# Package 'BART.sp'

September 27, 2020

**Type** Package

**Title** Spatially Adjusted Bayesian Additive Regression Trees

**Version** 0.0.0.8999

**Author** Tianyu Zhang [aut, cre],
Howard H. Chang [aut],
Robert McCulloch [aut],
Rodney Sparapani [aut],
Robert Gramacy [aut],
Charles Spanbauer [aut],
Matthew Pratola [aut],
Martyn Plummer [ctb],
Nicky Best [ctb],
Kate Cowles [ctb]

**Maintainer** Tianyu Zhang <tianyu.zhang.18@gmail.com>

**Description**

A spatially adjusted Bayesian Additive Regression Trees (BART) model that adds a spatial residual with Matern correlation to the model and fits spatial data better.

**URL** https://github.com/Tianyu00/BART.sp/

**BugReports** https://github.com/Tianyu00/BART.sp/issues/

**License** GPL-3

**Imports** Rcpp (>= 1.0.4.6),
Matrix (>= 1.2-18),
assertthat (>= 0.2.1),
dplyr (>= 1.0.2),
fastDummies (>= 1.6.2),
fields (>= 11.4)

**LinkingTo** Rcpp, RcppArmadillo, BH

**RoxygenNote** 7.1.1

**Encoding** UTF-8

**Suggests** knitr (>= 1.29),
rmarkdown (>= 2.3),
ggplot2 (>= 3.3.2),
matrixStats (>= 0.56.0),
BART (>= 2.7),
reshape2 (>= 1.4.4)

**VignetteBuilder** knitr

**NeedsCompilation** yes

**Depends** R (>= 2.10)

# R topics documented:

pm25 . . . . . . . . . . 2
predict.wbart_sp . . . . . . . . . . 2
wbart_sp . . . . . . . . . . 3

**Index** **8**

---

`pm25` *PM2.5 values in Southern US*

---

## Description

This data set contains the pm2.5 values in Southern US for the year 2010 and multiple variates.

## Format

A data frame of 5000 rows * 16 columns.

## References

Hu X, Waller LA, Lyapustin A, Wang Y, Al-Hamdan MZ, Crosson WL, Estes Jr MG, Estes SM, Quattrochi DA, Puttaswamy SJ, Liu Y. Estimating ground-level PM2. 5 concentrations in the Southeastern United States using MAIAC AOD retrievals and a two-stage model. Remote Sensing of Environment. 2014 Jan 1;140:220-32

---

`predict.wbart_sp` *Predicting new observations with a previously fitted wbart_sp model*

---

## Description

This function is used to predict outcomes of new observations with a previously fitted model with type `wbart_sp`.

## Usage

```
## S3 method for class 'wbart_sp'
predict(
  object,
  newdata,
  newloc,
  draw_from_total_distribution,
  block = 1,
  seed = 88,
  ...
)
```

### Arguments

| | |
|---|---|
| `object` | An `object` of type `wbart_sp` (fitted from `wbart_sp` function). |
| `newdata` | A data frame of covariates to predict for. |
| `newloc` | A data frame of location information of the obervations in `newdata`. It should have 2 columns (names should be exactly the same as those of the `coordinates_train` when fitting the model). |
| `draw_from_total_distribution` | Whether sampling from the total distribution when doing prediction on `newdata`. If so, it would be slower but utilize all the location information. |
| `block` | If `draw_from_total_distribution=FALSE`, how many locations should be drawn at the same time. |
| `seed` | Setting the seed for reproducibility. |

### Value

The return is a list containing these components:

`fhat.test` a matrix of drawings of $f$ corresponding to newdata. Each row corresponds to a draw of the spatial random effect and each column corresponds to a row of `newdata`

`yhat.test` a matrix of final predictions (sum of `fhat.test` and `what.test`) corresponding to newdata. Each row corresponds to a draw of the spatial random effect and each column corresponds to a row of `newdata`

`what.test` a matrix of drawings of spatial random effect corresponding to newdata. Each row corresponds to a draw of the spatial random effect and each column corresponds to a row of `newdata`.

`unique_test_locations` A data frame of unique test locations in `newdata` (order not necessary the same as in `newdata`).

---

`wbart_sp` *Spatially Adjusted Bayesian Additive Regression Trees for continuoues outcomes*

---

### Description

`wbart_sp` is a Bayesian "sum-of-trees" model designed for spatial data. It is built upon the original BART model (see ...) with an extra spatial random effect.

For a numeric continuous outcome $y$, we have $y = f(x) + e_s + e$, where $e_s$ is the spatial random effect and $e\ N(0, sigma^2)$. See ...

### Usage

```
wbart_sp(
  x_train,
  y_train,
  coordinates_train,
  coordinates_system,
  x_test = matrix(0, 0, 0, 0),
  coordinates_test = NULL,
  sparse = FALSE,
  theta = 0,
```

```
  omega = 1,
  a = 0.5,
  b = 1,
  rho = NULL,
  augment = FALSE,
  xinfo = matrix(0, 0, 0),
  usequants = FALSE,
  cont = FALSE,
  rm.const = TRUE,
  sigest = NA,
  sigdf = 3,
  sigquant = 0.9,
  k = 2,
  power = 2,
  base = 0.95,
  sigmaf = NA,
  lambda = NA,
  fmean = mean(y_train),
  w = rep(1, length(y_train)),
  ntree = 200L,
  numcut = 100L,
  ndpost = 1000L,
  nskip = 100L,
  keepevery = 1L,
  nkeeptrain = ndpost,
  nkeeptest = ndpost,
  nkeeptestmean = ndpost,
  nkeeptreedraws = ndpost,
  printevery = 100L,
  transposed = FALSE,
  logrange_select_sd = 0.3,
  logsmoothness_select_sd = 0.3,
  sigma2_prior_a = 10,
  sigma2_prior_b = 1,
  tau2_prior_a = 1,
  tau2_prior_b = 1,
  logrange_init = 0,
  logsmoothness_init = 0,
  tau2_init = 1,
  logrange_prior_mean = 1,
  logrange_prior_sd = 0.5,
  logsmoothness_prior_mean = 0,
  logsmoothness_prior_sd = 0.5,
  mc,
  mc.cores = 2,
  draw_from_total_distribution = TRUE,
  block = 50,
  seed = 88
)
```

## Arguments

x_train
: Explanatory variables for training (in sample) data. Must be a data frame, with rows corresponding to observations and columns to variables. If a variable is a factor in a data frame, it is replaced with dummies. Note that q dummies are created if q>2 and one dummy is created if q=2, where q is the number of levels of the factor. Location information can be either in `x_train` or not but must be in `coordinates_train`.

y_train
: Continuous dependent variable for training (in sample) data. Must be a vector whose length equals to the number of rows in `x_train`.

coordinates_train
: The location information of observations in `x_train`. Must be a dataframe of 2 columns and the same number of rows as `x_train`. If latitude and longitude are provided as location information, the 2 columns of `coordinates_train` must be named exactly 'lon' and 'lat' (order matter) and the argument `coordiantes` set as 'lonlat'. If locations information on the ground are provided, the 2 columns must be named exactly 'x' and 'y' (order matters) and argument `coordiantes` set as 'ground'. The distane between locations is calculated accordingly (see `coordinates_system`).

x_test
: Explanatory variables for test (out of sample) data. Should have same structure as `x_train` (Must be a data frame, with rows corresponding to observations and columns to variables). If provided, must also provide `coordinates_test`.

coordinates_test
: The location information of observations in `x_test`. Should have same structure as `coordinates_train`. It must be of the same kind of coordinate system as `coordinates_train` and named exactly the same as `coordinates_train` (order matters).

logrange_select_sd
: Spatial residual sampling parameter. logRange select SD in mcmc. (see ...)

logsmoothness_select_sd
: Spatial residual sampling parameter. logSmoothness select SD in mcmc. (see ...)

sigma2_prior_a
: Prior paramter for the random noise $e$. (see ...)

sigma2_prior_b
: Prior paramter for the random noise $e$. (see ...)

tau2_prior_a
: Prior paramter for the matern correlation function tau2 (see ...)

tau2_prior_b
: Prior paramter for the matern correlation function tau2 (see ...)

logrange_init
: Initial value for logrange in mcmc.

logsmoothness_init
: Initial value for logsmoothness in mcmc.

tau2_init
: Initial value for tau2 in mcmc.

logrange_prior_mean
: Prior paramter for the matern correlation function logrange mean (see ...)

logrange_prior_sd
: Prior paramter for the matern correlation function logrange sd (see ...)

logsmoothness_prior_mean
: Prior paramter for the matern correlation function logsmoothness mean (see ...)

logsmoothness_prior_sd
: Prior paramter for the matern correlation function logsmoothness sd (see ...)

`mc` Whether fitting the model in parallel. (which usually improves the model performance but requires multiple cores.) Please also set the number of threads in argument `mc.cores`.

`mc.cores` How many threads to use if fitting the model in parallel. If `mc=FALSE`, this argument does not matter. If `mc=TRUE`, how many threads to use.

`draw_from_total_distribution`
If draw from total distribution or ? distribution in the prediction. If no `x_test`, does not matter. Usually it would be slower but perserving and utilizing all the location information in the testing dataset to set `draw_from_total_distribution=TRUE` instead of `FALSE`.

`block` The spatial random effect of how many locations to predict at one time if `draw_from_total_distrib`
If `draw_from_total_distribution=FALSE`, `block` is not used.

`coordiantes_system`
What the `coordinates_train` are. Must be either 'lonlat' or 'ground'. If `coordinates_train` is the longitude and latitude information, `coordinates_train` should be 'lonlat' and the distance is calculated using grand circle distance with unit km. If `coordinates_train` is the location information on the ground, `coordinates_train` should be 'ground' and the distance is calculated using Euclidean distance.

## Details

`wbart_sp` is the only function (besides S3 method `predict.wbart_sp`) provided by this package.

`wbart_sp` implements the spatially adjusted Bayesian Additive Regression Trees (in single thread or multiple threads). S3 method `predict.wbart_sp` implements the prediction of a model of class wbart_sp.

The detailed information about the model please see: paper/github ...

## Value

`wbart_sp` returns an object of type `wbart_sp` which is a list. It has the following components:

`fhat.train` A matrix with ndpost rows and nrow(x_train) columns. Each row corresponds to a draw $f^*$ from the posterior of $f$ and each column corresponds to a row of x_train. The $(i, j)$ value is $f^*(x)$ for the $i^{th}$ kept draw of $f$ and the $j^{th}$ row of x.train. Burn-in is dropped. NOTICE: this is the not final prediction value, `yhat.train` is.

`fhat.test` Same as `fhat.train` but now the x's are the rows of the test data.

`yhat.train` A matrix with ndpost rows and nrow(x_train) columns. Each row corresponds to the final prediction (sum of a draw from $f(x)$ and a draw of the spatial random effect) and each column corresponds to a row of x_train.

`yhat.test` Same as `yhat.train` but now the x's are the rows of the test data.

`what.train` A matrix with ndpost rows and nrow(x_train) columns. Each row corresponds to a draw of the spatial random effect and each column corresponds to a row of x_train.

`what.test` Same as `what.train` but now the x's are the rows of the test data.

`sigma` post burn in draws of sigma, length = ndpost.

`sigma_all` A data frame of burn in draws and post burn in draws of sigma, dim = (nskip + ndpost/mc.cores) * (`mc.cores`). Can be used to inspect convergence.

`tau2` post burn in draws of sigma, length = ndpost.

`logrange` post burn in draws of logrange, length = ndpost.

logsmoothness post burn in draws of logsmoothness, length = ndpost.

nskip nskip

ndpost ndpost

mu mean of y_train

varcount a matrix with ndpost rows and nrow(x_train) columns. Each row is for a draw. For each variable (corresponding to the columns), the total count of the number of times that variable is used in a tree decision rule (over all trees) is given.

varprob a matrix with ndpost rows and nrow(x_train) columns. Each row is for a draw. For each variable (corresponding to the columns), the probability (frequency / total frequency) that variable is used in a tree decision rule (over all trees) is given.

sigest The rough error standard deviation ($\sigma$) used in the prior.

coordinates_system Coordiantes parameters for coordinates_train.

unique_train_lcations Unique locations in the training data.

unique_w sampled spatial random effects according for unique_train_locations.

proc.time processing time

## References

Chipman, H., George, E., and McCulloch R. (2010) Bayesian Additive Regression Trees. *The Annals of Applied Statistics*, **4,1**, 266-298 <doi:10.1214/09-AOAS285>.

Chipman, H., George, E., and McCulloch R. (2006) Bayesian Ensemble Learning. Advances in Neural Information Processing Systems 19, Scholkopf, Platt and Hoffman, Eds., MIT Press, Cambridge, MA, 265-272.

Friedman, J.H. (1991) Multivariate adaptive regression splines. *The Annals of Statistics*, **19**, 1–67.

Linero, A.R. (2018) Bayesian regression trees for high dimensional prediction and variable selection. *JASA*, **113**, 626–36.

# Index

∗ **datasets**
pm25, 2

pm25, 2
predict.wbart_sp, 2

wbart_sp, 3