

README for Reproducibility Package

“Scanning for Soil Health: Assessing the Accuracy and Scalability of Innovations for In-Situ Soil Measurement”

Karan S. Shakya[‡], Leah Bevis[±], and Sydney Gourlay[§]

Contents

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

Overview

This reproducibility package contains code and (where access is permitted) data needed to replicate the results presented in the working paper. The analysis was carried out in Stata (version 19.5) and R (via RStudio version 2025.05.01).

All replication materials are contained in the folder “Replication – Excluding Confidential Data”, organized into “Codes”, “Data”, and “Output” subfolders.

Data Availability

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

The raw data underlying this analysis cannot be made publicly available. While some of the raw data – including the complete set of survey microdata files and soil analysis results - will be made available on the World Bank’s Microdata Catalog following an embargo period to allow for the several ongoing research outputs to be completed, the geospatial coordinates of the agricultural plots being analyzed cannot be disseminated due to confidentiality requirements. This prevents public users from being able to extract data from the geospatial soil products at plot locations. Given this, to allow for replication of the analysis conducted in this paper, key variables extracted from the geospatial products utilized in this analysis

[‡] Department of Environmental Studies and Economics, Kenyon College; shakya1@kenyon.edu.

[±] Department of Agricultural, Environmental and Development Economics, The Ohio State University; bevis.16@osu.edu

[§] Living Standards Measurement Study (LSMS), Survey Unit, Development Data Group, World Bank; sgourlay@worldbank.org.

are disseminated as part of the processed dataset. The processed datasets used for analysis, which include variables constructed from confidential data as well as those derived from the survey microdata which is not yet publicly available, are made available to users in the replication package folder. The data was collected as part of the Uganda Climate, Land Area, and Soil Study (CLASS), which was designed and implemented by the World Bank, with a survey firm (EDI Global) as the implementing partner, and is owned by the World Bank. The Uganda CLASS is a sub-national methodological survey aimed at understanding the tradeoffs across different approaches to assessing soil health in the context of agricultural and household surveys.

Data Sources

- Filename: soil-complete_merge.dta; self-ranked.dta
 - Source: World Bank — Uganda Climate, Land Area, and Soil Study (CLASS). This file also includes select variables derived from [SoilGrids 2.0](#) and [iSDAsoil](#) digital soil map products (more on these below).
 - Access: Available with reproducibility package. The confidential data from which soil-complete_merge.dta and self-ranked.dta are constructed are not publicly available. The raw input files, such as survey microdata files (excluding identifying information), will be published in the World Bank Microdata Catalog following an embargo period. While the SoilGrids and iSDA soil map products are publicly available, the plot-level coordinates with which the survey data is merged with these products are strictly confidential. Therefore, the extracted SoilGrids and iSDA variables are included only as part of the processed dataset (soil-complete_merge.dta).
-
- Filename: uga_admbnda_adm2_ubos_20200824.shp
 - Date of access: April 2025
 - Source: Uganda Bureau of Statistics with Support from WHO
 - Access: The administrative boundaries are publicly available here: <https://data.humdata.org/dataset/cod-ab-uga>
-
- Filename: Uganda_lakes_rivers2005.shp
 - Source: Uganda Bureau of Statistics
 - Date of access: April 2025
 - Access: The shapefile containing the rivers and lakes of Uganda is publicly available here: <https://earthworks.stanford.edu/catalog/stanford-fh022bz4757>.
-
- Dataset name: SoilGrids version 2.0
 - Date of access: March 2025
 - Access: The SoilGrids 2.0 data are publicly available at: <https://soilgrids.org/>. This is a geospatial data product. Values for relevant variables were extracted from the SoilGrids 2.0 product based on the plot-level coordinates captured in the Uganda CLASS study. Given the confidentiality of the GPS coordinates in the Uganda CLASS study, these extracted variables

are made available rather than including the coordinates directly. The extracted variables are included in the soil-complete_merge.dta data file, with the suffix “_soilgrid”.

- Dataset name: iSDAsoil
- Date of access: May 2025
- Access: The iSDAsoil data are publicly available at: <https://www.isda-africa.com/isdasoil/>
This is a geospatial data product. Values for relevant variables were extracted from the iSDAsoil product based on the plot-level coordinates captured in the Uganda CLASS study. Given the confidentiality of the GPS coordinates in the Uganda CLASS study, these extracted variables are made available rather than including the coordinates directly. The extracted variables are included in the soil-complete_merge.dta data file, with the suffix “_isda”.

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

A. For users with access to the confidential datasets:

1. Launch R and Stata
2. In R, open and run 0_packages.R
3. In Stata, open 1-data_process.do. Update the “PROJ” global on line 10, ensuring it points to your local copy of the replication package.
4. Execute 1-data_process.do.
5. In Stata, open 2-data_merge.do. Update the “PROJ” global on line 10, ensuring it points to your local copy of the replication package.
6. Execute 2-data_merge.do.
7. In R, open 3-analysis.R
8. Update the working directory on line 6 to point to your local copy of the replication package.
9. Run 3-analysis.R
10. Outputs are written to the Outputs folder.

B. For users without access to the confidential datasets:

1. Launch R
2. In R, run 0_packages.R
3. In R, open 3-analysis.R

4. Update the working directory on line 6 to point to your local copy of the replication package.
5. Run 3-analysis.R
6. Outputs are written to the Outputs folder.

List of Exhibits

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

All tables and figures listed below are reproduced by the code, except for Table 1 and Appendix Tables A1 and A3. These tables are static, text-based tables and are not based on empirical analyses. Appendix Tables A1 and A3 are provided directly in the Outputs folder in excel-results.xlsx. Table 1 is written directly into the manuscript.

Exhibit	Output filename	Script	Notes
Table 1	N/A	N/A	Table is static text, not based on empirical analysis
Figure 1	map-district_sample.png	Codes/3-analysis.R	
Figure 2	scatterplot-wet_icraf-top.png	Codes/3-analysis.R	
Figure 3	mean-self_district.png	Codes/3-analysis.R	
Figure 4	combo-sankey_self.png	Codes/3-analysis.R	
Figure 5	cor-icraf_whole.png	Codes/3-analysis.R	
Figure 6	combo-ph1.png	Codes/3-analysis.R	
Figure 7	combo-clay1.png	Codes/3-analysis.R	
Figure 8	combo-soc1.png	Codes/3-analysis.R	
Figure 9	combo-totanit1.png	Codes/3-analysis.R	
Figure 10	combo-pot1.png	Codes/3-analysis.R	
Figure 11	combo-classification.png	Codes/3-analysis.R	
Table A1	excel-results.xlsx (worksheet "Table A1")	N/A	Table is static text, not based on empirical analysis
Table A2	wet_validation.tex	Codes/3-analysis.R	
Table A3	excel-results.xlsx (worksheet "Table A3")	N/A	Table is static text, not based on empirical analysis
Figure A1	self-three_criteria.png	Codes/3-analysis.R	
Figure A2	classification-ph.png	Codes/3-analysis.R	

Figure A3	classification-clay.png	Codes/3-analysis.R	
Figure A4	classification-soc.png	Codes/3-analysis.R	
Figure A5	classification-nit.png	Codes/3-analysis.R	
Figure A6	classification-pot.png	Codes/3-analysis.R	

Requirements

Computational and Software Requirements

- Software: Stata 19; R.
- Operating system: Any OS supported by Stata 19 and R (Windows/macOS/Linux).
- Additional Stata packages: install any user-written commands required by `01\_main\_analysis.do` as prompted by the script or documented within the do-file.
- Additional R packages: Installation and loading of all required R packages is included in the 0_packages.R script.
- Runtime and storage: Less than 15 mins; ensure sufficient write permissions and disk space for the Output folder.

Code Description

- Codes/0_packages.R
 - This R script installs and loads all relevant packages needed for running the 3_analysis.R script.
- Codes/1-data_process.do
 - This Stata do file works with the individual data sets stemming from each soil measurement method, harmonizing the naming and formatting each file such that they can be merged in the subsequent step. Select variables derived from the survey microdata files are also constructed here, such as crop yield. The input datasets used are confidential. This code is included for information but is not run as part of the package (unless the user has access to the confidential data) due to confidentiality of the underlying data. Following an embargo period, the survey microdata and soil analyses results will be released in the World Bank's microdata catalog.
- Codes/2-data_merge.do
 - This Stata do file works merges the individual data sets stemming from each soil measurement method, after they have been harmonized and restructured in 1-data_process.do. Additional variables are constructed in this file to (i) ensure the variables used in the analysis are in the same unit across soil measurement methods, (ii) that the soil variables used are reflective of the most comparable soil depths, and (iii) create categorical bins tied to each soil property of interest. The input datasets used are confidential. This code is included for information but is not run as part of the package (unless the user has access to the

confidential data) due to confidentiality of the underlying data. Following an embargo period, the survey microdata and soil analyses results will be released in the World Bank's microdata catalog.

- Codes/3-analysis.R
 - Main R script that loads the shared analysis datasets (soil-complete_merge.dta; self-ranked.dta), runs the empirical analysis, and generates all tables and figures saved in the Output folder.

Folder Structure

Codes

- ├── 0_packages.R
- ├── 1-data_process.do (not run if no access to confidential data)
- ├── 2-data_merge.do (not run if no access to confidential data)
- └── 3-analysis.R

Data

- ├── Raw - Confidential (suppressed for confidentiality)
 - └── Soil Data
- ├── Raw - Public (includes publicly available shapefiles)
- ├── Intermediate - Confidential (suppressed for confidentiality)
- └── Processed - Public

Output