Using free remotely sensed data to assess war-induced damage to agricultural cultivation: Evidence from Ukraine

JDE-version

2025-10-09

This README file provides an explanation of how to reproduce the results, including descriptive and regression tables and figures, from the paper titled "Using free remotely sensed data to assess war-induced damage to agricultural cultivation: Evidence from Ukraine," forthcoming in JDE.

Reproducibility Package

Introduction

This reproducibility package comprises essential elements that support the replication of the results described in the paper. The main components are as follows:

(i) Raw data: Raw data files from multiple sources are provided, along with necessary links and codes for downloading publicly available raw datasets that form the basis of the study.
(ii) Stata codes for processing raw datasets: The package includes Stata codes that merge data from various sources and perform comprehensive cleaning to produce analysis-ready data sets.
(iii) Stata codes to reproduce tables and figures: Dedicated Stata code files are available for each set of tables and figures, enabling users to reproduce statistical analyses and generate the corresponding outputs.
(iv) Village and Rayon shapefiles: The included shapefiles contain the required attribute fields for conducting spatial analysis and extracting data (such as winter and summer crop cover, daily temperature and precipitation) from raster maps.

Computational Requirements

Software requirements: Data cleaning and analysis are done using Stata 18. Raw data extraction and generation are done using ArcMap (used version 18.1) or QGIS as well as GEE and Python platforms.

Memory and Runtime Requirements: The Stata codes were last run on a Widows 10 laptop with 32GB of RAM and i7-1365U processor with 1.8 GHz.

File Structure

To ensure that each step of the workflow is clearly documented and reproducible, the package separates raw data, processing scripts, analysis code, and outputs into distinct folders. This organization allows users to easily follow the progression from raw data to final results, making it straightforward to rerun analyses or adapt the workflow as needed. The folder structure of the reproducibility package is outlined below:

```
/UKRCropReprod/
├── README.docx              # Project description and instructions
├── Analysis/                # Scripts and results for data analysis
│   ├── Code_JDE/                    # Analysis code files
│   └── Output_JDE/          # Output generated from analysis
├── DataManipulation/        # Raw data, manipulation and clening folders
│   ├── CleanData/           # Cleaned and processed datasets
│   ├── Code/                # Scripts for data merging and cleaning
│   ├── RawData/             # Raw, original datasets
│   └── Temp/                # Temporary or intermediate files
└── Paper/                   # Manuscript
```

- **DataManipulation/:**

This folder is dedicated to data preparation, cleaning, and management. It organizes all files and scripts needed to transform raw data into analysis-ready datasets.

- ✓ RawData/:
  Contains the original, unprocessed datasets as received from various data sources, organized into separate folders by theme and data sources. These files are not modified and serve as the starting point for all data manipulation.
- ✓ Code/:
  Includes scripts for importing raw data, merging datasets, cleaning variables, and generating intermediary files. Each script is named to indicate the dataset it processes and, when relevant, its specific function (e.g., regression data for winter and summer seasons, spatial data preparation). The recommended order for running these scripts is provided in the “master.do” file.
- ✓ CleanData/:
  Stores datasets that have been cleaned and processed, making them ready for analysis. These files are generated by running the scripts in the Code/ subfolder via the “master.do” file, and are used as input for the analysis scripts.
- ✓ Temp/:
  Contains temporary or intermediate files created during data processing. These files are used internally by the scripts and are not required for final analysis or output.

- **Analysis/:**

This directory includes all scripts and results (tables and figures) associated with the main data analysis.

- ✓ Code_JDE/:
  Contains analysis scripts that utilize cleaned and processed data to produce results (tables and figures) included in the manuscript. Each script is named according to the corresponding table or

figure caption in the manuscript and can be run independently once the data is prepared. The recommended order for running these scripts is specified in the "master.do" file.

- ✓ Output_JDE/:
  Contains all output generated by the analysis scripts, including summary statistics, regression results and figures. Each output file is labeled to match the table or figure number in the manuscript.

Data Availability

Outlined below are the sources for raw datasets, the procedures used for data generation and extraction, and, where available, the scripts utilized for data retrieval across various platforms.

Crop cover data: Annual crop maps (both for winter and summer crops) for all of Ukraine from 2019 to 2024 were produced using the machine learning methodology described by Kussul et al. (2025). Training data were collected through two rounds of in-situ sampling along main roads each year, except in 2022, when the conflict limited data collection to a single round. For each agricultural season, all available optical imagery from Sentinel-2 and Synthetic Aperture Radar (SAR) data from Sentinel-1 were used to train either a multilayer perceptron or a random forest model. These models were designed to classify specific crop types, rather than simply identifying the presence of vegetation.

The cultivated area for major agricultural crops was aggregated for both winter and summer seasons over six years for nearly 10,000 village communities (VCs) across Ukraine. This aggregated crop area serves as the primary outcome variable.

To create the winter and summer cropland maps used in this study, we used the winter cropland mask and annual crop classification maps for the target years. First, the winter cropland class for each target year was defined as areas under wheat, barley, or rapeseed (the main winter crop types) in the annual crop-specific classification map that intersected with the winter cropland mask. Other crop types, as well as wheat, barley, and rapeseed that did not intersect with the winter cropland mask (summer variants of these crops), were considered summer crops. Non-cultivated cropland was defined as cropland classified as a non-agriculture class in the annual crop-specific classification map but classified as cropland at least once within three years before the full-scale invasion. This definition is harmonized with the FAO definition of agricultural field abandonment, where active cropland fields are defined as those used for agriculture at least once in four years.

Harmonized raster crop maps (in ".tif" format) are available for the years 2019 to 2024 in the subfolder "…\DataManipulation\RawData\CropCover\Maps". To extract winter and summer crop areas, overlay these raster maps with village shapefile (ADMIN_3.shp) available at https://github.com/justinelliotmeyers/official_ukraine_administrative_boundary_shapefile/ using tools such as the QGIS zonal histogram function. [1]

All of these datasets are included in the reproducibility package, and are available on the Development Data Hub (DDH).

Damaged area due to military action based on hand-labeling: To create a granular time-varying measure of conflict-induced damage that is relevant for agriculture, we use freely available S2 imagery stacks to hand-label damage fields. This process relies on nearly one million filed boundaries generated by Sinergise, which serve as the basic unit of cultivation for each year from 2022 to 2024.The

[1] The original source of the data is OCHA Ukraine (http://www.ukrmap.com.ua/) covered under the license Creative Commons Attribution for Intergovernmental Organizations (http://creativecommons.org/licenses/by/3.0/igo/legalcode).

methodology for delineating field boundaries is described at https://area-monitoring.sinergise.com/docs/markers/field-delineation/.

Manual identification of damaged fields follows the approach outlined in Kussul et al. (2023). In nine conflict-affected oblasts—Kyiv, Chernihiv, Sumy, Kharkiv, Luhansk, Donetsk, Zaporizhja, Kherson, and Mykolaiv—trained university students labeled agriculturally suitable fields (whether cultivated or not at the time of damage) that were affected by military actions. These actions include crop burning, heavy vehicle movement, artillery shelling, and trench construction. Labeling was based on visual inspection of cloud-masked imagery for every two-week period following Russia's invasion in February 2022.

For an example of the hand-labeling process, see: "…\DataManipulation\RawData\Conflict_Damage\Example_Damages_Handlabel_Period28.docx".

All these datasets are included in the reproducibility package, and are available on the DDH.

Burned areas based on machine learning: To identify burned fields independently from other types of damage, we utilize short-wave infrared (SWIR) bands, which are sensitive to the high temperatures produced by active fires. SWIR bands with 20 m spatial resolution are available from Sentinel-2 (S2). For higher temporal frequency (several times per day) but coarser spatial resolution (375–1,000 m), we also use NASA's FIRMS system, which supports real-time wildfire monitoring and management globally.

Detection and area estimation are performed using custom code, available in a GitHub repository stored on Google Drive. Detailed instructions for use are provided in the accompanying Readme.md file. Access the repository at: https://drive.google.com/drive/folders/1Tjvr6nLpvd-uVEPbsnDHn2khWDpfPu-2?usp=drive_link and permission to access the code will be granted upon request. The methodology as well as the data generated for use in this research project are available on the DDH. The generated datasets on burned area are included in the reproducibility package.

Data from hand annotation experiments (HandAnnotation_JDE): To evaluate the reliability manually labeling damaged fields and estimating burned areas based on machine learning without ground truth data, we conducted two experiments: First, three annotators independently labeled agricultural fields in conflict-affected regions (Kharkiv, Luhansk, Donetsk, Zaporizhzhia, and Kherson). Candidate sample areas (with resulting data compiled in sample1.csv through sample5.csv) were selected using statistical thresholds for damage and fire activity, ensuring the inclusion of hotspots and a variety of damage types across different years. For each region, one cell (10 square km) was randomly selected, and each year included at least one damage hotspot among the samples. Second, a validation exercise for the year 2022 evaluated the burnt area detection algorithm across Donetsk, Luhansk, Zaporizhzhia, and Kherson. The process involved randomly sampling 1,400 polygons from both occupied and buffer zones, stratified by fire detections, FIRMS data, and field damage using standard map validation techniques. Key variables included elevation, slope, number of cloud-free images, and distance to the nearest burnt field. The data were compiled into the file "dataset_with_dist_annual.csv" and are available on the DDH.

Conflict indicators based on situational reports (public): Information regarding whether village councils were occupied, a scene of active conflict, or under Russian troop control is publicly available from daily situational reports published by local and regional military administrations at the settlement level (https://novy.tv/ru/news/2022/06/02/karta-bojovyh-dij-v-ukrayini/ and https://liveuamap.com/ru/time).

The analysis of the progression of conflict events, changes in frontlines, and the generation of village-level military activity indicators was conducted based on reports published by the Institute for the Study of War (https://understandingwar.org/research/russia-ukraine/) and the Ukrainian OSINT expert project DeepState (https://deepstatemap.live/). These reports are detailed, georeferenced, and highly credible for Ukrainian and international experts, journalists, and decision-makers. By-weekly indicators are constructed at the village council level (ADMIN_3.shp from https://github.com/justinelliotmeyers/official_ukraine_administrative_boundary_shapefile/) using these reports and can be found in the directory: "…\DataManipulation\RawData\Conflict_Public". All the datasets are included in the reproducibility package and are available on the DDH.

Incidence of armed conflict: Data on armed or inter-state organized violence is sourced from publicly available databases: the Armed Conflict Location & Event Data (ACLED) project and the Uppsala Conflict Data Program (UCDP). These datasets can be accessed at: https://acleddata.com/ and https://ucdp.uu.se/. ACLED event data for Ukraine was retrieved on April 19, 2024 (covering the period from January 1, 2018 to April 19, 2024) and again on April 12, 2025 (covering the period from April 12, 2022 to April 11, 2025) including all the daily records and data columns. The 2004 download is used only for the period until April 2022. The UCDP data was downloaded on May 15, 2025, including all the daily events (state-based violence, non-state violence, and one-sided violence) and data columns.

To associate each event from ACLED and UCDP with its corresponding village council, latitude and longitude coordinates are extracted and converted into point shapefiles using any standard GIS application. These point shapefiles are then overlaid on the village shapefile to determine village council attribution. The workflow is described in the relevant scripts within the data manipulation folder, specifically 3_4_ACLED_2024download.do, 3_4_2_ACLED_2025download.do, and 3_5_UCDP.do. The resulting ".dbf" files are uploaded to the appropriate data directories for ACLED and UCDP.

Data on primary and secondary road networks (OSM_JDE): Data on primary and secondary road networks for Ukraine were retrieved from open street map (OSM) on December 7, 2022 (https://data.humdata.org/dataset/hotosm_ukr_roads), and overlayed with village shape file to calculate distance from villages with damaged fields or conflict incidents reported by ACLED. The resulting data were compiled as "village_primary_dist.csv and "village_secondary_dist.csv and are available on the DDH.. The python codes for calculating distance are provided in the "Code" folder under the "DataManipulation" directory.

Gridded population data (WorldPop_JDE): These data are sourced from *Wroldpop* (https://hub.worldpop.org/geodata/summary?id=75850) and consist of *"gridded population estimates for Ukraine using UN COD-PS estimates 2022, version 3.0" (Bondarenko et al., 2022).* Population figures were aggregated at the village level from a 100m resolution pixel using the village shapefile, with the compiled results stored in "village_population.csv" and are available on the DDH..

Weather data: Daily precipitation and minimum/maximum temperature data at the Rayon level are extracted from the US National Center for Environmental Prediction Global Forecast System, which produces weather forecast every six hours. Google Earth Engine (GEE) scripts for data extraction available at: https://code.earthengine.google.com/7fe306e5e36e6cf154ff3ed3a60e826a for daily precipitation and https://code.earthengine.google.com/ffa4a235a2d6eb104d88adf703971795 for temperature. For this project, we extracted daily precipitation and temperature data ranging from 08/01/2018 to 12/31/2024, and data are available on the DDH. Please ensure to update the date parameters in the script as needed.

Admin codes: These datasets include: (i) village codes extracted from the village shapefile available at https://github.com/justinelliotmeyers/official_ukraine_administrative_boundary_shapefile/ (ADMIN_3.shp); (ii) rayon and village codes: this data includes village and rayon codes after overlaying village shapefile with rayon shapefile of State Service of Ukraine for Geodesy, Cartograph and Cadastre; (iii) region code includes official oblast codes (the first two digits of rayon codes) together with oblast names and region names; and (iv) occupied villages since 2014 include village codes that have been occupied by Russia since 2014 compiled based on situational reports. These datasets are included in the reproducibility package and are available on the DDH.

References

Bondarenko M., Priyatikanto R., McKeen T., Sorichetta A., Leasure D.R., and Tatem A.J. 2022 Gridded population estimates for Ukraine using UN COD-PS estimates 2022, version 3.0. WorldPop, University of Southampton. doi:10.5258/SOTON/WP00742.

Kussul, N., Drozd, S., Yailymova, H., Shelestov, A., Lemoine, G., & Deininger, K. (2023). "Assessing damage to agricultural fields from military actions in Ukraine: An integrated approach using statistical indicators and machine learning." International Journal of Applied Earth Observation and Geoinformation, 125: 103562. https://doi.org/10.1016/j.jag.2023.103562

Kussul, N., Shelestov, A., Yailymov, B., Yailymova, H., Lemoine, G., & Deininger, K. (2025). "Assessment of war-induced agricultural land use changes in Ukraine using machine learning applied to Sentinel satellite data." International Journal of Applied Earth Observation and Geoinformation, 140: 104551. https://doi.org/10.1016/j.jag.2025.104551