Data Science for Economics Research Project

Pierre Biscaye, Chaire de Professeur Junior
Email: pierre.biscaye@uca.fr

Université Clermont Auvergne
Masters 2 AED / Magistère 3

Spring 2026

Overview

In this research project, students will work in groups of 3–4 to apply tools from class and different data sources to analyze recent incidence of flooding in Nigeria. By synthesizing household surveys, administrative outlooks, web-scraped remote sensing archives, and additional geospatial data, students will build a multi-layered analysis of flood risk and incidence.

This work falls under my ongoing research agenda studying the economic impacts of floods and potential policy responses to mitigate adverse impacts. Please reach out to me if you would like to use any of the data from this project for your own research or if you are interested in potentially working with me on related research projects.

Project deliverables and evaluation

- January 26: ~5 page report including exhibits presenting and discussing the analysis for task 1 (15% of grade), and accompanying code (5% of grade).
- February 13: ~5 page report including exhibits presenting and discussing the analysis for task 2 (25% of grade), and accompanying code (10% of grade).
- March 2: ~15 page report including exhibits presenting and discussing the analysis for task 3 (25% of grade) along with a synthesis and discussion of all three tasks (10% of grade), and accompanying code for task 3 (10% of grade).

All deliverables should be written in English and submitted by email by 17h on the due date. Do not exceed the given page limit. If you have additional exhibits (tables or figures) you really think should be included you may use an appendix of no more than 2 pages to add them. The "% of grade" for each deliverable refers to the grade for this project. Work on this project will make up **50% of your final grade** in this class. Divide each component's share by two for the % of your total class grade.

The reports will be graded on both the clarity and organization of the presentation (writing and exhibits), on how well you completed the specific sub-tasks, and on the quality of the analysis. For each report, you should include an acknowledgement at the end of how you used generative AI (e.g., writing code, troubleshooting errors, drafting text, etc.). You will not be penalized for your use of generative AI, so be complete in your acknowledgement. This acknowledgement will not count against your page limit.

For each deliverable, you will submit a zip file of all code files used to produce paper exhibits, with a readme document. You do not have to use python for your analyses: use whatever tool you feel most comfortable using for each specific task. The code should be well-organized, well-comments, and structured to allow simple replication after changing local paths/directories at the top of selected code files. The readme should summarize all the code files used, the packages required to run each code file, and list the data inputs (along with information on where/how they were accessed) and the outputs (new or updated datasets, tables, and/or figures) for each file. The objective of these zip files is that a student from any other group in the class should be able to reproduce your outputs after accessing any data indicated in the readme and making minor changes to your code directories.

Note that you are expected to produce original code and reports that are the result of your own work. Any evidence of plagiarism or copying the work of others will result in a grade of 0 on the applicable deliverable.

Tasks

Task 1: Ground-Truth & Administrative Analysis

Focus: Cleaning and reconciling multi-level human flood reporting data
Key competencies: Data wrangling, cleaning, and visualization; descriptive analysis

Students will begin by establishing the "ground truth" of flood incidence using the Nigeria General Household Survey (GHSP) and the Nigeria Annual Flood Outlook (AFO) to understand how flooding is perceived and reported at the household, community, and administrative level.

Data: The GHSP data are available from the World Bank Microdata Catalog (Wave 5 is here). You will need to create a free account to be able to download the data. AFO data for 2023 can be accessed from the Project folder on the course website. "NIHSA_2023_Probables.xlsx" categorizes local government areas (LGAs) as probable or highly probable to flood in 2023 based on the Nigeria Hydrological Services Agency (NIHSA)'s flood modeling for the 2023 AFO. "NIHSA_LGA_Floods_2023.csv" lists all LGAs where flooding was reported in 2023 to the Nigeria Emergency Management Authority, as catalogued by NIHSA.

Sub-tasks:

1. Household measure: Derive 2023 flood indicators from the GHSP post-harvest household economic shocks module.
2. Community measure: Calculate the share of households reporting floods based on sub-task 1. Derive additional 2023 flood indicators from the GHSP post-harvest community survey community events module.
3. LGA measure: Clean NIHSA data to create flood risk and realization variables, and merge these with the survey data at the LGA level.
4. Correlation analysis: At the community level, analyze and discuss the consistency (or divergence) between household, community, and administrative flood reports.

Task 2: Web-Scraping & Remote Sensing

Focus: Extending datasets through automated collection, remote sensing, and spatial joins
Key competencies: data wrangling; web scraping; geospatial analysis

This task asks student to incorporate infrastructure and satellite observation data into the initial dataset. Students will investigate how physical infrastructure (dams) and pixel-based satellite detection relate to the survey flood reports analyzed in Task 1.

Data: The GHSP data are available from the World Bank Microdata Catalog (Wave 4 is here). Note that community coordinates are randomly offset by 0-5 km (keep this in mind for your discussion). Data on all dams in Nigeria including their locations and characteristics is available from dams.ng. A scraped spreadsheet of the data can be accessed from the Project folder on the course website. The NOAA-GMU VIIRS 5-day Flood Mapping (VFM) archive is available here. I suggest using the TIF files, but you are free to use other formats. The data are organized by month and day – note that not all months are available for 2023 (keep this in mind for your discussion). Information on this data source is available from Li, S., Sun, D., Goldberg, M. D., Sjoberg, B., Santek, D., Hoffman, J. P., DeWeese, M., Restrepo, P., Lindsey, S., & Holloway, E. (2018). Automatic near real-time flood detection using suomi-npp/viirs data. *Remote sensing of environment*, 204, 672–689.

Sub-tasks:

1. Proximity to dams: Merge GHSP wave 4 community coordinates (from nga_householdgeovars_y4.dta) with the GHSP wave 5 data. For each community, use the dams data to calculate the distance to the nearest dam of any type, and to the nearest dam used for flood control.
2. Web scraping: Scrape 2023 flood detection data from the VFM archive and create a shapefile of total days of detected flooding at the pixel level for all of Nigeria in 2023.
3. Calculating zonal statistics: For each survey community, calculate the share of pixels with any floods detected and the mean total days of detected flooding in the VFM within a 5 km radius.
4. Correlation analysis: Analyze the correlation between survey flood reports and satellite detection, and test whether proximity to a dam is correlated with flooding according to either source.

Task 3: Machine Learning & Spatial Prediction

Focus: Predictive modeling and feature engineering
Key competencies: Data wrangling, spatial analysis, machine learning, feature engineering, model evaluation

The final task requires students to build a predictive model. You will expand the temporal scope of the analysis (adding 2018 data) and integrate environmental covariates to predict flooding among GHSP communities. You will then use this model to predict flood incidence among earlier GHSP communities.

Data: This task uses all the above data and asks you to identify additional data sources you can incorporate in the machine learning model.

Sub-tasks:

1. Focus on the GHSP community survey flood reports. Extend the survey flood dataset by including community survey flood reports for 2018 from GHSP wave 4.
2. Extend the VFM dataset by creating shapefiles for remotely sensed flooding in 2018, and merge this with the community locations following the process in Task 2.
3. Identify and retrieve data on additional potential variables that could be useful in the analysis.
    a. Start with variables like elevation and slope from the GHSP wave 4 nga_householdgeovars_y4.dta file, making sure to collapse the data to the community level. Justify your rationale for selecting variables from this dataset.
    b. Find a dataset on rivers or bodies of water and calculate the distance from each community to the nearest body of water.
    c. Use CHIRPS or a similar dataset to measure precipitation around each community. Consider comparing to the historical precipitation data in nga_householdgeovars_y4.dta to identify precipitation deviations rather than absolute precipitation.
    d. Identify one additional variable that you can retrieve and add to the dataset that you think could improve the quality of the prediction.
4. Use a random forest model to predict community survey flood reports based on remotely sensed flooding and other variables. Discuss and justify all your modeling choices. Test and discuss the performance of your model.
5. Retrieve the coordinates and characteristics of GHSP wave 3 communities that were not included in the wave 4 and 5 samples from the World Bank Microdata Catalog. Use the model to predict 2018 and 2023 flood incidence in these communities.

<u>Synthesis Report</u>

Components:

1. Executive Summary: High-level findings on flood incidence and prediction.
2. Data: Documentation of data sources and how the data were prepared and merged for different analyses.
3. Descriptive Results: Presentation of analyses from tasks 1 and 2.
4. Flood Prediction Methodology: Documentation of logic for ML model features and other modeling decisions.
5. Flood Prediction Results: Presentation of analyses from task 3
6. Discussion: Reflections on challenges and approaches to flood detection and measurement, key predictors of flood incidence, and limitations of the analyses.
7. Conclusion: Key takeaways and suggestions for future research on flood mapping.