README for the Reproducibility Package for

A Data-Driven Approach for Early Detection of Food Insecurity in Yemen's Humanitarian Crisis

License

Materials in this reproducibility package are published under a BSD-3-Clause license.

Data availability statement

Datasets

Dataset	Publicly available	Data URL / Access instructions	Accessed in	Data files
FEWS NET Food Security data (2024)	Yes	https://datacatalog.worldbank.org/search/dataset/0064614	April 2024	FEWS February 2024 Update TrueBoundaries_04-19-24.csv
Yemen: Acute Food Insecurity Country Data	Yes	https://data.humdata.org/dataset/b70c2734-2339-4a4d-a69d-fa2bd3225156	July 2023	ipc_yem_area_wide.csv was obtained from the directly URL and manually collapsed by district and groups of months (in the column "Analysis Period") to obtain ipc_district_2018-2023.csv (included in the reproducibility package)
FAO Yemen Agricultural Calendar (2023)	No	Data was directly received from FAO and it is included in the reproducibility package. Users may contact Alemu Manni from FAO for direct data access (Alemu.Manni@fao.org).	July 2023	AgricultualCalendar_Yemen_V2.xlsx
Rainfall and SPI data	No	Data was directly received from FAO and it is included in the reproducibility package. Users may contact Alemu Manni from FAO for direct data access (Alemu.Manni@fao.org).	July 2023	dump_ye_rainfall_and_spi_data.csv
Armed Conflict Location & Event Data Project (ACLED)	Yes	https://acleddata.com Raw ACLED data cannot be redistributed. Users must review the terms of use in the license URL and download the data in the data URL. The dataset is updated weekly and it's possible that data points will not be the same as used by the authors.	July 2023	original_acled_conflict.csv (not included in the reproducibility package)
Monthly food price estimates by product and market	Yes	https://microdata.worldbank.org/index.php/catalog/4483 Data was received over email from the dataset authors along with the datasets “Monthly energy price estimates by product and market” and “Monthly currency exchange rate estimates by market” in a wide format by day and market. The dataset can be directly accessed through the dataset API URL: https://microdata.worldbank.org/index.php/api/tables/data/fcv/wld_2021_rtfp_v02_m	July 2023	world_bank_market_prices.csv
Monthly energy price estimates by product and market	Yes	https://microdata.worldbank.org/index.php/catalog/6134 Data was received over email from the dataset authors along with the datasets “Monthly food price estimates by product and market” and	July 2023	world_bank_market_prices.csv

		“Monthly currency exchange rate estimates by market” in a wide format by day and market. The dataset can be accessed through the dataset API URL: https://microdata.worldbank.org/index.php/api/tables/data/FCV/WLD_2023_RTEP_v01_M		
Monthly currency exchange rate estimates by market	Yes	https://microdata.worldbank.org/index.php/catalog/6160 Data was received over email from the dataset authors along with the datasets “Monthly energy price estimates by product and market” and “Monthly food price estimates by product and market” in a wide format by day and market. The dataset can be accessed through the dataset API URL: https://microdata.worldbank.org/index.php/api/tables/data/FCV/WLD_2023_RTFX_v01_M	July 2023	world_bank_market_prices.csv
Telegram Exchange Rate from the Yemen Economic Tracking Initiative	Yes	https://yemen.yeti.acaps.org/ Data was accessed with the API URL call: https://yemen.yeti.acaps.org/api/telegram_exchange_rates/?format=csv+%28non-paginated%29	July 2023	api_telegram_exchange_rates.csv
Yemen Normalized difference vegetation index (NDVI) at Subnational level	Yes	https://data.humdata.org/dataset/7907439a-0138-4b8e-a4cf-4f0233e0d694/resource/5a30f36a-aa4c-4261-a24b-22bda7639e4c/	July 2023	yem-ndvi-adm2-full.csv
Yemen Rainfall Indicators at Subnational Level	Yes	https://data.humdata.org/dataset/902bd76c-3f3c-42ad-8547-f9c31deb7937/resource/8ebd5130-715e-4bc9-920c-773d461b0754	July 2023	yem-rainfall-adm2-full.csv
Yemen - Baseline Assessment Round 37 (IOM)	Yes	https://dtm.iom.int/datasets/yemen-baseline-assessment-round-37	July 2023	DTM_YEMEN_Round_37_November_2018.xlsx
Yemen — Rapid Displacement Tracking Dataset (25 June - 01 July 2023) (IOM)	Yes	https://dtm.iom.int/datasets/yemen-rapid-displacement-tracking-dataset-25-june-01-july-2023	July 2023	04_RDT Dataset- 25 June 2023 - 01 July 2023 External District Level.xlsx
Yemen — Rapid Displacement Tracking Dataset (25 - 31 December 2022) (IOM)	Yes	https://dtm.iom.int/datasets/yemen-rapid-displacement-tracking-dataset-25-31-december-2022	July 2023	20230102_RDT Dataset- 25 December 2022 - 31 December 2022 External District Level.xlsx

Yemen — Rapid Displacement Tracking Dataset (02 - 08 January 2022) (IOM)	Yes	https://dtm.iom.int/datasets/yemen-rapid-displacement-tracking-dataset-02-08-january-2022	July 2023	20220109_RDT_Weekly Update- 02 January 2022 - 09 January 2022 External District Level.xlsx
Yemen — Rapid Displacement Tracking Dataset (01 January - 05 December 2020) (IOM)	Yes	https://dtm.iom.int/datasets/yemen-%E2%80%94-rapid-displacement-tracking-dataset-01-january-05-december-2020	July 2023	Yemen - Rapid Displacement Tracking Dataset (01 January - 05 Dec 2020).xlsx
Yemen — Rapid Displacement Tracking Dataset (01 Jan to 30 Nov 2019) (IOM)	Yes	https://dtm.iom.int/datasets/yemen-%E2%80%94-rapid-displacement-tracking-dataset-01-jan-02-nov-2019	July 2023	DTM Yemen_ Rapid Displacement Tracking (RDT) Dataset _01 Jan to 30 Nov 2019.xlsx
Yemen Agriculture Zone by Level-2 Administrative Area	No	Data was directly received from FAO and it is included in the reproducibility package. Users may contact Alemu Manni from FAO for direct data access (Alemu.Manni@fao.org).	July 2023	lookupagri_zone_type.csv
Yemen - Subnational Administrative Boundaries	Yes	https://data.humdata.org/dataset/cod-ab-yem	May 2023	yem_adm_govyem_cso_ochayemen_20191002_GPKG.zip was obtained from the data URL and was manually merged with other datasets of the project to produce the data files: "acled_pcodes.csv", "adm2_correction-Grid view.csv", "ipc_pcodes.csv", and "lookup/yemen_pcodes.csv". This data was also manually processed in QGIS to produce datasets with neighboring areas in the files "lookup/neighbouring_adm2_1deg.csv" and "lookup/neighbouring_adm2_2deg.csv", and to produce datasets with spatial joins in "lookup/world_bank_fews_pcodes.csv" and "lookup/yemen_adm2_population.csv".
High Resolution Population Density Maps from Facebook for Yemen	No	https://dataforgood.facebook.com/dfg/tools/high-resolution-population-density-maps Data for Yemen is not available for download in the data URL. The team received the data for Yemen from Alex Pompe from Meta (alexpompe@fb.com).	July 2023	The data was manually processed in QGIS by mapping it to geographic district data from OCHA and collapsing it at the district and date level, producing the file "lookup/yemen_adm2_population.csv". This file is not included in the reproducibility package. The file includes observations for the date January first and each district for the years: 2004 and 2009 until 2023. All the columns in the file are date, year, month, admin-1 level unit name, admin-1 level unit

				code, admin-2 level unit name, admin-2 level unit code, and estimated population
Yemen: Areas of control	Yes	https://data.humdata.org/m/dataset/yemen-areas-of-control?	June 2023	Data was manually collapsed by the district for the faction controlling the area (column "actors") and the monetary system used (column "irg_or_dfa") in June 2023 to produce the file "lookups/yemen_areas_of_control_adm2.csv", which is included in the reproducibility package.

Setup

- Make sure you have Python install (3.10 or higher)
- Make sure you have Jupyter Notebook installed
- Create a folder for this project somewhere on your local machine, we recommend using a virtual env in pip or conda. Make sure to activate it.
- If you are using a pip or conda environment, run “pip install -r requirements.txt -y” to install all required Python packages. Otherwise you will have to install them manually
- Users need to have Power BI installed to run step 6 below.

Reproducing the paper exhibits

Step 1: Create master dataset from individual datasets

- Go to the folder "countries\yemen"

- Run the file “yemen_create_jmr_master_dataset.py”, which constructs the main dataset based on all datasets in the folder "data\preparing master dataset".

- Output is a csv file called "yemen_jmr_master_dataset_model_repro.csv", which is written to the "data\preparing master dataset" folder

- You may read the excel file "Data availability statement data sources.xlsx" to see where the individual datasets are taken from, and from which dates they are.

Step 2: Threshold model

- Open the jupyter notebook "yemen_threshold_modeling_academic_report.ipynb" in the folder "countries\yemen"

Step 2.1: Initialize model

- Run the cell under the heading: "Initialize model", which reads the csv file "yemen_jmr_master_dataset_model_repro.csv" and applies some filtering to it, as well as setting an outcome indicator.

- Parameters can be set here, like 'weight', 'high_risk_cutoff' and 'escalations_only'. Leave at default (0.5, 4 and True) to get the results as in the Academic paper.

Step 2.2: Univariate analysis

- This part is split into two parts: "Run all indicator options" and "Run for specific indicator option".

- The cells under "Run all indicator options" run for each indicator group all possible indicator options.

- The cells under "Run for specific indicator option" run for the choosen optimal indicator per indicator group, and these are the input for the combined model.

Step 2.2.1: Run all indicator options

Step 2.2.1.1: Food prices all

- Runs all possible indicators for the indicator group food prices. Output is an excel file called: "high_risk_cutoff_{high_risk_cutoff}_escalations_only_{escalations_only}_food_prices.xlsx", where the parameters high_risk_cutoff and escalations_only are set in the first cell of this notebook (step 2.1). The excel file is written to the "data\threshold modeling" folder.

Step 2.2.1.2: Fuel prices all

- Runs all possible indicators for the indicator group fuel prices. Output is an excel file called: "high_risk_cutoff_{high_risk_cutoff}_escalations_only_{escalations_only}_fuel_prices.xlsx", where the parameters high_risk_cutoff and escalations_only are set in the first cell of this notebook (step 2.1). The excel file is written to the "data\threshold modeling" folder.

Step 2.2.1.3: Exchange Rate all

- Runs all possible indicators for the indicator group exchange rate. Output is an excel file called: "high_risk_cutoff_{high_risk_cutoff}_escalations_only_{escalations_only}_exchange_rate.xlsx", where the parameters high_risk_cutoff and escalations_only are set in the first cell of this notebook (step 2.1). The excel file is written to the "data\threshold modeling" folder.

Step 2.2.1.4: Displacements all

- Runs all possible indicators for the indicator group displacements. Output is an excel file called: "high_risk_cutoff_{high_risk_cutoff}_escalations_only_{escalations_only}_displacements.xlsx", where the parameters high_risk_cutoff and escalations_only are set in the first cell of this notebook (step 2.1). The excel file is written to the "data\threshold modeling" folder.

Step 2.2.1.5: Conflict all

- Runs all possible indicators for the indicator group conflict. Output is an excel file called: "high_risk_cutoff_{high_risk_cutoff}_escalations_only_{escalations_only}_conflict.xlsx", where the parameters high_risk_cutoff and escalations_only are set in the first cell of this notebook (step 2.1). The excel file is written to the "data\threshold modeling" folder.

Step 2.2.1.6: Drought all

- Runs all possible indicators for the indicator group drought. Output is an excel file called: "high_risk_cutoff_{high_risk_cutoff}_escalations_only_{escalations_only}_drought.xlsx", where the parameters high_risk_cutoff and escalations_only are set in the first cell of this notebook (step 2.1). The excel file is written to the "data\threshold modeling" folder.

Step 2.2.2: Trimming down solution space

- This cell reads each of the excel files created in the steps 2.2.1.1 - 2.2.1.6 and filters out the non-feasible solutions. The output is one excel file for each of the indicator groups called: "high_risk_cutoff_{high_risk_cutoff}_escalations_only_{escalations_only}_{indicator_group}_trimmed.xlsx". These files are written to the "data\threshold modeling" folder.

Step 2.2.3: Run for specific indicator option

Step 2.2.3.1: Food prices specific

- Plots the optimal solution, and the solution used in the combined model, for the indicator group food prices.

Step 2.2.3.2: Fuel prices specific

- Plots the optimal solution, and the solution used in the combined model, for the indicator group fuel prices.

Step 2.2.3.3: Exchange Rate specific

- Plots the optimal solution, and the solution used in the combined model, for the indicator group exchange rate.

Step 2.2.3.4: Displacments specific

- Plots the optimal solution, and the solution used in the combined model, for the indicator group displacements.

Step 2.2.3.5: Conflict specific

- Plots the optimal solution, and the solution used in the combined model, for the indicator group conflict.

Step 2.2.3.6: Drought specific

- Plots the optimal solution, and the solution used in the combined model, for the indicator group drought.

Step 3: Multivariate analysis

- This part combines the individual indicators into one model, using a GLM (Generalized Linear Model)

Step 3.1: Initialize combined model

- Selects per indicator group the optimal found solution.

- Constructs Table 3 as a csv file.

- Defines eight different models, where each model looks at a different situation (alerts only, alarms only, north-south interaction, ...)

Step 3.2: Correlation plots

- Creates three correlation matrices, which are saved in the folder "data\visuals"

Step 3.3: Kernel density plot

- Creates a kernel density visual, which is saved in the folder "data\visuals"

Step 3.4: Generalized Linear Model and Recursive Feature Selection

- First cell initializes the function that will be called by the next 8 cells

Step 3.4.1: GLM Model 1

- Outputs the Recursive Feature Elimination and the regression table for model 1

Step 3.4.2: GLM Model 2

- Outputs the Recursive Feature Elimination and the regression table for model 2

Step 3.4.3: GLM Model 3

- Outputs the Recursive Feature Elimination and the regression table for model 3

Step 3.4.4: GLM Model 4

- Outputs the Recursive Feature Elimination and the regression table for model 4

Step 3.4.5: GLM Model 5

- Outputs the Recursive Feature Elimination and the regression table for model 5

Step 3.4.6: GLM Model 6

- Outputs the Recursive Feature Elimination and the regression table for model 6

Step 3.4.7: GLM Model 7

- Outputs the Recursive Feature Elimination and the regression table for model 7

Step 3.4.8: GLM Model 8

- Outputs the Recursive Feature Elimination and the regression table for model 8

Step 4: Filling in master dataset

- Fills in the entire dataset given the selected model (model 8)

- Starts with again running for model 8

- Writes the output as a csv file called "yemen_jmr_master_dataset_model_filled_in.csv" to the folder "yemen"

- This file will be the main input for our Power Bi Dashboard

Step 5: Estimated percentage of population at risk of deterioration into IPC4+

- Ouputs 4 csv files, called: "yemen_estimated_population_in_ipc4_plus_country_{region}_{m}.csv" for region in [IRG, AA] and m in [perc_of_locf_avg, per_of_locf_hmean]. Outputs are wriiten to "data\estimated population at risk"

Step 6: Power Bi Dashboard

- The dashboard "Yemen JMR monitoring dashboard Academic Report.pbix" can be found in the folder "yemen\dashboard"

- It shows some of the figures used in the Academic Report

Exhibits in the paper

In order of occurrence in Academic report:

- Table 1: Power Bi Dashboard, sheet "Table 1"
- Figure 1: Power Bi Dashboard, sheet "Figure 1"
- Table 2: Manually constructed, not an output of code or visualised in the dashboard
- Figure 2: Power Bi Dashboard, sheets "Figure 2 left" and "Figure 2 right"
- Table 3: Created in Step 3.1, location of csv file is "data\visuals\table_3.csv", note that the order in which the indicators are shown are different
- Figure 3: Output of Step 3.3, manually save to the "data\visuals" folder as "figure_3"
- Figure 4: Power Bi Dashboard, sheet "Figure 4"
- Table 4: Manual combination of the regression tables from step 3.4.1 to 3.4.3
- Figure 5: Visuals generated be step 3.4.2 and 3.4.3. Plots are saved in "data\visuals" folder as "figure_5_model_2" and "figure_5_model_3"

- Figure 6: Visuals generated be step 3.4.7 and 3.4.8. Plots are saved in "data\visuals" folder as "figure_6_model_7" and "figure_6_model_8"
- Table 5: Manual combination of the regression tables from step 3.4.4 and 3.4.6
- Table 6: Manual combination of the regression tables from step 3.4.7 and 3.4.8 (note, per model, there are two outputs, one for Alerts and one for Alarms)
- Figure 7: Power Bi Dashboard, sheet "Figure 7"
- Figure 8: Power Bi Dashboard, sheet "Figure 8 top" and sheet "Figure 8 bottom"
- Table A 1a: Comes from Step 2.2.2. Summary of the file "high_risk_cutoff_4_escalations_only_True_food_prices_trimmed.xlsx" in the folder "data\threshold modeling". It shows the top result per indicator
- Table A 1b: Comes from Step 2.2.2. Summary of the file "high_risk_cutoff_4_escalations_only_True_food_prices_trimmed.xlsx" in the folder "data\threshold modeling". It shows the top result per method applied to basket indicators
- Table A 2: Comes from Step 2.2.2. Summary of the file "high_risk_cutoff_4_escalations_only_True_fuel_prices_trimmed.xlsx" in the folder "data\threshold modeling". It shows the top result per indicator method
- Table A 3: Comes from Step 2.2.2. Summary of the file "high_risk_cutoff_4_escalations_only_True_exchange_rate_trimmed.xlsx" in the folder "data\threshold modeling". It shows the top result per method
- Table A 4: Comes from Step 2.2.2. Summary of the files "high_risk_cutoff_4_escalations_only_True_drought_crop_calendar_trimmed.xlsx" and "high_risk_cutoff_4_escalations_only_True_drought_no_crop_calendar_trimmed.xlsx"in the folder "data\threshold modeling". It shows the top result per method.
- Table A 5: Comes from Step 2.2.2. Summary of the file "high_risk_cutoff_4_escalations_only_True_conflict_trimmed.xlsx" in the folder "data\threshold modeling". It shows the top result per method. The last entry of the table can be found in "high_risk_cutoff_4_escalations_only_True_conflict.xlsx", this row is the optimal solution for this method, however, it is filtered out in the trimming step (2.2.2), since the loss is above 0.5.
- Table A 6: Comes from Step 2.2.2. Summary of the file "high_risk_cutoff_4_escalations_only_True_displacements_trimmed.xlsx" in the folder "data\threshold modeling". It shows the top result per method. The last entry of the table can be found in "high_risk_cutoff_4_escalations_only_True_displacements.xlsx", this row is the optimal solution for this method, however, it is filtered out in the trimming step (2.2.2), since the loss is above 0.5.
- Table A 7: Compiled based on expert consultation. Those who contributed have been mentioned in the acknowledgements. The timeline provide an indicator than can be used to cross-check the performance of the alerts and alarms depicted in the visuals in this paper, see Table A 8.
- Table A 8: Power Bi Dashboard, sheet "Figure 4". We have zoomed in at the given dates per visual
- Table A 9: Comes from Step 3.2. Output is saved as a csv file called "table_A_9.csv" in the folder "data\visuals"
- Table A 10: Comes from Step 3.2. Output is saved as a csv file called "table_A_10.csv" in the folder "data\visuals"
- Table A 11: Comes from Step 3.2. Output is saved as a csv file called "table_A_11.csv" in the folder "data\visuals"