README

This document details how to rerun the analysis pipeline described in:

> "Current water contact and *Schistosoma mansoni* infection have distinct determinants: a data-driven population-based study in rural Uganda" by Fabian Reitzug, Narcis B. Kabatereine, Anatol M. Byaruhanga, Fred Besigye, Betty Nabatte, Goylette F. Chami

System requirements

This code was run on the University of Oxford high-performance Biomedical Research Computing (BMRC) computing cluster on August 24, 2024 on 1 CPU core with 24 GB RAM (approximate run time 24 hours).

The following software modules on the BMRC cluster were required:

- `R/4.1.0`
- `SQLite/3.38.3-GCCcore-11.3.0`
- `PROJ/9.0.0-GCCcore-11.3.0`
- `GEOS/3.10.3-GCC-11.3.0`
- `GDAL/3.3.0-foss-2021a`
- `rgdal/1.5-23-foss-2021a-R-4.1.0`
- `MPFR/4.1.0-GCCcore-11.3.0`

Installation guide

To run this code, installation of `R >= 4.1.0` is required.

All required R packages are loaded in `/code/prep/01_paths_pkgs.R` (any packages not installed already can be installed via the `install.package` function).

Typical install time on a normal desktop computer should be less than 30 minutes.

Directory structure

Directory name	Content	Purpose
`/dict`	data dictionary	contains variable definitions, used for selecting variable sets for the analysis
`/do`	main scripts	contains scripts to reproduce the main figures (Figs 3-9)
`/funs`	functions	contains all helper functions
`/out`	analysis outputs	contains sub-folders for saving main figures, supplements and variable selection outputs, and compile the supplement
`/path`	directory paths	contains script to define paths to the data directory

Directory name	Content	Purpose
`/prep`	data preparation scripts	contains scripts to load packages and helper function, load the data, and run/load variable selection results

Demo

Instructions to run on data

The following scripts may need to be modified to successfully run the scripts on a local computer:

- Set the working directory to the code directory using the `setwd` command in R.
- Set the directory paths so that they point to the directory where the demo data is located.

The entire analysis pipeline can be run by executing the `/code/RUN.R` script, which runs all scripts required to reproduce the results.

Expected output

- **Variable selection output**: Outputs from the variable selection process (via likelihood ratio tests and Bayesian variable selection) are saved in the `/code/out/var_sel/` directory (the variable selection is run on the confidential raw data, thus only selection outputs are publicly available).
- **Main figures**: Figs. 3-9 are written to the `/code/out/main/` directory (Fig. 1 is not created programmatically, and Fig. 2 has latitude and longitude columns and requires an external waterbody dataset that is not included with the demo data).
- **Supplementary tables**: All supplementary tables are written to the `/code/out/main/s_tabs/` directory.
- **Supplementary figures**: All supplementary figures are written to the `/code/out/main/s_figs/` directory.
- **Supplementary file**: All supplementary figures and tables are wrapped together using LaTeX (by means of the `/code/out/s_file/s_file.Rnw` script which generates a PDF saved in the same folder).

Expected run time for demo on a "normal" desktop computer

- Expected runtime of the project should be less than two hours.

Instructions for use

To run the code on a different dataset with a similar structure, the following modifications would be required:

- The `/code/prep/03_read.R` file would need to be modified to load in the desired datasets.
- All preprocessing steps for the data (subsequent to loading) should be done by scripts in the `/code/prep/` directory, which is aimed to contain all data preparation scripts.
- A data dictionary of the same format as the one saved in `/code/dict/` (in `.csv` format) would be required to label the main datasets and specify the variables which should be included in the candidate variable set (this is done in the `/code/prep/12_dict.R` and the `/code/prep/13_applylabs.R` scripts).