Package 'MRSea'

January 14, 2014

Title Marine Renewables Strategic environmental assessment

Description Examines animal survey data for signs of changes in animal abundance and distribution following marine renewables development. The functions of this package can be used to analyse segmented line transect data and nearshore vantage point data. Non-parametric bootstrapping can be used to estimate uncertainty. Several model assessment tools are available. This review constitutes work carried out at the Centre for Research into Ecological and Environmental Modelling (CREEM) at the University of St. Andrews, performed under contract for Marine Scotland (SB9 (CR/2012/05)).

Version 0.1.2

Date 2014-01-14

Author Lindesay Scott-Hayward <lass@st-and.ac.uk>, Cornelia Oedekoven, Monique Mackenzie, Cameron Walker <cameron.walker@auckland.ac.nz>

Maintainer Lindesay Scott-Hayward <lass@st-and.ac.uk>

Depends R (>= 3.0.0), calibrate (>= 1.7.2), car (>= 2.0-19), fields (>= 6.8), geepack (>= 1.1-6), ggplot2 (>= 0.9.3.1), lawstat (>= 2.4), Matrix (>= 1.0-12), mrds (>= 2.1.4), mvtnorm (>= 0.9-9996), splines (>= 3.0.1)

License GPL-2

LazyData true

Note Scott-Hayward LAS, Oedekoven CS, Mackenzie ML, Walker, CG and Rexstad E (2013). ``MRSea package (version 0.1.2): Statistical Modelling of bird and cetacean distributions in offshore renewables development areas". University of St. Andrews: Contract with Marine Scotland: SB9 (CR/2012/05)

URL http://creem2.st-and.ac.uk/software.aspx

Collate 'LocalRadialFunction.R' 'MRSea-package.r' 'SALSA1DCode.R' 'SALSA2DCode.R' 'bootstrap.orig.data.R' 'create.NHAT.R' 'create.bootcount.data.R' 'create.bootstrap.data.R' 'create.count.data.R' 'dis.data.de.r' 'dis.data.no.r' 'dis.data.re.r' 'do.bootstrap.cress.R' 'do.bootstrap.gam.r' 'functions.R' 'getCV_cress.R' 'getCVfoldID.R' 'getDifferences.R' 'getPlotDimensions.R' 'getPvalues.R' 'knotgrid.ns.r' 'knotgrid.off.r' 'makeBootCIs.R' 'makesplineParams.R' 'ns.data.de.r' 'ns.data.no.r'

'ns.data.re.r' 'ns.predict.data.de.r' 'ns.predict.data.no.r'
'ns.predict.data.re.r' 'plotCumRes.R' 'plotRunsProfile.R'
'predict.data.de.r' 'predict.data.no.r' 'predict.data.re.r'
'runACF.R' 'runDiagnostics.R' 'runInfluence.R'
'runInfluenceCheck.R' 'runPartialPlots.R' 'runSALSA.R'
'runSALSA1D_withremoval.R' 'runSALSA2D.R' 'which.bin.R'

R topics documented:

`acffunc` *calculate correlation for residuals by block*

Description

calculate correlation for residuals by block

Usage

```
acffunc(block, model)
```

Arguments

`block`	Vector of blocks that identify data points that are correlated
`model`	Fitted model object (glm or gam)

`bootstrap.orig.data` *Obtaining a data frame of bootstrapped data using resamples*

Description

This function extracts the records corresponding to each resample from the original distance data and pastes them together in a new data frame which is returned.

Usage

```
bootstrap.orig.data(orig.data, resample, new.resamples, resamples.no)
```

Arguments

`orig.data`	Original data to be bootstrapped
`resample`	Specifies the resampling unit for bootstrapping, default is `transect.id`. Must match a column name in `orig.data` exactly
`new.resamples`	String of resampled units from `data[,"resample"]`. Created by `create.bootstrap.data()`
`resamples.no`	Length of new.resamples

Value

Returns bootstrapped data. Internal function called by function `create.bootstrap.data`.

Examples

```
data(dis.data.re)
resample<-"transect.id"
samples<-unique(dis.data.re[,resample])
resamples.no<-length(samples)
new.resamples<-sample(samples,resamples.no,replace=TRUE)
bootstrap.data<-bootstrap.orig.data(dis.data.re,resample,new.resamples,resamples.no)
```

---

checkfactorlevelcounts

Factor level response check

Description

This function checks that there are some non-zero counts in each level of each factor variable for consideration in a model

Usage

```
checkfactorlevelcounts(factorlist, data, response)
```

Arguments

factorlist	Vector of factor variables specified in model. Specified so that a check can be made that there are non-zero counts in all levels of each factor.
data	Data frame containing columns of covariates listed in factorlist. Column names must match with names in factorlist
response	A vector of response values

Examples

```
# load data
data(ns.data.re)

checkfactorlevelcounts(factorlist=c(floodebb, impact), ns.data.re,
     ns.data.re$birds)
```

---

create.bootcount.data *Aggregate bootstrapped distance data into count data*

Description

This function creates a new data set where dis.data is aggregated for each visit to a segment. For bootstrapped data, the column with the ids for visits to a segment is segment.id2 which is created by create.bootstrap.data using the default for argument rename. The sum of the estimated number of individuals for each segment from dis.data$NHAT is given in the column NHAT in the new data. All other columns from the observation layer should be discarded. This is achieved by specifying the columns that should be retained using the argument column.numbers. Generally, all columns from the segment and higher levels should be kept. If the default is used, column.numbers=NULL, the columns distance, object, size, distbegin and distend from the observation level are automatically discarded. Note that for those columns from the observation layer that are kept, only the first recorded value will be transferred.

Usage

```
create.bootcount.data(dis.data, column.numbers = NULL)
```

Arguments

dis.data	Data frame containing distance data (one row for each detection). Expects a column NHAT, i.e. size of detection divided by its probability of detection (see create.NHAT) and that and that ids in segment.id2 are unique regardless of what resampled transect they belong to.
column.numbers	Optional argument: vector of integers indicating which columns other than NHAT from dis.data should be retained in the returned data.

Value

This function returns bootstrapped count data that is suited for second stage count modelling of distance sampling data. The data includes the columns NHAT and area which are the response and the offset required by functions concerned with second stage modelling from this package.

Examples

```
data(dis.data.re)
# bootstrap data without stratification
dis.data.re$survey.id<-paste(dis.data.re$season,dis.data.re$impact,sep="")
result<-ddf(dsmodel=~mcds(key="hn", formula=~1), data=dis.data.re, method="ds",
            meta.data=list(width=250))
dis.data.re<-create.NHAT(dis.data.re,result)

bootstrap.data<-create.bootstrap.data(dis.data.re)

bootcount.data<-create.bootcount.data(bootstrap.data)
```

---

create.bootstrap.data *Create bootstrap data for non-parametric bootstrapping*

Description

This function creates one realisation of bootstrapped data based on dis.data. The default resampling unit is transect.id which may be modified using the argument resample.

Usage

```
create.bootstrap.data(dis.data, resample = "transect.id",
  rename = "segment.id", stratum = NULL)
```

Arguments

dis.data	Original data to be bootstrapped. Requires a column that matches argument resample exactly.
resample	Specifies the resampling unit for bootstrapping, default is transect.id. Must match a column name in dis.data exactly

rename	A vector of column names for which a new column needs to be created for the bootstrapped data. This defaults to segment.id for line transects, however others might be added A new column with new ids will automatically be created for the column listed in resample
stratum	The column name in dis.data that identifies the different strata. The default NULL returns un-stratified bootstrap data. If stratum is specified, this requires a column in dis.data that matches argument stratum exactly

Value

Returns one realisation of bootstrapped distance data. Note that a new column (in addition to those listed under argument rename) is created. If the default for resample is used, a column with new unique ids called transect.id2. Note that a new column is created with renamed bootstrap resamples to preserve the number of unique bootstrap resamples. If the default for resample is used, i.e. transect.id, this new column is called transect.id2. In addition, a new column segment.id2 is created which is required for other bootstrap functions.

Examples

```
data(dis.data.re)
# run distance analysis to create NHATS
dis.data.re$survey.id<-paste(dis.data.re$season,dis.data.re$impact,sep="")
result<-ddf(dsmodel=~mcds(key="hn", formula=~1), data=dis.data.re, method="ds",
            meta.data=list(width=250))
dis.data.re<-create.NHAT(dis.data.re,result)

# bootstrap data without stratification
bootstrap.data<-create.bootstrap.data(dis.data.re)
# boostrap data with stratification (here by survey which is composed of
# season and impact)
dis.data.re$survey.id<-paste(dis.data.re$season,dis.data.re$impact,sep="")
bootstrap.data.str<-create.bootstrap.data(dis.data.re, stratum = "survey.id")
```

---

create.count.data *Aggregate distance data into count data*

Description

This function creates a new data set where dis.data is aggregated for each visit to a segment (segment.id). The sum of the estimated number of individuals for each segment from dis.data$NHAT is given in the column NHAT in the new data. Only columns from the segment or higher layers should be carried over into count.data from dis.data. Use argument column.numbers to identify these.

Usage

```
create.count.data(dis.data, column.numbers = NULL)
```

Arguments

dis.data	Data frame containing distance data (one row for each detection). Expects a column NHAT, i.e. size of detection divided by its probability of detection (see create.NHAT) and that ids in segment.id are unique regardless of what transect they belong to

`column.numbers` Optional argument: vector of integers indicating which columns other than `NHAT` from `dis.data` should be retained in the returned data. Generally all columns from the segment and higher levels should be kept while those from the observation level should be discarded. If the default is used, `column.numbers=NULL`, the columns `distance`, `object`, `size`, `distbegin` and `distend` from the observation level are automatically discarded. Note that for those columns from the observation layer that are kept, only the first recorded value will be transferred.

Value

This function returns count data that is suited for second stage count modelling of distance sampling data. The data includes the columns `NHAT` and `area` which are the response and the offset required by functions concerned with second stage modelling from this package.

Examples

```
data(dis.data.re)
dis.data.re$survey.id<-paste(dis.data.re$season,dis.data.re$impact,sep="")
result<-ddf(dsmodel=~mcds(key="hn", formula=~1), data=dis.data.re, method="ds",
           meta.data=list(width=250))
dis.data.re<-create.NHAT(dis.data.re,result)
count.data<-create.count.data(dis.data.re)
```

---

create.NHAT *Estimated number of individuals for each detection*

Description

This function creates a new column in `data` which contains the estimated number of animals for each detection. This is the number of observed individuals divided by their probability of detection using MCDS methods (size/detection probability). In the case that no `size` column is given in `dis.data`, it is assumed that detections were made of individuals and `size` is set to 1 for all detections. The values for `size` and `NHAT` are set to zero in case the distance was larger than the truncation distance `w` specified in `det.fct.object`. In addition, a new column `area` is created which is used as the offset in the second stage count model (segment length * (truncation distance/1000) * 2). The truncation distance is divided by 1000 to convert it from metres to km. It is assumed that the segment data represents two-sided surveys. In case the survey was one-sided, this column needs to be divided by 2 after the call to this function.

Usage

```
create.NHAT(data, ddf.obj)
```

Arguments

`data`	distance data object used with `det.fct` to estimate probabilities of detection
`ddf.obj`	detection function object created by `ddf`

Examples

```
data(dis.data.re)
result<-ddf(dsmodel=~mcds(key="hn", formula=~1), data=dis.data.re,method="ds",
     meta.data= list(width=250,binned=FALSE))
dis.data<-create.NHAT(dis.data.re,result)
```

`dis.data.de` *Line transect data with decrease post-impact*

Description

A simulated dataset containing the observed perpendicular distances, the effort data and other variables of segmented line transect data. The variables are as follows:

Format

A data frame with 10759 rows and 12 variables

Details

- `transect.id` Identifier for the individual visits to the transects
- `transect.label` Labels for transects
- `season` Numerical indicator for the four different seasons
- `impact` Numerical indicator for before (0) and after (1) impact
- `segment.id` Identifier for individual visits to the segment
- `segment.label` Label for segments
- `length` Length of segment in km
- `x.pos` spatial location in the horizontal axis in UTMs
- `y.pos` spatial location in the vertical axis in UTMs
- `depth` Depth in m
- `object` Id for detected object
- `distance` Perpendicular distance from the line

`dis.data.no` *Line transect data with no post-impact consequence*

Description

A simulated dataset containing the observed perpendicular distances, the effort data and other variables of segmented line transect data. The variables are as follows:

Format

A data frame with 10771 rows and 12 variables

Details

- `transect.id` Identifier for the individual visits to the transects
- `transect.label` Labels for transects
- `season` Numerical indicator for the four different seasons
- `impact` Numerical indicator for before (0) and after (1) impact
- `segment.id` Identifier for individual visits to the segment
- `segment.label` Label for segments
- `length` Length of segment in km
- `x.pos` spatial location in the horizontal axis in UTMs
- `y.pos` spatial location in the vertical axis in UTMs
- `depth` Depth in m
- `object` Id for detected object
- `distance` Perpendicular distance from the line

`dis.data.re` *Line transect data with redistribution post-impact*

Description

A simulated dataset containing the observed perpendicular distances, the effort data and other variables of segmented line transect data. The variables are as follows:

Format

A data frame with 10951 rows and 12 variables

Details

- `transect.id` Identifier for the individual visits to the transects
- `transect.label` Labels for transects
- `season` Numerical indicator for the four different seasons
- `impact` Numerical indicator for before (0) and after (1) impact
- `segment.id` Identifier for individual visits to the segment
- `segment.label` Label for segments
- `length` Length of segment in km
- `x.pos` spatial location in the horizontal axis in UTMs
- `y.pos` spatial location in the vertical axis in UTMs
- `depth` Depth in m
- `object` Id for detected object
- `distance` Perpendicular distance from the line

`do.bootstrap.cress` *Bootstrapping function without model selection using CReSS/SALSA for fitting the second stage count model*

Description

This fuction performs a specified number of bootstrapping iterations using CReSS/SALSA for fitting the second stage count model. See below for details.

Usage

```
do.bootstrap.cress(orig.data, predict.data, ddf.obj = NULL, model.obj,
  splineParams, g2k, resample = "transect.id", rename = "segment.id",
  stratum = NULL, B, name = NULL, save.data = FALSE, nhats = FALSE,
  seed = 12345, nCores = 1)
```

Arguments

`orig.data`	The original data. In case `ddf.obj` is specified, this should be the original distance data. In case `ddf.obj` is `NULL`, it should have the format equivalent to `count.data` where each record represents the summed up counts at the segments.
`predict.data`	The prediction grid data
`ddf.obj`	The ddf object created for the best fitting detection model. Defaults to `NULL` for nearshore data.
`model.obj`	The best fitting `CReSS` model for the original count data
`splineParams`	The object describing the parameters for fitting the one and two dimensional splines
`g2k`	(N x k) matrix of distances between all prediction points (N) and all knot points (k)
`resample`	Specifies the resampling unit for bootstrapping, default is `transect.id`. Must match a column name in `dis.data` exactly
`rename`	A vector of column names for which a new column needs to be created for the bootstrapped data. This defaults to `segment.id` for line transects (which is required for `create.bootcount.data`), others might be added. A new column with new ids will automatically be created for the column listed in `resample`. In case of nearshore data, this argument is ignored.
`stratum`	The column name in `orig.data` that identifies the different strata. The default `NULL` returns un-stratified bootstrap data. In case of nearshore data, this argument is ignored.
`B`	Number of bootstrap iterations
`name`	Analysis name. Required to avoid overwriting previous bootstrap results. This name is added at the beginning of "predictionboot.RData" when saving bootstrap predictions.
`save.data`	If TRUE, all created bootstrap data will be saved as an RData object in the working directory at each iteration, defaults to FALSE

nhats	(default = FALSE). If you have calculated bootstrap NHATS because there is no simple ddf object then a matrix of these may be fed into the function. The number of columns of data should >= B. The rows must be equal to those in orig.data and d2k and *must* be in matching order.
seed	Set the seed for the bootstrap sampling process.
nCores	Set the number of computer cores for the bootstrap process to use (default = 1). The more cores the faster the proces but be wary of over using the cores on your computer. If nCores > (number of computer cores - 2), the function defaults to nCores = (number of computer cores - 2). Note: On a Mac computer the parallel code does not compute so use nCores=1.

Details

In case of distance sampling data, the following steps are performed for each iteration:

- the original data is bootstrapped

- a detection function is fitted to the bootstrapped data

- a count model is fitted to the bootstrapped data

- coefficients are resampled from a multivariate normal distribution defined by MLE and COV from count model

- predictions are made to the prediction data using the resampled coefficients

In case of count data, the following steps are performed for each iteration:

- coefficients are resampled from a multivariate normal distribution defined by MLE and COV from the best fitting count model

- predictions are made to the prediction data using the resampled coefficients

Value

The function returns a matrix of bootstrap predictions. The number of rows is equal to the number of rows in predict.data. The number of columns is equal to B. The matrix may be very large and so is stored directly into the working directory as a workspace object: '"name"predictionboot.RObj'. The object inside is called bootPreds.

Examples

```
# ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
# offshore redistribution data
# ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
data(dis.data.re)
data(predict.data.re)
data(knotgrid.off)
# ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
# distance sampling
dis.data.re$survey.id<-paste(dis.data.re$season,dis.data.re$impact,sep="")
result<-ddf(dsmodel=~mcds(key="hn", formula=~1), data=dis.data.re, method="ds",
        meta.data=list(width=250))
dis.data.re<-create.NHAT(dis.data.re,result)
count.data<-create.count.data(dis.data.re)

# ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
# spatial modelling
splineParams<-makesplineParams(data=count.data, varlist=c(depth))
```

```
#set some input info for SALSA
count.data$response<- count.data$NHAT
# make distance matrices for datatoknots and knottoknots
distMats<-makeDists(cbind(count.data$x.pos, count.data$y.pos), na.omit(knotgrid.off))
# choose sequence of radii
r_seq<-getRadiiChoices(8,distMats$dataDist)
# set initial model without the spatial term
initialModel<- glm(response ~ as.factor(season) + as.factor(impact) + offset(log(area)),
                family=quasipoisson, data=count.data)
# make parameter set for running salsa2d
salsa2dlist<-list(fitnessMeasure = QICb, knotgrid = knotgrid.off, startKnots=4, minKnots=4,
                maxKnots=20, r_seq=r_seq, gap=4000, interactionTerm="as.factor(impact)")
salsa2dOutput_k6<-runSALSA2D(initialModel, salsa2dlist, d2k=distMats$dataDist,
                   k2k=distMats$knotDist, splineParams=splineParams)

splineParams<-salsa2dOutput_k6$splineParams
# specify parameters for local radial function:
radiusIndices <- splineParams[[1]]$radiusIndices
dists <- splineParams[[1]]$dist
radii <- splineParams[[1]]$radii
aR <- splineParams[[1]]$invInd[splineParams[[1]]$knotPos]
count.data$blockid<-paste(count.data$transect.id, count.data$season, count.data$impact, sep=)
# Re-fit the chosen model as a GEE (based on SALSA knot placement) and GEE p-values
geeModel<- geeglm(formula(salsa2dOutput_k6$bestModel), data=count.data, family=poisson, id=blockid)
dists<-makeDists(cbind(predict.data.re$x.pos, predict.data.re$y.pos), na.omit(knotgrid.off),
       knotmat=FALSE)$dataDist

# ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
# bootstrapping
do.bootstrap.cress(dis.data.re, predict.data.re, ddf.obj=result, geeModel, splineParams,
              g2k=dists, resample=transect.id, rename=segment.id, stratum=survey.id,
              B=4, name="cress", save.data=FALSE, nhats=NULL, nCores=1)
load("cresspredictionboot.RData") # loading the bootstrap predictions into the workspace
# look at the first 6 lines of the bootstrap predictions (on the scale of the response)
head(bootPreds)

## Not run:
# In parallel (Note: windows machines only)
require(parallel)
do.bootstrap.cress(dis.data.re, predict.data.re, ddf.obj=result, geeModel, splineParams,
                g2k=dists, resample=transect.id, rename=segment.id, stratum=survey.id,
                B=4, name="cress", save.data=FALSE, nhats=NULL, nCores=4)
load("cresspredictionboot.RData") # loading the bootstrap predictions into the workspace
# look at the first 6 lines of the bootstrap predictions (on the scale of the response)
head(bootPreds)

## End(Not run)

# ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
# nearshore redistribution data
# ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
## Not run:
do.bootstrap.cress(ns.data.re, ns.predict.data.re, ddf.obj=NULL, geeModel, splineParams,
             g2k=dists, resample=transect.id, rename=segment.id, stratum=NULL,
             B=2, name="cress", save.data=FALSE, nhats=NULL)
load("cresspredictionboot.RData") # loading the predictions into the workspace
# look at the first 6 lines of the bootstrap predictions (on the scale of the response)
```

```
head(bootPreds)
## End(Not run)
```

`do.bootstrap.gam` *Bootstrapping function without model selection using* `gam` *as the second stage count model*

Description

This fuction performs a specified number of bootstrapping iterations using gams for fitting the second stage count model. See below for details.

Usage

```
do.bootstrap.gam(orig.data, predict.data, ddf.obj = NULL, model.obj,
  resample = "transect.id", rename = "segment.id", stratum = NULL, B,
  name = NULL, save.data = FALSE, nhats = NULL)
```

Arguments

`orig.data`	The original data. In case `ddf.obj` is specified, this should be the original distance data. In case `ddf.obj` is `NULL`, it should have the format equivalent to `count.data` where each record represents the summed up counts at the segments.
`predict.data`	The prediction grid data
`ddf.obj`	The ddf object created for the best fitting detection model. Defaults to `NULL` for nearshore data.
`model.obj`	The best fitting `gam` model for the original count data
`resample`	Specifies the resampling unit for bootstrapping, default is `transect.id`. Must match a column name in `dis.data` exactly
`rename`	A vector of column names for which a new column needs to be created for the bootstrapped data. This defaults to `segment.id` for line transects (which is required for `create.bootcount.data`), others might be added. A new column with new ids will automatically be created for the column listed in `resample`. In case of nearshore data, this argument is ignored.
`stratum`	The column name in `orig.data` that identifies the different strata. The default `NULL` returns un-stratified bootstrap data. In case of nearshore data, this argument is ignored.
`B`	Number of bootstrap iterations
`name`	Analysis name. Required to avoid overwriting previous bootstrap results. This name is added at the beginning of "predictionboot.RData" when saving bootstrap predictions.
`save.data`	If TRUE, all created bootstrap data will be saved as an RData object in the working directory at each iteration, defaults to FALSE
`nhats`	(default = FALSE). If you have calculated bootstrap NHATS because there is no simple ddf object then a matrix of these may be fed into the function. The number of columns of data should >= B. The rows must be equal to those in `orig.data` and `d2k` and *must* be in matching order.

Details

In case of distance sampling data, the following steps are performed for each iteration:

- the original data is bootstrapped

- a detection function is fitted to the bootstrapped data

- a count model is fitted to the bootstrapped data

- coefficients are resampled from a multivariate normal distribution defined by MLE and COV from count model

- predictions are made to the prediction data using the resampled coefficients

In case of count data, the following steps are performed for each iteration

- coefficients are resampled from a multivariate normal distribution defined by MLE and COV from the best fitting count model

- predictions are made to the prediction data using the resampled coefficients

Value

The function returns a matrix of bootstrap predictions. The number of rows is equal to the number of rows in predict.data. The number of columns is equal to B. The matrix may be very large and so is stored directly into the working directory as a workspace object: '"name"predictionboot.RObj'. The object inside is called bootPreds.

Examples

```
# offshore redistribution data
data(dis.data.re)
data(predict.data.re)
dis.data.re$survey.id<-paste(dis.data.re$season,dis.data.re$impact,sep="")
result<-ddf(dsmodel=~mcds(key="hn", formula=~1), data=dis.data.re, method="ds",
            meta.data=list(width=250))
dis.data.re<-create.NHAT(dis.data.re,result)
count.data<-create.count.data(dis.data.re)
require(mgcv)
gam.2<-gam(NHAT~as.factor(impact)+s(x.pos,y.pos,by=as.factor(impact))+offset(log(area)),
           data=count.data,family=quasipoisson)
do.bootstrap.gam(dis.data.re,predict.data.re,ddf.obj=result,model.obj=gam.2,resample="transect.id",
            rename="segment.id",stratum=survey.id,1,name=gam,save.data=FALSE,nhats=NULL)
load("gampredictionboot.RData") # loading the predictions into the workspace
# look at the first 6 lines of the predictions on the response scale
head(bootPreds)


## Not run: # nearshore redistribution data
data(ns.data.re)
data(ns.predict.data.re)
require(mgcv)
gam.ns2=gam(birds~as.factor(impact)+s(x.pos,y.pos,by=as.factor(impact))+offset(log(area)),
         data=ns.data.re,family=quasipoisson)
do.bootstrap.gam(ns.data.re,ns.predict.data.re,ddf.obj=NULL,model.obj=gam.ns2,resample=NULL,
                rename=NULL,stratum=NULL,1,name=ns.gam,save.data=FALSE,nhats=NULL)
# load the replicate predictions into the workspace
load("ns.gampredictionboot.RData")
# look at the first 6 lines of the predictions on the response scale
head(bootPreds)
## End(Not run)
```

getCVids *IDs for running cross validation*

Description

This function creates a string of integers which will be used for pointing to the right subsets of data for cross validation of regression objects

Usage

```
getCVids(data, folds, block = NULL)
```

Arguments

data	data used in regression model
folds	integer number of validation data sets
block	column in data indicating the blocking structure for cross-validation (if block = NULL, individual observations will be used as blocks)

Details

The function returns a random sequence of 1:folds of the same length as observations in data. It is called by other functions, e.g. getCV_CReSS.

Examples

```
# load data
data(ns.data.re)

CVids<-getCVids(ns.data.re, 5)
```

getCV_CReSS *Calculate cross-validation score for a CReSS type model*

Description

Calculate cross-validation score for a CReSS type model

Usage

```
getCV_CReSS(data, baseModel, splineParams)
```

Arguments

data	Data frame containing columns of covariates contained in baseModel.
baseModel	glm or CReSS type model object
splineParams	list object containing information for fitting one and two dimensional splines. See makesplineParams for more details.

Details

There must be a column in the data called `foldid`, which can be created using `getCVids`. This column defines the folds of data for the CV calculation.

Examples

```
# load data
data(ns.data.re)
# load prediction data
data(ns.predict.data.re)

splineParams<-makesplineParams(data=ns.data.re, varlist=c(observationhour, DayOfMonth))
# set some input info for SALSA
ns.data.re$response<- ns.data.re$birds
salsa1dlist<-list(fitnessMeasure = QICb, minKnots_1d=c(2,2), maxKnots_1d = c(20, 20),
               startKnots_1d = c(2,2), degree=c(2,2), maxIterations = 10, gaps=c(1,1))

# set initial model without the spline terms in there
# (so all other non-spline terms)
initialModel<- glm(response ~ as.factor(floodebb) + as.factor(impact) + offset(log(area)),
                  family=quasipoisson,data=ns.data.re)

# run SALSA
salsa1dOutput<-runSALSA1D(initialModel, salsa1dlist, varlist=c(observationhour,DayOfMonth),
          factorlist=c(floodebb, impact), ns.predict.data.re, splineParams=splineParams)

# make blocking structure and fold structure
ns.data.re$blockid<-paste(ns.data.re$GridCode, ns.data.re$Year, ns.data.re$MonthOfYear,
                    ns.data.re$DayOfMonth, sep=)
ns.data.re$blockid<-as.factor(ns.data.re$blockid)
ns.data.re$foldid<-getCVids(ns.data.re, folds=5, block=blockid)

# calculate CV
cv1<-getCV_CReSS(ns.data.re, salsa1dOutput$bestModel, salsa1dOutput$splineParams)
```

`getDifferences` *Identify any significant differences between predicted data before an impact event and predicted data after an impact event*

Description

Identify any significant differences between predicted data before an impact event and predicted data after an impact event

Usage

```
getDifferences(beforePreds, afterPreds, quants = c(0.025, 0.975))
```

Arguments

`beforePreds`	Matrix of bootstrap predictions (n x B) to each grid cell before impact (same length and order as `afterPreds`)

afterPreds	Matrix of bootstrap predictions (n x B) to each grid cell after impact (same length and order as beforePreds)
quants	(default = =c(.025,.975)) Quantile for significance.

Details

This function finds the differences for every predicted grid cell for every bootstrap replicate. Quantiles are used to determine whether each difference is significantly different from zero and if so, in what direction.

Value

A list is returned consisting of

mediandiff	Vector of the median difference for each grid cell
lowerci	Vector of the lower 2.5% difference for each grid cell
upperci	Vector of the upper 97.5% difference for each grid cell
significanceMarker	Vector of significance. 0: not significant, 1: significant and positive, -1: significant and negative

Examples

```
## Not run:
getDifferences(beforePreds, afterPreds)
## End(Not run)
```

---

getPlotdimensions *find the plotting dimensions for quilt.plot when using a regular grid*

Description

find the plotting dimensions for quilt.plot when using a regular grid

Usage

```
getPlotdimensions(x.pos, y.pos, segmentWidth, segmentLength)
```

Arguments

x.pos	Vector of x-coordinates in dataset
y.pos	Vector of y-coordinates in dataset
segmentWidth	Width of each grid cell of data (in same units as x.pos)
segmentLength	Length of each grid cell of data (in same units as y.pos)

Examples

```
# # load data
data(ns.data.re)

getPlotdimensions(ns.data.re$x.pos, ns.data.re$y.pos, segmentWidth=500, segmentLength=500)
```

getPvalues *Calculate marginal p-values from a* model.

Description

An ANOVA is fitted repeatedly with each covariate being the last so that the output is marginal. varlist and factorlist are optional and shorten the variable names in the output.

Usage

```
getPvalues(model, varlist = NULL, factorlist = NULL)
```

Arguments

model	Fitted model object (gee)
varlist	(default =NULL). Vector of covariate names (continous covariates only) used to make the output table names shorter. Useful if spline parameters are specified in the model.
factorlist	(default =NULL). Vector of covariate names (factor covariates only) used to make the output table names shorter. Useful if spline parameters are specified in the model.

Value

Print out table of each variable and its associated marginal p-value.

Examples

```
# load data
data(ns.data.re)

# make blocking structure
ns.data.re$blockid<-paste(ns.data.re$GridCode, ns.data.re$Year, ns.data.re$MonthOfYear,
                    ns.data.re$DayOfMonth, sep=)
ns.data.re$blockid<-as.factor(ns.data.re$blockid)

initialModel<- geeglm(birds ~ as.factor(floodebb) + as.factor(impact) + observationhour + x.pos +
              y.pos + offset(log(area)), family=poisson,data=ns.data.re, id=blockid)

getPvalues(initialModel, varlist=c(observationhour, x.pos, y.pos),
            factorlist=c(floodebb, impact))

getPvalues(initialModel)
```

`getRadiiChoices` *Function for obtaining a sequence of range parameters for the CReSS smoother*

Description

Function for obtaining a sequence of range parameters for the CReSS smoother

Usage

```
getRadiiChoices(numberofradii = 8, distMatrix)
```

Arguments

`numberofradii`	The number of range parameters for SALSA to use when fitting the CReSS smooth. The default is 8. Remember, the more parameters the longer SALSA will take to find a suitable one for each knot location.
`distMatrix`	Matrix of distances between data locations and knot locations (n x k). May be Euclidean or geodesic distances. Euclidean distances created using `makeDists`.

Details

The range parameter determines the range of the influence of each knot. Small numbers indicate local influence and large ones, global influence.

Value

This function returns a vector containing a sequence of range parameters.

References

Scott-Hayward, L.; M. Mackenzie, C.Donovan, C.Walker and E.Ashe. Complex Region Spatial Smoother (CReSS). Journal of computational and Graphical Statistics. 2013. DOI: 10.1080/10618600.2012.762920

Examples

```
# load data
data(ns.data.re)
# load knot grid data
data(knotgrid.ns)

# make distance matrices for datatoknots and knottoknots
distMats<-makeDists(cbind(ns.data.re$x.pos, ns.data.re$y.pos), na.omit(knotgrid.ns))

# choose sequence of radii
r_seq<-getRadiiChoices(8, distMats$dataDist)
```

knotgrid.ns *Knot grid data for nearshore example*

Description

Knot grid data for nearshore example

knotgrid.off *Knot grid data for offshore example*

Description

Knot grid data for offshore example

LocalRadialFunction *Function for creating an exponential basis function for a spatial smooth using the CReSS method.*

Description

This function calculates a local radial exponential basis matrix for use in runSALSA2D.

Usage

```
LocalRadialFunction(radiusIndices, dists, radii, aR)
```

Arguments

radiusIndices	Vector of length startKnots identifying which radii (splineParams[[1]]$radii) will be used to initialise the model
dists	Matrix of distances between data locations and knot locations (n x k). May be Euclidean or geodesic distances.
radii	Sequence of range parameters for the CReSS basis from local (small) to global (large). Determines the range of the influence of each knot.
aR	Index of knot locations. The index contains numbers selected by SALSA from 1 to the number of legal knot locations na.omit(knotgrid). Used to specify which columns of dists should be used to construct the basis matrix.

Details

Calculate a local radial basis matrix for use in runSALSA2D. The distance matrix input may be Euclidean or geodesic distances.

Value

Returns a basis matrix with one column for each knot in aR and one row for every observation (i.e. same number of rows as dists)

Examples

```
# load data
data(ns.data.re)
# load knot grid data
data(knotgrid.ns)

splineParams<-makesplineParams(data=ns.data.re, varlist=c(observationhour))

#set some input info for SALSA
ns.data.re$response<- ns.data.re$birds

# make distance matrices for datatoknots and knottoknots
distMats<-makeDists(cbind(ns.data.re$x.pos, ns.data.re$y.pos), na.omit(knotgrid.ns), knotmat=FALSE)

# choose sequence of radii
r_seq<-getRadiiChoices(8, distMats$dataDist)

# using the fourth radius and picking 5 knots
basis<-LocalRadialFunction(radiusIndices=rep(4, 5), dists=distMats$dataDist, radii = r_seq,
        aR=c(3, 10, 15, 28, 31))
```

`makeBootCIs` *Calculate percentile confidence intervals from a matrix of bootstrapped predictions*

Description

Calculate percentile confidence intervals from a matrix of bootstrapped predictions

Usage

```
makeBootCIs(preds, quants = c(0.025, 0.975))
```

Arguments

`preds`	matrix of bootstrap predictions where each column is a bootstrap realisation
`quants`	(`default = c(0.025, 0.975)`. Vector of length two of quantiles.

Examples

```
## Not run:
makeBootCIs(bootPreds)

## End(Not run)
```

makeDists — *Make Euclidean distance matrices for use in CReSS and SALSA model frameworks*

Description

This function makes two Euclidean distance matrices. One for the distances between all spatial observations and all spatial knot locations. The other, if specified, is the distances between knot locations.

Usage

```
makeDists(datacoords, knotcoords, knotmat = TRUE)
```

Arguments

datacoords	Coordinates of the data locations
knotcoords	Coordinates of the legal knot locations
knotmat	(default=TRUE). Should a matrix of knot-knot distances be created

Details

The data-knot matrix is used in the CReSS basis and the knot-knot matrix is used in SALSA to determine where a nearest knot to 'move' should be.

Examples

```
# load data
data(ns.data.re)
# load knot grid data
data(knotgrid.ns)

# make distance matrices for datatoknots and knottoknots
distMats<-makeDists(cbind(ns.data.re$x.pos, ns.data.re$y.pos), na.omit(knotgrid.ns))
```

makesplineParams — *Constructing an object of spline parameters*

Description

This function makes a list object containing all of the information to fit splines to continuous data.

Usage

```
makesplineParams(data, varlist, predictionData = NULL, degree = NULL)
```

Arguments

data	Data frame containing columns of covariates listed in varlist. Column names must match with names in varlist
varlist	Vector of variable names for the covariates of interest
predictionData	Data frame containing columns of covariates listed in varlist. Column names must match with those in varlist. This parameter is used to find the maximum range of covariates between the data and prediction data. If predictionData is NULL then the range of the data is used.
degree	Vector specifying the degree of the spline. If unspecified, degree 2 is stored.

Details

The information is stored in list slots [[2]] and onward (slot [[1]] is reserved for a spatial term). Specifically:

covar. Name of covariate.

explanatory. Vector of covariate data.

knots. Knot(s) for spline fitting. This function initialises with a knot at the mean covariate value.

bd. This specifies the boundary knots. If predictionData is NULL then this is the range of the covariate data. Otherwise, the boundary knots are the maximum combined range of the data and prediction data.

degree. The degree of a B-spline. This function retuns 2 by default.

See runSALSA2D for details on the spatial slot ([[1]])

Examples

```
# load data
data(ns.data.re)
# load prediction data
data(ns.predict.data.re)

splineParams<- makesplineParams(ns.data.re, varlist=c(observationhour, DayOfMonth),
                predictionData=ns.predict.data.re)
```

MRSea *MRSea*

Description

MRSea

ns.data.de *Nearshore data with decrease post-impact*

Description

A simulated dataset containing the observed counts, the effort data and other variables of grid data. The variables are as follows:

Format

A data frame with 27798 rows and 12 variables

Details

- `x.pos` spatial location in the horizontal axis in UTMs
- `y.pos` spatial location in the vertical axis in UTMs
- `area` area surveyed in the gridcell in km squared
- `floodebb` 3 level factor covariate for tides
- `observationhour` hour of observation
- `GridCode` identifier for the different grids that were surveyed
- `Year` Year of the survey
- `DavOfMonth` Day of the survey
- `MonthOfYear` Month of the survey
- `impact` numerical indicator for before (0) and after (1) impact
- `birds` observed number of birds
- `cellid` identifier for the individual records

ns.data.no *Nearshore data with no effect of impact*

Description

A simulated dataset containing the observed counts, the effort data and other variables of grid data. The variables are as follows:

Format

A data frame with 27798 rows and 12 variables

Details

- `x.pos` spatial location in the horizontal axis in UTMs
- `y.pos` spatial location in the vertical axis in UTMs
- `area` area surveyed in the gridcell in km squared
- `floodebb` 3 level factor covariate for tides
- `observationhour` hour of observation
- `GridCode` identifier for the different grids that were surveyed
- `Year` Year of the survey
- `DavOfMonth` Day of the survey
- `MonthOfYear` Month of the survey
- `impact` numerical indicator for before (0) and after (1) impact
- `birds` observed number of birds
- `cellid` identifier for the individual records

`ns.data.re` *Nearshore data with redistribution post-impact*

Description

A simulated dataset containing the observed counts, the effort data and other variables of grid data. The variables are as follows:

Format

A data frame with 27798 rows and 12 variables

Details

- `x.pos` spatial location in the horizontal axis in UTMs
- `y.pos` spatial location in the vertical axis in UTMs
- `area` area surveyed in the gridcell in km squared
- `floodebb` 3 level factor covariate for tides
- `observationhour` hour of observation
- `GridCode` identifier for the different grids that were surveyed
- `Year` Year of the survey
- `DavOfMonth` Day of the survey
- `MonthOfYear` Month of the survey
- `impact` numerical indicator for before (0) and after (1) impact
- `birds` observed number of birds
- `cellid` identifier for the individual records

ns.predict.data.de *Prediction grid data for nearshore post-impact decrease*

Description

A simulated prediction dataset containing the true counts, the effort data and other variables of grid data. The variables are as follows:

Format

A data frame with 27798 rows and 11 variables

Details

- `x.pos` spatial location in the horizontal axis in UTMs
- `y.pos` spatial location in the vertical axis in UTMs
- `area` Area surveyed in the gridcell in km squared
- `floodebb` 3 level factor covariate for tide state
- `observationhour` hour of observation
- `GridCode` identifier for the different grids that were surveyed
- `Year` Year of the survey
- `DavOfMonth` Day of the survey
- `MonthOfYear` Month of the survey
- `impact` numerical indicator for before (0) and after (1) impact
- `birds` true density of birds

ns.predict.data.no *Prediction grid data for nearshore no post-impact consequence*

Description

A simulated prediction dataset containing the true counts, the effort data and other variables of grid data. The variables are as follows:

Format

A data frame with 27798 rows and 11 variables

Details

- `x.pos` spatial location in the horizontal axis in UTMs
- `y.pos` spatial location in the vertical axis in UTMs
- `area` Area surveyed in the gridcell in km squared
- `floodebb` 3 level factor covariate for tide state
- `observationhour` hour of observation
- `GridCode` identifier for the different grids that were surveyed
- `Year` Year of the survey
- `DavOfMonth` Day of the survey
- `MonthOfYear` Month of the survey
- `impact` numerical indicator for before (0) and after (1) impact
- `birds` true density of birds

`ns.predict.data.re` *Prediction grid data for nearshore post-impact redistribution*

Description

A simulated prediction dataset containing the true counts, the effort data and other variables of grid data. The variables are as follows:

Format

A data frame with 27798 rows and 11 variables

Details

- `x.pos` spatial location in the horizontal axis in UTMs
- `y.pos` spatial location in the vertical axis in UTMs
- `area` Area surveyed in the gridcell in km squared
- `floodebb` 3 level factor covariate for tide state
- `observationhour` hour of observation
- `GridCode` identifier for the different grids that were surveyed
- `Year` Year of the survey
- `DavOfMonth` Day of the survey
- `MonthOfYear` Month of the survey
- `impact` numerical indicator for before (0) and after (1) impact
- `birds` true density of birds

plotacf *run functions to create acf matrix and plot the results*

Description

run functions to create acf matrix and plot the results

Usage

```
plotacf(acfmat)
```

Arguments

acfmat	Matrix of output from acffunc (blocks x max block length).

plotCumRes *Calculate cumulative residuals and plot.*

Description

The output is plots of cumulative residuals.

Usage

```
plotCumRes(model, varlist, d2k = NULL, splineParams = NULL, label = "",
  save = FALSE)
```

Arguments

model	Fitted model object (glm or gam)
varlist	Vector of covariate names (continous covariates only)
d2k	(default=NULL). Distance matrix of data to knot points. Used only if there is a LocalRadialFunction smooth in the model formula
splineParams	(default =NULL) List object containing output from runSALSA/runSALSA2D required for updating model. Used only if there is a LocalRadialFunction smooth in the model formula. See makesplineParams for details of this object.
label	Label printed at the end of the plot name to identify it if save=TRUE.
save	(default=FALSE). Logical stating whether plot should be saved into working directory.

Value

Cumulative residual plots are returned for residuals ordered by each covariate in varlist, predicted value and index of observations (temporally). The blue dots are the residuals The black line is the line of cumulative residual. On the covariate plots (those in varlist) the grey line indicates what we would expect from a well fitted covariate. i.e. one that is fitted with excessive knots.

Note: if the covariate is discrete in nature (like the example below), there will be a lot of overplotting of residuals.

Examples

```
# load data
data(ns.data.re)

model<-glm(birds ~ observationhour + as.factor(floodebb) + as.factor(impact),
           family=quasipoisson, data=ns.data.re)

plotCumRes(model, varlist=c(observationhour))
```

`plotRunsProfile` *Calculate runs test and plot profile plot. The output is a plot of runs profiles (with p-value to indicate level of correlation)*

Description

Calculate runs test and plot profile plot. The output is a plot of runs profiles (with p-value to indicate level of correlation)

Usage

```
plotRunsProfile(model, varlist, label = "", save = FALSE)
```

Arguments

`model`	Fitted model object (glm or gam)
`varlist`	Vector of covariate names (continous covariates only)
`label`	Label printed at the end of the plot name to identify it when `save=TRUE`.
`save`	(`default=FALSE`). Logical stating whether plot should be saved into working directory.

Value

Runs profile plots are returned for residuals ordered by each covariate in `varlist`, predicted value and index of observations (temporally).

The black line is the line of sequences of positive or negative residuals. The vertical lines are the change between a sequence of positive to negative residuals (or vice versa).

The p-values are from a `runs.test` and indicate whether there is correlation in the residuals ($p<0.05$) or independence ($p>0.05$). The test statistic determines the type of correlation (positive/negative) and the result printed at the bottom of the figure.

Note: if the covariate is discrete in nature (like the example below), there will be a lot of overplotting of runs. Some jittering occurs at each discrete value (for covariates with <= 25 unique values).

Examples

```
# load data
data(ns.data.re)

model<-glm(birds ~ observationhour + as.factor(floodebb) + as.factor(impact),
           family=quasipoisson, data=ns.data.re)

plotRunsProfile(model, varlist=c(observationhour))
```

predict.data.de *Prediction grid data for post-impact decrease*

Description

A simulated dataset containing the true number of birds, the effort data and other variables of prediction grid data. The variables are as follows:

Format

A data frame with 37928 rows and 8 variables

Details

- `area` area surveyed in the gridcell in km squared
- `x.pos` spatial location in the horizontal axis in UTMs
- `y.pos` spatial location in the vertical axis in UTMs
- `depth` depth in m
- `segment.id` Identifier for individual visits to the segment
- `season` Numerical indicator for the four different seasons
- `impact` Numerical indicator for before (0) and after (1) impact
- `truth` number of birds

predict.data.no *Prediction grid data for no post-impact consequence*

Description

A simulated dataset containing the true number of birds, the effort data and other variables of prediction grid data. The variables are as follows:

Format

A data frame with 37928 rows and 8 variables

Details

- `area` area surveyed in the gridcell in km squared
- `x.pos` spatial location in the horizontal axis in UTMs
- `y.pos` spatial location in the vertical axis in UTMs
- `depth` depth in m
- `segment.id` Identifier for individual visits to the segment
- `season` Numerical indicator for the four different seasons
- `impact` Numerical indicator for before (0) and after (1) impact
- `truth` number of birds

predict.data.re *Prediction grid data for post-impact redistribution*

Description

A simulated dataset containing the true number of birds, the effort data and other variables of prediction grid data. The variables are as follows:

Format

A data frame with 37928 rows and 8 variables

Details

- `area` area surveyed in the gridcell in km squared
- `x.pos` spatial location in the horizontal axis in UTMs
- `y.pos` spatial location in the vertical axis in UTMs
- `depth` depth in m
- `segment.id` Identifier for individual visits to the segment
- `season` Numerical indicator for the four different seasons
- `impact` Numerical indicator for before (0) and after (1) impact
- `truth` number of birds

return.reg.spline.fit *Code for adaptively spacing knots for a given covariate.*

Description

Code for adaptively spacing knots for a given covariate.

Usage

```
return.reg.spline.fit(response, explanatory, degree, minKnots, maxKnots,
  startKnots, gap, winHalfWidth, fitnessMeasure = "BIC",
  maxIterations = 100, initialise = TRUE, initialKnots = NULL,
  baseModel = NULL, bd, spl)
```

Arguments

`response`	vector of response data for the modelling process
`explanatory`	vector of covariate to find knots for
`degree`	degree of the spline to be used
`minKnots`	minimum number of knots to fit
`maxKnots`	maximum number of knots to fit
`startKnots`	number of equally spaced knots to start with (between minKnots and maxKnots)
`gap`	minimum gap between knots (in unit of measurement of `explanatory`)

fitnessMeasure	(default=BIC). Measure used to evaluate the fit. Other options are AIC, AICc, BIC, QAIC, QAICc, QICb (Quasi-Likelihood Information Criterion with log(n) penalty)
maxIterations	exchange/improve heuristic will terminate after maxIterations if still running
initialise	(default = TRUE). Logical stating whether or not to start with equally spaced knots (TRUE) or user specified locations (FALSE)
initialKnots	If initialise=FALSE then the start locations for the knots are specified in initialKnots
baseModel	starting model for SALSA to use. Must not contain the covariate in explanatory
bd	the x-coordinate of the boundary knots of explanatory
spl	"bs" uses b-spline, "cc" uses cyclic cubic, "ns" uses natural cubic spline for fitting smooth to explanatory
winHalfWidth	Half-width of window used to calculate region with biggest average residual magnitude

Author(s)

Cameron Walker, Department of Enginering Science, University of Auckland.

---

return.reg.spline.fit.2d

Code for adaptively spacing knots for a spatial smooth. The smoothing process uses a CReSS basis.

Description

Code for adaptively spacing knots for a spatial smooth. The smoothing process uses a CReSS basis.

Usage

```
return.reg.spline.fit.2d(splineParams, startKnots, winHalfWidth,
  fitnessMeasure = "BIC", maxIterations = 10, tol = 0, baseModel = NULL,
  radiusIndices = NULL, initialise = TRUE, initialKnots = NULL,
  interactionTerm = NULL, knot.seed = 10)
```

Arguments

splineParams	List object where the first element [[1]] contains a list of objects for the 2D SALSA fitting process: knotDist, radii, dist, gridresp, grid, datacoords, response, knotgrid, minKnots, maxKnots, gap
startKnots	number of space-filled knots to start with (between minKnots and maxKnots)
fitnessMeasure	(default=BIC). Measure used to evaluate the fit. Other options are AIC, AICc, BIC, QICb (Quasi-Likelihood Information Criterion with log(n) penalty)
maxIterations	exchange/improve heuristic will terminate after maxIterations if still running
baseModel	starting model for SALSA to use. Must not already contain a spatial smooth.
radiusIndices	vector of length startKnots identifying which radii (splineParams[[1]]$radii) will be used to initialise the model

initialise	(default = TRUE). Logical stating whether or not to start with space-filled knots (TRUE) or user specified locations (FALSE)
initialKnots	If initialise=FALSE then the start locations for the knots are specified in initialKnots. Must be coordinates.
interactionTerm	(default=NULL). Specifies which term in baseModel the spatial smooth will interact with. If NULL no interaction term is fitted
winHalfWidth	Half-width of window used to calculate region with biggest average residual magnitude
tol	Tolerance for difference between fit measures. E.g. tol=2 means that the calculated fitness measures must be 2 units apart to be considered different

Details

The following are the details of the splineParams[[1]] objects. Note. If salsa1D has been run then details for those covariates will sit in splineParams[[2]] and onward.

knotDist: matrix of knot to knot distances (k x k). May be Euclidean or geodesic distances. Must be square and the same dimensions as nrows(na.omit(knotgrid))

radii Sequence of range parameters for the CReSS basis from local (small) to global (large). Determines the range of the influence of each knot.

dist: matrix of distances between data locations and knot locations (n x k). May be Euclidean or geodesic distances.

gridresp The first column of knotgrid

grid Index of knotgrid locations. Should be same length as knotgrid but with x=integer values from 1 to number of unique x-locations and y= integer values from 1 to number of unique y-locations.

datacoords: Coordinates of the data locations

response: vector of response data for the modelling process

knotgrid: grid of legal knot locations. Must be a regular grid with c(NA, NA) for rows with an illegal knot

minKnots: minimum number of knots to fit

maxKnots: maximum number of knots to fit

gap: Minimum gap between knots (in unit of measurement of datacoords)

Author(s)

Cameron Walker, Department of Enginering Science, University of Auckland.

---

runACF *run functions to create acf matrix and plot the results*

Description

run functions to create acf matrix and plot the results

Usage

```
runACF(block, model, store = FALSE, save = F)
```

Arguments

block	Vector of blocks that identify data points that are correlated
model	Fitted model object (glm or gam)
store	(default=F). Logical stating whether a list of the matrix of correlations is stored (output from acffunc.)
save	(default=FALSE). Logical stating whether plot should be saved into working directory.

Value

Plot of lag vs correlation. Each grey line is the correlation for each individual block in block. The red line is the mean values for each lag.

If store=TRUE then the matrix of correlations (nblocks x length_max_block) is returned and plotacf may be used to plot the acf.

Examples

```
# load data
data(ns.data.re)

model<-glm(birds ~ observationhour + as.factor(floodebb) + as.factor(impact),
           family=quasipoisson, data=ns.data.re)

ns.data.re$blockid<-paste(ns.data.re$GridCode, ns.data.re$Year, ns.data.re$MonthOfYear,
                    ns.data.re$DayOfMonth, sep=)
ns.data.re$blockid<-as.factor(ns.data.re$blockid)

runACF(ns.data.re$blockid, model)
```

---

runDiagnostics *functions to create observed vs fitted and fitted vs scaled pearsons residual plots*

Description

functions to create observed vs fitted and fitted vs scaled pearsons residual plots

Usage

```
runDiagnostics(model, plotting = "b", save = FALSE)
```

Arguments

model	Fitted model object (glm or gam)
plotting	Plotting options (default=b). b: returns both plots, f: returns observed vs fitted only and r: returns scale pearsons residual plot only.
save	(default=FALSE). Logical stating whether plot should be saved into working directory.

Value

Two plots:

`Observed vs Fitted`
: Plot of observed vs fitted with concordence correlation and marginal R-squared printed in the plot title.

`Fitted vs scaled Pearsons residuals`
: The red line is a locally weighted least squares regression line of all of the residuals.

Examples

```
# load data
data(ns.data.re)

model<-glm(birds ~ observationhour + as.factor(floodebb) + as.factor(impact),
           family=quasipoisson, data=ns.data.re)

runDiagnostics(model)
```

`runInfluence` *Assessing the influece of each correlated block on both the precision of the parameter estimates (COVRATIO statistics) and the sensitivity of model predictions (PRESS statistics).*

Description

Assessing the influece of each correlated block on both the precision of the parameter estimates (COVRATIO statistics) and the sensitivity of model predictions (PRESS statistics).

Usage

```
runInfluence(model, id, d2k = NULL, splineParams = NULL, save = FALSE)
```

Arguments

`model`	Fitted model object (glm or gam)
`id`	blocking structure
`d2k`	(`default=NULL`). (n x k) Matrix of distances between all data points in `model` and all valid knot locations.
`splineParams`	(`default=NULL`). List object containng output from runSALSA (e.g. knot locations for continuous covariates). See `makesplineParams` for more details of this object.
`save`	(`default=FALSE`). Logical stating whether plot should be saved into working directory.

Details

Always run `timeInfluenceCheck` first to see how long it will take to produce the plots.

Value

Two plots one each for COVRATIO and PRESS statistics, giving the influence of each block on precision of the parameter estimates and the sensitivity of model predictions. List object:

`influenceData` List of `blocks`, COVRATIO statistics and PRESS statistics used for making the plot of PRESS and COVRATIO statistics.

`influencePoints` Row id of blocks in `influenceData` that lie outside the 95% quantile of COVRATIO statistics and above the 95% quantile of PRESS statistics.

Examples

```
# load data
data(ns.data.re)

ns.data.re$blockid<-paste(ns.data.re$GridCode, ns.data.re$Year, ns.data.re$MonthOfYear,
                    ns.data.re$DayOfMonth, sep=)
ns.data.re$blockid<-as.factor(ns.data.re$blockid)

model<-geeglm(birds ~ observationhour + as.factor(floodebb) + as.factor(impact),
              family=poisson, data=ns.data.re, id=blockid)

timeInfluenceCheck(model, ns.data.re$blockid)

## Not run:
# **WARNING** this example takes a long time
influences<-runInfluence(model, ns.data.re$blockid)

## End(Not run)
```

---

runPartialPlots *Plot partial plots for each of the variables listed in* `factorlist` *or* `varlist`.

Description

Plot partial plots for each of the variables listed in `factorlist` or `varlist`.

Usage

```
runPartialPlots(model, data, factorlist = NULL, varlist = NULL,
  showKnots = FALSE, save = FALSE)
```

Arguments

`model`	Fitted model object (glm or gam)
`data`	Data frame of data information used to fit `model`
`factorlist`	(`default=NULL`). Vector or names of factor variables
`varlist`	(`default=NULL`). Vector of names of continuous variables
`showKnots`	(`default=FALSE`). Logical stating whether knot locations should be plotted.
`save`	(`default=FALSE`). Logical stating whether plot should be saved into working directory.

Value

Partial plots, one for each covariate in `factorlist` and `varlist`

Examples

```
# # load data
data(ns.data.re)

model<-glm(birds ~ observationhour + as.factor(floodebb) + as.factor(impact),
           family=quasipoisson, data=ns.data.re)

runPartialPlots(model, ns.data.re, factorlist=c(floodebb, impact),
                varlist=c(observationhour))
```

---

runSALSA1D *Running SALSA for continuous one-dimensional covariates.*

Description

This function finds spatially adaptive knot locations for one or more continuous one-dimensional covariates.

Usage

```
runSALSA1D(initialModel, salsa1dlist, varlist, factorlist = NULL,
  predictionData, varlist_cyclicSplines = NULL, splineParams = NULL)
```

Arguments

`initialModel`	The best fitting `CReSS` model with no continuous covariates specified
`salsa1dlist`	Vector of objects required for `runSALSA1D`: `fitnessMeasure`, `minKnots_1d`, `maxKnots_1d`, `startKnots_1d` `degree`, `maxIterations` `gap`.
`varlist`	Vector of variable names for the covariates required for knot selection
`factorlist`	vector of factor variables specified in `initialModel`. Specified so that a check can be made that there are non-zero counts in all levels of each factor. Uses the function `checkfactorlevelcounts`. Default setting is NULL.
`predictionData`	The data to be predicted for. column names correspond to the data in `initialModel`
`varlist_cyclicSplines`	Vector of variable names for covariates to be modelled with cyclic cubic splines. This must be a subset of `varlist`.The default is `NULL`
`splineParams`	List object containing information for fitting splines to the covariates in `varlist`. If not specified (`NULL`) this object is created and returned. See `makesplineParams` for details.

Details

There must be a column called `response` in the data, which is the response variable used in the initial model to be fitted.

The object `salsa1dlist` contains parameters for the `runSALSA1D` function.

`fitnessMeasure`. The criterion for selecting the 'best' model. Available options: AIC, AIC_c, BIC, QIC_b.

`minKnots_1d`. Minimum number of knots to be tried.

`maxKnots_1d`. Maximum number of knots to be tried.

`startKnots_1d`. Starting number of knots (spaced at quantiles of the data).

`degree`. The degree of the B-spline. Does not need to be specified if `splineParams` is a parameter in `runSALSA1D`.

`maxIterations`.The exchange/improve steps will terminate after maxIterations if still running.

`gaps`. The minimum gap between knots (in unit of measurement of explanatory).

`minKnots_1d`, `maxKnots_1d`, `startKnots_1d` and `gaps` are vectors the same length as `varlist`. This enables differing values of these parameters for each covariate.

The initial model contains all the factor level covariates and any covariates of interest that are not specified in the `varlist` argument of `runSALSA1D`

Note: The algorithm will not remove variables in `varlist`. If there is no better model than with a knot at the mean, the output will include that covariate with a knot at the mean. The user must decide if the covariate is required in the model as a linear term instead.

Value

A list object is returned containing 4 elements:

`bestModel`	A glm model object from the best model fitted
`modelFits1D`	A list object with an element for each new term fitted to the model. The first element is a model fitted with a knot at the mean for each of the covariates in `varlist`. Within the first element, the model term of interest, the current fit, knots and formula. The second element is the result of SALSA on the first term in `varlist`. Within this element, the knots chosen and the improvement in model fit are presented `$modelfits`. This continues till all covariates in `varlist` have been through SALSA.
`splineParams`	The updated spline parameter object, with the new (if chosen) knot locations for each covariate in `varlist`
`fitstat`	The final fit statistic of `bestModel`. The type of statistic was specified in `salsa1dlist`.

References

Walker, C.; M. Mackenzie, C. Donovan and M. O'Sullivan. SALSA - a Spatially Adaptive Local Smoothing Algorithm. Journal of Statistical Computation and Simulation, 81(2):179-191, 2010

Examples

```
# load data
data(ns.data.re)
# load prediction data
data(ns.predict.data.re)
```

```
splineParams<-makesplineParams(data=ns.data.re, varlist=c(observationhour, DayOfMonth))
#set some input info for SALSA
ns.data.re$response<- ns.data.re$birds

# # set initial model without the spline terms in there
# (so all other non-spline terms)
initialModel<- glm(response ~ as.factor(floodebb) + as.factor(impact) + offset(log(area)),
                   family=quasipoisson,data=ns.data.re)

salsa1dlist<-list(fitnessMeasure = QICb, minKnots_1d=c(2,2), maxKnots_1d = c(20, 20),
                  startKnots_1d = c(2,2), degree=c(2,2), maxIterations = 10, gaps=c(1,1))
# run SALSA
salsa1dOutput<-runSALSA1D(initialModel, salsa1dlist, varlist=c(observationhour, DayOfMonth),
            factorlist=c(floodebb, impact), ns.predict.data.re, splineParams=splineParams)
```

`runSALSA1D_withremoval` *Running SALSA for continuous one-dimensional covariates.*

Description

This function finds spatially adaptive knot locations for one or more continuous one-dimensional covariates. It differs to `runSALSA1D` in that if the CV score of a model does not improve with the addition of a covariate in `varlist` then that term is either reduced to linear or removed from the model.

Usage

```
runSALSA1D_withremoval(initialModel, salsa1dlist, varlist, factorlist = NULL,
  predictionData, varlist_cyclicSplines = NULL, splineParams = NULL)
```

Arguments

`initialModel`	The best fitting `CReSS` model with no continuous covariates specified
`salsa1dlist`	Vector of objects required for `runSALSA1D`: `fitnessMeasure`, `minKnots_1d`, `maxKnots_1d`, `startKnots_1d` `degree`, `maxIterations` `gap`.
`varlist`	Vector of variable names for the covariates required for knot selection
`factorlist`	vector of factor variables specified in `initialModel`. Specified so that a check can be made that there are non-zero counts in all levels of each factor. Uses the function `checkfactorlevelcounts`. Default setting is NULL.
`predictionData`	The data to be predicted for. column names correspond to the data in `initialModel`
`varlist_cyclicSplines`	Vector of variable names for covariates to be modelled with cyclic cubic splines. This must be a subset of `varlist`.The default is `NULL`
`splineParams`	List object containing information for fitting splines to the covariates in `varlist`. If not specified (`NULL`) this object is created and returned. See `makesplineParams` for details.

Details

There must be columns called `response` (response variable) and `foldid` (for cross-validation calculation) in the data used in the initial model to be fitted.

The object `salsa1dlist` contains parameters for the `runSALSA1D` function.

`fitnessMeasure`. The criterion for selecting the 'best' model. Available options: AIC, AIC_c, BIC, QIC_b.

`minKnots_1d`. Minimum number of knots to be tried.

`maxKnots_1d`. Maximum number of knots to be tried.

`startKnots_1d`. Starting number of knots (spaced at quantiles of the data).

`degree`. The degree of the B-spline. Does not need to be specified if `splineParams` is a parameter in `runSALSA1D`.

`maxIterations`.The exchange/improve steps will terminate after maxIterations if still running.

`gaps`. The minimum gap between knots (in unit of measurement of explanatory).

`minKnots_1d`, `maxKnots_1d`, `startKnots_1d` and `gaps` are vectors the same length as `varlist`. This enables differing values of these parameters for each covariate.

The initial model contains all the factor level covariates and any covariates of interest that are not specified in the `varlist` argument of `runSALSA1D`

Note: The algorithm may remove variables in `varlist` (but not the variables in `factorlist`. If there is no better model than with a knot at the mean, the output will include that covariate with a knot at the mean. The best model with a given smooth term is tested both against a model with the term as linear or removed. Cross-Validation is used in the selection process.

Value

A list object is returned containing 4 elements:

`bestModel` A glm model object from the best model fitted

`modelFits1D` A list object with an element for each new term fitted to the model. The first element is a model fitted with a knot at the mean for each of the covariates (startmodel) in `varlist`. Within the first element, the current fit and formula of the start model.

The second element is the result of SALSA on the first term in `varlist`. Within this element:

- `term`: term of interest
- `kept`: Statement of whether the term is kept in the model (yes- initial knots, yes - new knots, yes -linear or no)
- `basemodelformula`: the resulting model formula. If `kept=yes` or `kept=linear` then the term of interest is included in the model otherwise it is removed.
- `knotSelected`: the knots chosen for the term of interest (NA if term removed or linear)
- `baseModelFits`: fit statistics for the resulting formula
- `modelfits`: fit statistics for the model with the term included (same as resulting formula if `kept=yes`)

This continues till all covariates in `varlist` have been through SALSA.

`splineParams` The updated spline parameter object, with the new (if chosen) knot locations for each covariate in `varlist`

`fitstat` The final fit statistic of `bestModel`. The type of statistic was specified in `salsa1dlist`.

`keptvarlist` The covariates from `varlist` that have been retained in the model

References

Walker, C.; M. Mackenzie, C. Donovan and M. O'Sullivan. SALSA - a Spatially Adaptive Local Smoothing Algorithm. Journal of Statistical Computation and Simulation, 81(2):179-191, 2010

Examples

```
# load data
data(ns.data.re)
# load prediction data
data(ns.predict.data.re)

splineParams<-makesplineParams(data=ns.data.re, varlist=c(observationhour, DayOfMonth))

# make column with foldid for cross validation calculation
ns.data.re$blockid<-paste(ns.data.re$GridCode, ns.data.re$Year, ns.data.re$MonthOfYear, ns.data.re$DayOfMont
ns.data.re$blockid<-as.factor(ns.data.re$blockid)
ns.data.re$foldid<-getCVids(ns.data.re, folds=5, block=blockid)

# # set initial model without the spline terms in there
# (so all other non-spline terms)
ns.data.re$response<- ns.data.re$birds
initialModel<- glm(response ~ as.factor(floodebb) + as.factor(impact) + offset(log(area)),
                   family=quasipoisson,data=ns.data.re)

#set some input info for SALSA
salsa1dlist<-list(fitnessMeasure = QICb, minKnots_1d=c(2,2), maxKnots_1d = c(5, 5),
                  startKnots_1d = c(2,2), degree=c(2,2), maxIterations = 10, gaps=c(1,1))

# run SALSA
salsa1dOutput<-runSALSA1D_withremoval(initialModel, salsa1dlist, varlist=c(observationhour, DayOfMonth),
           factorlist=c(floodebb, impact), ns.predict.data.re, splineParams=splineParams)
```

`runSALSA2D` *Running SALSA for a spatial smooth with a CReSS basis*

Description

This function fits a spatially adaptive two dimensional smooth of spatial coordinates with knot number and location selected by SALSA.

Usage

```
runSALSA2D(model, salsa2dlist, d2k, k2k, splineParams = NULL, tol = 0)
```

Arguments

`model`	A model with no spatial smooth
`salsa2dlist`	Vector of objects required for `runSALSA2D`: `fitnessMeasure`, `knotgrid`, `startKnots`, `minKnots`, codemaxKnots, `r_seq`, `gap`, `interactionTerm`.
`d2k`	(n x k) Matrix of distances between all data points in `model` and all valid knot locations specified in `knotgrid` #'
`k2k`	(k x k) Matrix of distances between all valid knot locations specified in `knotgrid`

splineParams (default =NULL) List object containng output from runSALSA (e.g. knot locations for continuous covariates)

Details

There must be a column called response in the data, which is the response variable used in the initial model to be fitted.

The object salsa2dlist contains parameters for the runSALSA2D function.

fitnessMeasure. The criterion for selecting the 'best' model. Available options: AIC, AIC_c, BIC, QIC_b.

knotgrid. A grid of legal knot locations. Must be a regular grid with c(NA, NA) for rows with an illegal knot. An illegal knot position may be outside the study region or on land for a marine species for example.

knotdim. The dimensions of the knot grid as a vector. (x, y)

startknots. Starting number of knots (initialised as spaced filled locations).

minKnots. Minimum number of knots to be tried.

maxKnots. Maximum number of knots to be tried.

r_seq. Sequence of range parameters for the CReSS basis from local (small) to global (large). Determines the range of the influence of each knot. Sequence made using getRadiiChoices.

gap. The minimum gap between knots (in unit of measurement of coordinates). interactionTerm. Specifies which term in baseModel the spatial smooth will interact with. If NULL no interaction term is fitted.

Value

The spline paramater object that is return now contains a list in the first element (previously reserved for the spatial component). This list contains the objects required for the SALSA2D fitting process:

knotDist	Matrix of knot to knot distances (k x k). May be Euclidean or geodesic distances. Must be square and the same dimensions as nrows(na.omit(knotgrid)). Created using makeDists.
radii	Sequence of range parameters for the CReSS basis from local (small) to global (large). Determines the range of the influence of each knot.
dist	Matrix of distances between data locations and knot locations (n x k). May be Euclidean or geodesic distances. Euclidean distances created using makeDists.
gridresp	The first column of knotgrid.
grid	Index of knotgrid locations. Should be same length as knotgrid but with x=integer values from 1 to number of unique x-locations and y= integer values from 1 to number of unique y-locations.
datacoords	Coordinates of the data locations
response	Vector of response data for the modelling process
knotgrid	Grid of legal knot locations. Must be a regular grid with c(NA, NA) for rows with an illegal knot.
minKnots	Minimum number of knots to be tried.
maxKnots	Maximum number of knots to be tried.
gap	Minimum gap between knots (in unit of measurement of datacoords)

radiusIndices	Vector of length startKnots identifying which radii (splineParams[[1]]$radii) will be used for each knot location (splineParams[[1]]$knotPos)
knotPos	Index of knot locations. The index identifies which knots (i.e. which rows) from knotgrid were selected by SALSA
invInd	This is a vector of length the number of rows of knotgrid. It is used to translate between knotgrid (used in SALSA) and na.omit(knotgrid) (used in dist and LocalRadialFunction).

Author(s)

Cameron Walker

References

Scott-Hayward, L.; M. Mackenzie, C.Donovan, C.Walker and E.Ashe. Complex Region Spatial Smoother (CReSS). Journal of computational and Graphical Statistics. 2013. doi: 10.1080/10618600.2012.762920

Scott-Hayward, L.. Novel Methods for species distribution mapping including spatial models in complex regions: Chapter 5 for SALSA2D methods. PhD Thesis, University of St Andrews. 2013

Examples

```
# load data
data(ns.data.re)
# load prediction data
data(ns.predict.data.re)
# load knot grid data
data(knotgrid.ns)

splineParams<-makesplineParams(data=ns.data.re, varlist=c(observationhour))

#set some input info for SALSA
ns.data.re$response<- ns.data.re$birds

# make distance matrices for datatoknots and knottoknots
distMats<-makeDists(cbind(ns.data.re$x.pos, ns.data.re$y.pos), na.omit(knotgrid.ns))

# choose sequence of radii
r_seq<-getRadiiChoices(8, distMats$dataDist)

# set initial model without the spatial term
# (so all other non-spline terms)
initialModel<- glm(response ~ as.factor(floodebb) + as.factor(impact) + offset(log(area)),
                   family=quasipoisson, data=ns.data.re)

# make parameter set for running salsa2d
salsa2dlist<-list(fitnessMeasure = QICb, knotgrid = knotgrid.ns, knotdim = c(7, 9),
                  startKnots=6, minKnots=4, maxKnots=20, r_seq=r_seq, gap=1,
                  interactionTerm="as.factor(impact)")

salsa2dOutput_k6<-runSALSA2D(initialModel, salsa2dlist, d2k=distMats$dataDist,
                             k2k=distMats$knotDist, splineParams=splineParams)
```

timeInfluenceCheck *Timing check to see how long it will take to run* runInfluence.

Description

Timing check to see how long it will take to run runInfluence.

Usage

```
timeInfluenceCheck(model, id, d2k = NULL, splineParams = NULL)
```

Arguments

model	Fitted model object (glm or gam)
id	blocking structure
d2k	(default=NULL). (n x k) Matrix of distances between all data points in model and all valid knot locations.
splineParams	(default=NULL). List object containng output from runSALSA (e.g. knot locations for continuous covariates). See makesplineParams for more details of this object.

Examples

```
# load data
data(ns.data.re)

ns.data.re$blockid<-paste(ns.data.re$GridCode, ns.data.re$Year, ns.data.re$MonthOfYear,
                    ns.data.re$DayOfMonth, sep=)
ns.data.re$blockid<-as.factor(ns.data.re$blockid)
model<-geeglm(birds ~ observationhour + as.factor(floodebb) + as.factor(impact),
               family=poisson, data=ns.data.re, id=blockid)

timeInfluenceCheck(model, ns.data.re$blockid)
```

which.bin *Determining the distance bin*

Description

For a vector of perpendicular (or radial) distances, this function determines which distance bin it belongs to (given the input of cut points) and adds the beginning and end points of the respective distance bins in new colunns in dis.data called "distbegin" and "distend".

Usage

```
which.bin(dis.data, cutpoints)
```

Arguments

dis.data	A data frame with distance data for which perpendicular (or radial) distances are recorded in the distance column
cutpoints	A vector of cut points of the intervals (this function is not set up to deal with left-truncation)

Details

If a value in dis.data$distance matches a cut point in cutpoints exactly, the value of dis.data.re$distance will be attributed to the bin that is closer to the line/point unless the value of dis.data.re$distance is 0.

E.g. if cutpoints=c(0,1,2,3), dis.data$distance=2 will be attributed to interval 2 (and not 3).

Value

The dis.data data frame to which columns "distbegin" and "distend" were added giving the beginning and end cutpoints of the bin that the respective dis.data$distance belongs to.

Index