Data documentation for "Integrated causal-predictive machine learning models for tropical cyclone epidemiology"

Different data structures are used for the causal and predictive model components, and we describe both below. The structures are described in the context of the synthetic data provided for reproducibility, but they are analogous to the structures used in the real data analysis as well. In the master script, the causal model data are generated in step 1 ('1-create_panel_data.R') and the predictive model data are generated in step 2 ('2-create_predictors.R').

I. **Causal model data**. As in the real data, for each TC in our synthetic data we have a set of analytic treated and control counties, and a time series of 10 counts of the outcome for each county during the TC-specific study period, which are structured into a panel data matrix. Treatment occurs for treated counties at the last time point in the time series (time T). The data used in the causal inference models is contained in a list object called panel_list (stored in file 'panel_list.RData'). Each element of this list is named for a given TC and contains all the causal model data for that TC, stored within a sub-list. The sub-list is composed of the following named objects, some of which contain redundant information that is included only for convenience:

Object Name	Description
county_id	Vector of all county identifiers
county_id_trt	Vector of county identifier for treated counties only
Y0_obs	Y(0) matrix described in the manuscript with missing entries for treated counties during treatment. Ordering of counties in the rows of the matrix corresponds to the ordering of county_id.
Y0_full	Full Y(0) matrix without any missingness. Ordering of counties in the rows of the matrix corresponds to the ordering of county_id. Note: in real data this is not observed, only available for the synthetic data.
Y1_obs	Vector of observed Y(1) values for treated counties at time T. Order corresponds to ordering of county_id_trt.
iee	Vector of true IEEs for treated counties at time T. Order corresponds to ordering of county_id_trt. Note: in real data this is not observed, only available for the synthetic data.
iee_rate	Vector of true IEEs, rate variant, for treated counties during treatment. Order corresponds to ordering of county_id_trt. Note: in real data this is not observed, only available for the synthetic data.
pop_size	Vector of county population sizes (assumed constant over TC study period). Order corresponds to ordering of county_id.

pop_size_trt	Vector of county population sizes for treated counties only (assumed constant over TC study period). Order corresponds to ordering of county_id_trt.

II. **Predictive model data**. In the predictive model, the full set of posterior samples of the TC- and county-specific IEEs (rate version) output from the causal models are used as imputed outcomes, and they are regressed on two synthetic predictive variables representing county or TC features (a spline is used on one of the predictors). The IEE rate posterior samples from the causal model are exported to a file called ‘iee_postsamples.RData’, which contains a matrix called iee_rate_psamp, with posterior samples in the rows and TC-county pairs in the columns. The predictors are in a data frame stored in file 'pred.RData', with TC-county pairs in the rows and predictors in the columns. Columns in the data frame are described below:

Column Name	**Description**
tc	TC name
county_id	Exposed county id
x1	Predictor 1
x2	Predictor 2