

Producing Fast and Convenient Machine Learning Benchmarks in R with the stressor Package

The README.pdf file contains the brief descriptions of the supplementary files and folders.

The stressor.pdf file is the package vignette which contains a description of how to get started with using the **stressor** package. This includes how to install the correct version of **python** as well as a troubleshooting section.

Data

This folder contains two files `apmc.RDS`, `corn_yield.RDS`, `apmc_table.Rdata`, `yield.Rdata`, and `data_joint.RDS`.

`apmc.RDS`

This data set contains 135,714 observations of a crop presence/absence (AP) for the following crops: corn, winter wheat, soybeans, cotton, spring wheat, alfalfa, and hay. The following are the variables in the data set, all temperature variables are in degrees in Celsius:

- AP: The crop was planted/never planted in the cell from 2008 to 2019.
- B1_MEAN_ANNUAL_TEMP: Mean annual temperature
- B2_MEAN_DIURNAL_RANGE: Mean diurnal range (mean of max temp - min temp)
- B3_ISOTHERMALITY: Isothermality ($B2/B7 * 100$)
- B4_TEMP_SEASONALITY: Temperature seasonality (standard deviation * 100)
- B5_MAXTEMP_WARM: Max temperature of warmest month
- B6_MINTEMP_COLD: Min temperature of coldest month
- B7_TEMP_RANGE: Temperature annual range ($B5 - B6$)
- B8_MEANTEMP_WET: Mean temperature of wettest quarter
- B9_MEANTEMP_DRY: Mean temperature of driest quarter
- B10_MEANTEMP_WARM: Mean temperature of warmest quarter
- B11_MEANTEMP_COLD: Mean temperature of coldest quarter
- B12_TOTAL_PPT: Total (annual) precipitation
- B13_PPT_WET: Precipitation of wettest month
- B14_PPT_DRY: Precipitation of driest month
- B15_PPT_SEASONALITY: Precipitation seasonality (coefficient of variation)
- B16_PPT_WETQ: Precipitation of wettest quarter
- B17_PPT_DRYQ: Precipitation of driest quarter
- B18_PPT_WARMQ: Precipitation of warmest quarter
- B19_PPT_COLDQ: Precipitation of coldest quarter
- IRR_02: Percent irrigated in 1km pixel in 2002.
- IRR_07: Percent irrigated in 1km pixel in 2007.
- IRR_12: Percent irrigated in 1km pixel in 2012.
- IRR_17: Percent irrigated in 1km pixel in 2017.
- IRR_AVG: Average percent irrigated in 1km pixel of the years 2002, 2007, 2012, and 2017.
- IRR_ZERO: A 1 if the average is above zero, zero otherwise.
- IRR_50: A 1 if the average is above fifty, zero otherwise.
- SLOPE: Slope represents the change of elevation over a specific area.
- ELEVATION: The land height, in meters, above mean sea level.
- S_PH_H20: Subsoil pH (in water)

- **T_BS**: Topsoil Base Saturation
- **T_CAC03**: Topsoil Calcium Carbonate
- **T_CEC_CLAY**: Cation exchange capacity of the clay fraction in the topsoil
- **T_CEC_SOIL**: Cation exchange capacity (total nutrient fixing capacity of a soil; topsoil with low CEC have little resilience and cannot build up stores of nutrients).
- **T_CLAY**: Topsoil clay fraction
- **T_ESP**: Topsoil Sodidity (ESP, %)
- **T_GRAVEL**: Topsoil gravel fraction
- **T_OC**: Topsoil organic carbon
- **T_PH_H2O**: Topsoil pH (in H₂O)
- **T_REF_BULK_DENSITY**: Topsoil reference bulk density
- **T_SILT**: Topsoil silt fraction
- **T_TEB**: Topsoil TEB (cmol/kg)
- **T_ECE**: Topsoil Salinity (Elco, %)
- **T_SAND**: Topsoil sand fraction
- **age**: Average farmer age in the county, based on farmer census data.
- **chem**: Total expense of chemicals, including insecticides, herbicides, fungicides, and other pesticides, and the cost of custom application (excludes commercial fertilizer purchased), per agricultural acre; measured as total expense in USD \$ and standardized by the total number of agricultural acres operated, per county.
- **fert**: Total expense of fertilizers, including lime and soil conditioners, rock phosphate and gypsum, and the cost of custom application, per agricultural acre; measured as total expense in USD \$ and standardized by the total number of agricultural acres operated, per county.
- **gvt_prog**: Total cash receipts of government programs, per agricultural acre; measured in USD.
- **labor_expense**: Total expense of all laborers, per agricultural acre; measured as the total expense of laborers (hired, contract, and migrant) in USD \$ and standardized by the total number of agricultural acres operated, per county.
- **machinery**: Total asset value of agricultural machinery, per agricultural acre; measured as total machinery assets in USD \$ and standardized by the total number of agricultural acres operated, per county.
- **occup**: Percentage of operators in a county whose primary occupation is farming, standardized by the total number of operators in the county.
- **tenant**: Percentage of agricultural acres operated by tenants (producers who operate land they rent from others and/or land they worked on shares for others); measured as the number of agricultural acres operated by tenants and standardized by the total number of agricultural acres operated, per county.
- **acres_per_op**: Median farm size (in acres per operation) in a county.
- **exp**: Average number of years experience on present operation.
- **income**: Average farmer income in the county, based on census data.
- **insur_acres**: Percentage of agricultural acres with crop insurance; measured as the number of crop acres with insurance and standardized by the total number of acres, per county.
- **PERC_IRR**: Percentage of agricultural land in every county utilizing irrigation (includes all land irrigated by artificial/controlled means, including lagoon wastewater distributed by sprinkler or flood system).

corn_yield.RDS

This data set contains 16,983 observation of Corn Yield (YIELD). Along with 20 other variables that are similar to the `apmc.RDS` data set. The following are the variables for this data set:

- **YIELD**: Corn bushels per acre
- **YEAR**: The year the crop was recorded ranges from 2008 - 2018.
- **SDI_CDL_AG**: A measure of landscape diversity; measured as the proportional abundance of each land use category in a county and used as a relative index to compare across landscapes or the same landscape at different times. SDI increases as richness and evenness increase.
- **SLOPE**: Slope represents the change of elevation over a specific area.

- **ELEVATION**: The land height, in meters, above mean sea level.
- **PERC_IRR**: Percentage of agricultural land in every county utilizing irrigation (includes all land irrigated by artificial/controlled means, including lagoon wastewater distributed by sprinkler or flood system).
- **GDD**: an indicator of cumulative temperature exposure; the sum of maximum daily temperatures within a crop-specific tolerance range (10°C to 30°C for corn) over the growing season
- **BV2**: Mean diurnal range (mean of max temp - min temp)
- **BV4**: Temperature seasonality (standard deviation * 100)
- **BV8**: Mean temperature of wettest quarter
- **BV9**: Mean temperature of driest quarter
- **BV15**: Precipitation seasonality (coefficient of variation)
- **BV18**: Precipitation of warmest quarter
- **BV19**: Precipitation of coldest quarter
- **TP**: Sum of precipitation (in millimeters) throughout the growing season.
- **S_PH_H2O**: Subsoil pH (in water)
- **T_CEC_SOIL**: Cation exchange capacity (total nutrient fixing capacity of a soil; topsoil with low CEC have little resilience and cannot build up stores of nutrients.
- **T_REF_BULK_DENSITY**: Topsoil reference bulk density
- **T_OC**: Topsoil organic carbon
- **lon**: Longitude of county centroids
- **lat**: Latitude of county centroids

apmc_table.Rdata

This contains `mlm_apmc_cv` and `mlm_apmc_scv` both of these are `data.frame` objects that contain the predictions done from cross validation and spatial cross validation, found on lines 33 – 38 in `code.R`.

yield.Rdata

This contains `mlm_yield_cv`, `mlm_yield_scv`, and `mlm_yield_scv_latlon` all of these are `data.frame` objects that contain the predictions done from cross validation, spatial cross validation and spatial cross validation with grouping on latitude and longitude, found on lines 44 – 52 in `code.R`.

data_joint.RDS

This is a `data.frame` that contains the results from the cross validation, spatial cross validation, and the spatial cross validation with grouping on the latitude and longitude replication found in `code.R` lines 61 – 109.

Code

This folder contains a file named `code.R`. `code.R` is the R script necessary to generate all Tables and Figures included in the paper. This script has also been commented to show which chunk creates which table or figure. Note that Figure 1 and 2 from the paper were created with `tikz` pictures.