metrics.xlsx`, `Figure3 and Figure5.xlsx`	`05_s2s_imputation.R`	Code generates source outputs; final manuscript formatting have been completed outside the script
Figure 4	`lasso_qc_calibration_deciles_test.xlsx`, `Figure4.xlsx`	`05_s2s_imputation.R`	Code generates source outputs; final manuscript formatting have been completed outside the script
Figure 5	`lasso_qc_train_test_metrics.xlsx`, `Figure3 and Figure5.xlsx`	`05_s2s_imputation.R`	Needs manual check against manuscript caption
Figure 6	`mean_ci_by_AEZ.csv`, `mean_fertilizer_shares_by_AEZ.png`	`05_s2s_imputation.R`	Automatically reproduced
Figures 7–10	`histogram_urea_comparison.png`, `histogram_tsp_comparison.png`, `histogram_mop_comparison.png`, `histogram_dap_comparison.png`	`05_s2s_imputation.R`	Automatically reproduced
Figures 11–13	`distributional_assessment/results/aux_outcomes.xlsx`	`5_incidence.do`	Code exports raw incidence outputs; final chart formatting

			have been completed in Excel
Figure 2	not reproduced by code	manual	Conceptual figure

Any in-text numerical statements should be verified against the raw outputs listed above.

5. Requirements

Software requirements

- Stata 17 or above is recommended for the full package.
- R 4.3.2 or later is recommended for the R workflow.

R packages used in Stage 2

- ``data.table`, `dplyr`, `tidyr`, `ggplot2`, `haven`, `readr`, `glmnet`, `Matrix`, `foreach`, `doParallel`, `StatMatch`, `tibble`, `writexl``
- ``05_s2s_imputation.R`` installs missing R packages automatically if needed.

Stata user-written commands used in Stage 3

- ``groupfunction`` (installed by ``00_master_stage3.do`` if missing), ``winsor2``, ``quantiles``, ``distributional_assessment/do/ados/costpush.ado``

Computational notes

- Stage 2 uses parallel processing and runtime will depend on the number of available cores.
- No separate storage benchmark is provided in the package.

6. Code Description

- ``00_master_stage1_harmonization.do``
Runs the Stage 1 Stata workflow in sequence.
- ``01_harmonize_hies.do``
Builds the harmonized HIES 2022 household-level file and writes ``HIES_fertilizer_harmonized.dta``.
- ``02_harmonize_bihs.do``
Builds the harmonized BIHS 2018–2019 household-level donor file and writes ``BIHS_fertilizer_harmonized.dta``.
- ``04_table1_hies_bihs.do``
Generates the Table 1 source outputs and writes Stata and Excel versions of the summary-statistics files.
- ``03_figure1_kernel_density.R``
Generates the kernel-density comparison of total fertilizer use.
- ``05_s2s_imputation.R``
Runs the survey-to-survey imputation workflow, saves the preferred imputed HIES dataset, and writes the validation and QC outputs.
- ``00_master_stage3.do``
Runs the Stage 3 incidence workflow in sequence.

- ``1_prepare_inc.do``
Builds the auxiliary income file for Stage 3.
- ``2_referece_prices.do``
Reads ``input_prices.xlsx`` and stores the required price parameters.
- ``3_indirect_effect.do``
Uses the I-O matrix and expenditure bridge to estimate indirect effects.
- ``4_direct_effect.do``
Uses the imputed fertilizer file to estimate direct effects.
- ``5_incidence.do``
Produces the incidence outputs and exports
``distributional_assessment/results/aux_outcomes.xlsx``.

7. Folder Structure

```
fertilizer_s2s_reproducibility/
├── 01. data/
│   ├── 01. raw/
│   │   ├── HIES_2022/
│   │   ├── BIHS_2019/
│   │   └── BIHS_2015/
│   ├── 02. intermediate/
│   └── 03. final/
├── 02. code/
│   ├── 01. stata/
│   └── 02. r/
├── 03. documentation/
├── 04. output/
│   ├── 01. tables/
│   ├── 02. figures/
│   └── 03. logs/
├── distributional\_assessment/
│   ├── data/
│   ├── do/
│   │   └── ados/
│   └── results/
└── README.md
```

8. Contact

Questions about this package can be directed to FNU Jonaed at fjonaed@worldbank.org.