Package 'SpatialML'

May 9, 2019

Version 0.1.3

Date 2019-05-09

Type Package

Title Spatial Machine Learning

Author Stamatis Kalogirou [aut, cre],
Stefanos Georganos [aut, ctb]

Maintainer Stamatis Kalogirou <stamatis@lctools.science>

Depends R (>= 3.5.0), randomForest (>= 4.6-14)

Description Implements a spatial extension of the random forest algorithm (Georganos et al. (2019) <doi:10.1080/10106049.2019.1595177>). Future updates include more local machine learning methods as well as a geographically weighted random forest.

License GPL (>= 2)

Encoding UTF-8

LazyData true

URL http://lctools.science/

NeedsCompilation no

Repository CRAN

Date/Publication 2019-05-09 15:10:07 UTC

R topics documented:

grf 2
Income 4
predict.grf 5
random.test.data 7

Index **8**

`grf` *Geographical Random Forest*

Description

This function refers to a geographical (local) version of the popular Random Forest algorithm.

Usage

```
grf(formula, dframe, bw, kernel, coords, ntree=500, mtry=NULL,
          importance=TRUE, forests = TRUE)
```

Arguments

`formula`	the local model to be fitted using the same syntax used in the `randomForest` function of the R package `randomForest`. This is a string that is passed to the sub-models' `randomForest` function. For more details look at the class `formula`.
`dframe`	a numeric data frame of at least two suitable variables (one dependent and one independent)
`bw`	a positive number that may be an integer in the case of an "adaptive kernel" or a real in the case of a "fixed kernel". In the first case, the integer denotes the number of nearest neighbours, whereas in the latter case the real number refers to the bandwidth (in meters if the coordinates provided are Cartesian).
`kernel`	the kernel to be used in the regression. Options are "adaptive" or "fixed".
`coords`	a numeric matrix or data frame of two columns giving the X,Y coordinates of the observations
`ntree`	an integer referring to the number of trees to grow for each of the local random forests.
`mtry`	Number of variables randomly sampled as candidates at each split. Note that the default values is p/3, where p is number of variables in the formula
`importance`	Feature importance of the dependent variables used as input at the random forest. The measures used are the Mean Increase in Mean Squared Error (incMSE) if a predictor would be randomly permuted or the decrease in node impurities (IncNodePurity) from splitting on the variable, averaged over all trees. Both measures are derived from the Out of Bag (OOB) error.
`forests`	a option to save and export (TRUE) or not (FALSE) all the local forests

Details

Geographical Random Forest (GRF) is a spatial analysis method using a local version of the famous Machine Learning algorithm. It allows for the investigation of the existence of spatial non-stationarity, in the relationship between a dependent and a set of independent variables. The latter is possible by fitting a sub-model for each observation in space, taking into account the neighbouring observations. This technique adopts the idea of the Geographically Weighted Regression, Kalogirou (2003). The main difference between a tradition (linear) GWR and GRF is that we can model non-stationarity coupled with a flexible non-linear model which is very hard to overfit due to its bootstrapping nature, thus relaxing the assumptions of traditional Gaussian statistics. Essential it was designed to be a bridge between machine learning and geographical models, combining inferential and explanatory power. Additionally, it is suited for datasets with numerous predictors, due to the robust nature of the random forest algorithm in high dimensionality.

Value

`Locations` a numeric matrix or data frame of two columns giving the X,Y coordinates of the observations

`Local.Pc.IncMSE` a numeric data frame with the local feature importance (IncMSE) for each predictor in each local random forest model

`Local.IncNodePurity` a numeric data frame with the local IncNodePurity for each predictor in each local random forest model

`LGofFit` a numeric data frame with residuals and local goodness of fit statistics (training and OOB).

`Forests` all local forests.

`lModelSummary` Local Model Summary and goodness of fit statistics (training and OOB).

Warning

Large datasets may take long to calibrate. A high number of observations may result in a voluminous forests output.

Note

This function is under development. There should be improvements in future versions of the package `SpatialML`. Any suggestion is welcome!

Author(s)

Stamatis Kalogirou <stamatis@lctools.science>, Stefanos Georganos <sgeorgan@ulb.ac.be>

References

Stefanos Georganos, Tais Grippa, Assane Niang Gadiaga, Catherine Linard, Moritz Lennert, Sabine Vanhuysse, Nicholus Odhiambo Mboga, Eléonore Wolff & Stamatis Kalogirou (2019) Geographical Random Forests: A Spatial Extension of the Random Forest Algorithm to Address Spatial Heterogeneity in Remote Sensing and Population Modelling, Geocarto International, DOI: 10.1080/10106049.2019.1595177

See Also

`predict.grf`

Examples

```
data(Income)
Coords<-Income[ ,1:2]
grf <- grf(Income01 ~ UnemrT01 + PrSect01, dframe=Income, bw=60,
          kernel="adaptive", coords=Coords)
```

Income *Mean household income at lcoal authorities in Greece in 2011*

Description

Municipality centroids and socioeconomic variables aggregated to the new local authority geography in Greece (Programme Kallikratis).

Usage

```
data(Income)
```

Format

A data frame with 325 observations on the following 5 variables.

`X` a numeric vector of x coordinates

`Y` a numeric vector of y coordinates

`UnemrT01` a numeric vector of total unemployment rate in 2001 (Census)

`PrSect01` a numeric vector of the proportion of economically active working in the primary financial sector (mainly agriculture; fishery; and forestry in 2001 (Census))

`Foreig01` a numeric vector of proportion of people who do not have the Greek citizenship in 2001 (Census)

`Income01` a numeric vector of mean recorded household income (in Euros) earned in 2001 and declared in 2002 tax forms

Details

The X,Y coordinates refer to the geometric centroids of the new 325 Municipalities in Greece (Programme Kallikratis) in 2011.

Source

The original shapefile of the corresponding polygons is available from the Hellenic Statistical Authority (EL.STAT.) at http://www.statistics.gr/el/digital-cartographical-data. The population, employment, citizenship and employment sector data is available from the Hellenic Statistical Authority (EL.STAT.) at http://www.statistics.gr/en/home but were aggregated to the new municipalities by the author. The income data are available from the General Secretariat of Information Systems in Greece at http://www.gsis.gr/ at the postcode level of geography and were aggregated to the new municipalities by the author.

References

Kalogirou, S., and Hatzichristos, T. (2007). A spatial modelling framework for income estimation. Spatial Economic Analysis, 2(3), 297-316. http://www.tandfonline.com/doi/abs/10.1080/17421770701576921

Kalogirou, S. (2010). Spatial inequalities in income and post-graduate educational attainment in Greece. Journal of Maps, 6(1), 393-400.http://www.tandfonline.com/doi/abs/10.4113/jom.2010.1095

Kalogirou, S. (2013) Testing geographically weighted multicollinearity diagnostics, GISRUK 2013, Department of Geography and Planning, School of Environmental Sciences, University of Liverpool, Liverpool, UK, 3-5 April 2013. http://gisc.gr/?mdocs-file=1140&mdocs-url=false

Examples

```
data(Income)
boxplot(Income$Income01)
hist(Income$PrSect01)
```

---

predict.grf — *Predict Method for Geographical Random Forest*

Description

Prediction of test data using the geographical random forest.

Usage

```
## S3 method for class 'grf'
predict(object, new.data, x.var.name, y.var.name, local.w=1, global.w=0,...)
```

Arguments

`object`	an object that created by the function grf that includes all local forests.
`new.data`	a data frame containing new data.
`x.var.name`	the name of the variable with X coordinates.
`y.var.name`	the name of the variable with Y coordinates.
`local.w`	weight of the local model predictor allowing semi-local predictions. Default value is 1.
`global.w`	weight of the global model predictor allowing semi-local predictions. Default value is 0.
`...`	for other arguments of the generic predict functions. Not used currently.

Details

A Geographical Random Forest prediction on unknown data. The nearest local random forest model in coordinate space is used to predict in each unknown y-variable location.

Value

vector of predicted values

Note

This function is under development. There should be improvements in future versions of the package SpatialML. Any suggestion is welcome!

Author(s)

Stamatis Kalogirou <stamatis@lctools.science>, Stefanos Georganos <sgeorgan@ulb.ac.be>

References

Stefanos Georganos, Tais Grippa, Assane Niang Gadiaga, Catherine Linard, Moritz Lennert, Sabine Vanhuysse, Nicholus Odhiambo Mboga, Eléonore Wolff & Stamatis Kalogirou (2019) Geographical Random Forests: A Spatial Extension of the Random Forest Algorithm to Address Spatial Heterogeneity in Remote Sensing and Population Modelling, Geocarto International, DOI: 10.1080/10106049.2019.1595177

See Also

grf

Examples

```
#Load the sample data
data(Income)

#Create the vector of XY coordinates
Coords<-Income[,1:2]

#Fit local model
grf <- grf(Income01 ~ UnemrT01 + PrSect01, dframe=Income, bw=60,
          kernel="adaptive", coords=Coords)

#Create New Random Data - XY coordinates inside the sample data map extend
x<-runif(20, min = 142498, max = 1001578)
y<-runif(20, min = 3855768, max = 4606754)
u<-runif(20, min = 5, max = 50)
p<-runif(20, min = 0, max = 100)
f<-runif(20, min = 2, max = 30)
df2<-data.frame(X=x, Y= y,  UnemrT01=u,  PrSect01=p,  Foreig01=f)

#Make predictions using the local model
predict.grf(grf, df2, x.var.name="X", y.var.name="Y", local.w=1, global.w=0)
```

`random.test.data` *Radmom data generator*

Description

Generates datasets with random data for modelling including a dependent variable, independent variables and X,Y coordinates.

Usage

```
random.test.data(nrows = 10, ncols = 10, vars.no = 3, dep.var.dis = "normal",
                 xycoords = TRUE)
```

Arguments

`nrows`	an integer referring to the number of rows for a regular grid
`ncols`	an integer referring to the number of columns for a regular grid
`vars.no`	an integer referring to the number of independent variables
`dep.var.dis`	a character referring to the distribution of the dependent variable. Options are "normal" (default) and "poisson"
`xycoords`	a logical value indicating whether X,Y coordinates will be created (default) or not.

Details

The creation of a random dataset was necessary here to provide examples to some functions. However, random datasets may be used in simulation studies.

Value

a dataframe

Author(s)

Stamatis Kalogirou <stamatis@lctools.science>

Examples

```
RDF <- random.test.data(12,12,3)
```

Index

∗Topic **Greek Municipalities**
Income, 4
∗Topic **Income**
Income, 4
∗Topic **datasets**
Income, 4
∗Topic **local random forest**
predict.grf, 5
∗Topic **predictive analytics**
grf, 2
∗Topic **random data**
random.test.data, 7
∗Topic **spatial random forest**
grf, 2

formula, *2*

grf, 2, *6*

Income, 4

predict.grf, *3*, 5

random.test.data, 7
randomForest, *2*