

Data and Code for: “Yield Gains from Balancing Fertilizer Use: Evidence from Eastern India”

Julian Arteaga and Klaus Deininger

Data and Code Availability Statement

The following list describes the datasets used as source inputs in this project. All of them are available in different sublocations of the folder “./dta/src/”.

1. Cereal Systems Initiative for South Asia (CSISA):

The paper uses public, non-confidential data from farmer-level surveys collected and harmonized between 2017 and 2019 by several research teams affiliated with the *Cereal Systems Initiative for South Asia* (CSISA), and originally compiled by (Coggins et al., 2025). The data was obtained from <https://hdl.handle.net/11529/10549105>.

Data Source	Filename	Located at
CSISA	NUE_survey_dataset_v2.csv	./dta/src/CSISA/nue_survey_multicountry/dataverse_files
CSISA	NUE_survey_dataset.csv	./dta/src/CSISA/nue_survey_multicountry/dataverse_files
CSISA	Variable_Details_v2.csv	./dta/src/CSISA/nue_survey_multicountry/dataverse_files
CSISA	Variable_Details.csv	./dta/src/CSISA/nue_survey_multicountry/dataverse_files

2. Crop Cultivation Surveys (CCS):

The paper uses public, non-confidential data from the 2006–2020 rounds of the *Crop Cultivation Surveys* (CCS) conducted by the Department of Agriculture in India. This dataset collects detailed input-use information at the farmer-crop level across all seasons over three-year rounds, each of which is publicly available at <https://eands.da.gov.in/Plot-Level-Summary-Data.htm>. Opening this webpage might require a VPN connection to a location in India. Specific files used are:

- P2006-07.xls, P2007-08.xls, ..., P2021-22.xls. Downloaded from the ‘Summary data’ tab.
- P2014-17.xls, P2017-20.xls. Downloaded from the ‘Zone/District/Villages Selected Year’ tab.
 - o These files renamed as ‘selected_villages_2014_17.xls’ and ‘selected_villages_2017_20’ in the “./dta/src/CCS” folder.

Data Source	Filename	Located at
CCS	P2006-07.xls	./dta/src/CCS
CCS	P2007-08.xls	./dta/src/CCS
CCS	P2008-09.xls	./dta/src/CCS
CCS	P2009-10.xls	./dta/src/CCS
CCS	P2010-11.xls	./dta/src/CCS
CCS	P2011-12.xls	./dta/src/CCS
CCS	P2012-13.xls	./dta/src/CCS
CCS	P2013-14.xls	./dta/src/CCS
CCS	P2014-15.xls	./dta/src/CCS
CCS	P2015-16.xls	./dta/src/CCS
CCS	P2016-17.xls	./dta/src/CCS
CCS	P2017-18.xlsx	./dta/src/CCS
CCS	P2018-19.xlsx	./dta/src/CCS

CCS	P2019-20.xlsx	./dta/src/CCS
CCS	P2020-21.xlsx	./dta/src/CCS
CCS	P2021-22.xlsx	./dta/src/CCS
CCS	selected_villages_2011_14.xls	./dta/src/CCS
CCS	selected_villages_2014_17.xlsx	./dta/src/CCS
CCS	selected_villages_2017_20.xlsx	./dta/src/CCS
CCS	selected_villages_2020_23.xlsx	./dta/src/CCS

3. Socioeconomic high-resolution Rural-Urban Geographic Platform for India (*SHRUG*):

The paper uses public, non-confidential data on village-level shapefiles and administrative codes constructed by the Socioeconomic high-resolution Rural-Urban Geographic Platform for India (*SHRUG*) project (Asher et al., 2021). These shapefiles were obtained at https://www.devdatalab.org/shrug_download/. In this webpage, open the “Open Polygons and Spatial Statistics” module tab. The .shp files used correspond to files:

- PC11 Village Polygons (“2011 population census village-level geometries”)
- Shrid Polygons (“Shrid-level polygon geometries and data quality fields”)

Data Source	Filename	Located at
SHRUG	village_modified.cpg	./dta/src/SHRUG/shrug-pc11-village-poly-shp
SHRUG	village_modified.dbf	./dta/src/SHRUG/shrug-pc11-village-poly-shp
SHRUG	village_modified.prj	./dta/src/SHRUG/shrug-pc11-village-poly-shp
SHRUG	village_modified.shx	./dta/src/SHRUG/shrug-pc11-village-poly-shp
SHRUG	shrid_loc_names.dta	./dta/src/SHRUG/shrug-shrid-keys-dta
SHRUG	shrid1_shrid2_key.dta	./dta/src/SHRUG/shrug-shrid-keys-dta
SHRUG	shrid2_spatial_stats.dta	./dta/src/SHRUG/shrug-shrid-keys-dta
SHRUG	viirs_2023_7_5_500_ua_shrid2_key.dta	./dta/src/SHRUG/shrug-shrid-keys-dta

4. Location of all major international ports in India:

The paper uses public, non-confidential information on the location of all major international port locations in India, obtained from the 2015 version of the *Manual on Port Statistics* produced by the Transport Research Wing of the Ministry of Road Transport & Highways (GoI). It can be accessed at: <https://shipmin.gov.in/sites/default/files/MANUAL%202015.pdf>. Coordinates of all major ports are found in pages 6-7 of the document.

Data Source	Filename	Located at
IV	seaport location MANUAL 2015.pdf	./dta/src/IV
IV	major_port_locations.csv	./dta/src/IV
IV	major_port_locations.xlsx	./dta/src/IV

5. Location of all major urea manufacturing plants in India:

The paper uses public, non-confidential information on the location of all major urea manufacturing plants in India, obtained from Unstarred Question No. 2172 from the Ministry of Chemicals and Fertilizers to the Lok Sabha on 07/29/2022. The document can be accessed at <https://sansad.in/getFile/loksabhaquestions/annex/179/AU2172.pdf?source=pqals>. Information on State,

name of producer, and manufacturing unit location for all major processing plants are found in pages 3-4 of the document. Each individual location name was then georeferenced using Google Maps.

Data Source	Filename	Located at
IV	fertilizer_manufacturing_plants_location.csv	./dta/src/IV
IV	fertilizer_manufacturing_plants_location.xlsx	./dta/src/IV
IV	fertilizer_manufacturing_plants_location_ureaonly.csv	./dta/src/IV

6. Open Source Routing Machine (*OSRM*)

The paper uses public, non-confidential information on the minimum road distance between sample villages and both main ports and urea plants, computed from the Open Source Routing Machine (*OSRM*) <https://project-osrm.org/>, an online open source geo-localization platform. Data from the platform was requested on 12/13/2024 and 02/24/2025 and results are available in the “./dta/src/IV” subfolder as (.shp) shape files.

Data Source	Filename	Located at
IV	major_ports.cpg	./dta/src/IV
IV	major_ports.dbf	./dta/src/IV
IV	major_ports.prj	./dta/src/IV
IV	major_ports.qmd	./dta/src/IV
IV	major_ports.shp	./dta/src/IV
IV	major_ports.shx	./dta/src/IV
IV	route_plant_ccsvill.dbf	./dta/src/IV
IV	route_plant_ccsvill.prj	./dta/src/IV
IV	route_plant_ccsvill.shp	./dta/src/IV
IV	route_plant_ccsvill.shx	./dta/src/IV
IV	route_plant_nuevill.dbf	./dta/src/IV
IV	route_plant_nuevill.prj	./dta/src/IV
IV	route_plant_nuevill.shp	./dta/src/IV
IV	route_plant_nuevill.shx	./dta/src/IV
IV	route_ports_ccsvill.dbf	./dta/src/IV
IV	route_ports_ccsvill.prj	./dta/src/IV
IV	route_ports_ccsvill.shp	./dta/src/IV
IV	route_ports_ccsvill.shx	./dta/src/IV
IV	route_ports_nuevill.dbf	./dta/src/IV
IV	route_ports_nuevill.prj	./dta/src/IV
IV	route_ports_nuevill.shp	./dta/src/IV
IV	route_ports_nuevill.shx	./dta/src/IV
IV	urea_plants.cpg	./dta/src/IV
IV	urea_plants.dbf	./dta/src/IV
IV	urea_plants.prj	./dta/src/IV
IV	urea_plants.qmd	./dta/src/IV
IV	urea_plants.shp	./dta/src/IV
IV	urea_plants.shx	./dta/src/IV
IV	India_Country_Boundary.cpg	./dta/src/IV/India_states_shp
IV	India_State_Boundary.dbf	./dta/src/IV/India_states_shp
IV	India_Country_Boundary.dbf	./dta/src/IV/India_states_shp
IV	India_State_Boundary.prj	./dta/src/IV/India_states_shp
IV	India_Country_Boundary.prj	./dta/src/IV/India_states_shp
IV	India_State_Boundary.sbn	./dta/src/IV/India_states_shp
IV	India_Country_Boundary.sbn	./dta/src/IV/India_states_shp
IV	India_State_Boundary.sbx	./dta/src/IV/India_states_shp

IV	India_Country_Boundary.sbx	./dta/src/IV/India_states_shp
IV	India_State_Boundary.shp	./dta/src/IV/India_states_shp
IV	India_Country_Boundary.shp	./dta/src/IV/India_states_shp
IV	India_State_Boundary.shp.xml	./dta/src/IV/India_states_shp
IV	India_Country_Boundary.shp.xml	./dta/src/IV/India_states_shp
IV	India_State_Boundary.shx	./dta/src/IV/India_states_shp
IV	India_Country_Boundary.shx	./dta/src/IV/India_states_shp
IV	India_State_Boundary.cpg	./dta/src/IV/India_states_shp

7. ICRISAT-TCI District-Level Database (DLD):

The paper uses public, non-confidential information on of fertilizer application rates by nutrient at the district level from the *ICRISAT-TCI* District-Level Database (DLD) for Indian Agriculture and Allied Sectors. The dataset can be accessed at <http://data.icrisat.org/dld/index.html>. Specific files used are:

- Land utilization national level 1950-2021: <http://data.icrisat.org/dld/src/additional.html> (All India tab - Land Utilization file)
- Fertilizer consumption national level 1950-2020: <http://data.icrisat.org/dld/src/additional.html> (All India tab - Fertilizer consumption file)
- Fertilizer consumption state level 1966-2020: <http://data.icrisat.org/dld/src/additional.html> (State tab - Fertilizer consumption file)
- Fertilizer use district level 1966-2016: <http://data.icrisat.org/dld/src/inputs.html> under options
i) Select Categories: inputs; ii) Sub Category: Fertilizer consumption; iii) Unapportioned districts; iv) All available states, districts, elements, and items; v) years: All years.

Data Source	Filename	Located at
ICRISAT_TCI	crop-production-district_unapportioned.csv	./dta/src/ICRISAT_TCI
ICRISAT_TCI	fertilizer-consumption-district_until2016.csv	./dta/src/ICRISAT_TCI
ICRISAT_TCI	fertilizer-consumption-national.xlsx	./dta/src/ICRISAT_TCI
ICRISAT_TCI	fertilizer-consumption-state.xlsx	./dta/src/ICRISAT_TCI
ICRISAT_TCI	land-use-district_unapportioned.csv	./dta/src/ICRISAT_TCI
ICRISAT_TCI	land-utilization-national.xlsx	./dta/src/ICRISAT_TCI

8. Fertilizer Association of India (FAI):

The paper uses public, non-confidential data on national level production and consumption of fertilizers by nutrient type from the Fertilizer Association of India (FAI). The data is available at <https://www.faidelhi.org/statistics/statistical-database>. The specific files used are:

- Fertilizer production: <https://www.faidelhi.org/general/prodn-np.pdf>
- Fertilizer consumption: <https://www.faidelhi.org/general/con-npk.pdf>

Data Source	Filename	Located at
FAIDELHL	con-npk.xlsx	./dta/src/FAIDELHI
FAIDELHL	prodn-np.xlsx	./dta/src/FAIDELHI

9. Rejuvenating Watersheds for Agricultural Resilience through Innovative Development (REWARD):

The paper uses anonymized data from a baseline survey collected in the state of Odisha as part of the impact evaluation of the *Rejuvenating Watersheds for Agricultural Resilience through Innovative Development* (REWARD) watershed management project. For more information on the *REWARD* project visit: https://rewardiiswc.in/about_reward.php, and <https://projects.worldbank.org/en/projects-operations/project-detail/P172187>.

Data Source	Filename	Located at
REWARD	reward_estimation_sample_allvars.dta	./dta/src/REWARD

Computational requirements

Software: The majority of files were run on Stata 18.5. A few files were run on R (v. 4.4.1) and Python (v. 3.11.7). Individual packages might have to be installed before running scripts.

OS: Windows 11

CPU: Intel(R) Core(TM) Ultra 7 165U 2.10 GHz

Installed RAM: 32.0 GB

Description of programs/code

The outline of the folder structure of the replication package is as follows:

```
/projdir
| -- do/
|   | -- build/
|   | -- out/
| -- dta/
|   | -- src/
|   | -- cln/
| -- out
```

All code is in subfolder “./do/”, and is divided between code that processes raw data (“./do/build/”) and uses processed data to output results (“./do/build/”). The two scripts in “./do/” outside either of these folders are “./do/master.do”, and “./do/settings.do” which, respectively, describe the order in which files should be run and set the working directory

- *build*: All scripts in this folder process and compute data either directly from the raw data files in the “./dta/src/” subfolders or from the already processed datasets in the “./dta/cln/” subfolders. Each script in

the “./do/build” folder follows a naming convention specifying whether the file i) imports raw data, ii) processes imported data and builds clean data for estimation, iii) computes intermediary data to be used in estimation later on. Every file is also named according to the specific dataset from it relates to from the list of source data described above. All final data used for estimation and output can be obtained from running the files in the “build” folder following the numbering order. Given the computational requirements above, running all scripts included in this folder should take about 35 minutes.

In addition to describing the order in which the scripts should be run in order to build the final datasets, the “./do/master.do” do-file also documents what dataset in the “./dta/cln” folder is outputted by each script.

Two auxiliary, unnumbered, scripts, “./do/build/_aux_harmonize_village_names_ccs_shrug.do”, “./do/build/_aux_label_vars.do” are called within the main scripts and do not need to be run independently.

- *out*: Each individual script outputs a single exhibit included in the main body of the paper or in the appendix. After having built all required datasets using the scripts in the “./do/build/” folder, each of the scripts in “./do/out” can be run separately in any order. All output is saved in the “./out” folder.

Output-Exhibit Correspondence

All files outputted from the scripts in the “./do/out” folder have the name of the corresponding figure or table shown in the main body of the paper or in the appendix. A few exhibits (Figure 3, Table 3, Table A2) are split in two output files with corresponding names (fig_a, fig_b, tab_a, tab_b).

References:

- Asher, S., Lunt, T., Matsuura, R., & Novosad, P. (2021). Development Research at High Geographic Resolution: An Analysis of Night-Lights, Firms, and Poverty in India Using the SHRUG Open Data Platform. *The World Bank Economic Review*, 35(4), 845-871.
- Coggins, S., McDonald, A. J., Silva, J. V., Urfels, A., Nayak, H. S., Sherpa, S. R., . . . Craufurd, P. (2025). Data-driven strategies to improve nitrogen use efficiency of rice farming in South Asia. *Nature Sustainability*.