# Package 'gamair'

July 2, 2014

**Version** 0.0-8

**Author** Simon Wood <simon.wood@r-project.org>

**Maintainer** Simon Wood <simon.wood@r-project.org>

**Title** Data for ``GAMs: An Introduction with R"

**Description** Data sets and scripts used in the book ``Generalized Additive Models: An Introduction with R", Wood (2006) CRC.

**Depends** R (>= 2.10)

**License** GPL (>= 2)

**Repository** CRAN

**Date/Publication** 2012-03-27 14:03:03

**NeedsCompilation** no

## R topics documented:

---

gamair-package *Data + scripts for 'GAMs: An Introduction with R'*

---

## Description

This package contains the data sets used in the book *Generalized Additive Models: An Introduction with R*, which covers linear and generalized linear models, GAMs as implemented in package mgcv and mixed model extensions of these.

The script files for the book can be found in the 'scripts' folder of the 'inst' folder of the source package. They have been modified slightly to work with recent versions of mgcv (e.g. >= 1.7-0).

## Details

Each dataset has its own help page, which describes the dataset, and gives the original source and associated references. All datasets have been reformatted into standard R data frames. Some smaller datasets from the book have not been included. Datasets from other R packages have not been included, with the exception of a distillation of one set from the NMMAPSdata package.

Index:

| | |
|---|---|
| bird | Bird distribution data from Portugal |
| blowfly | Nicholson's Blowfly data |
| brain | Brain scan data |
| cairo | Daily temperature data for Cairo |
| chicago | Chicago air pollution and death rate data |
| chl | Chlorophyll data |
| co2s | Atmospheric CO2 at South Pole |
| coast | European coastline from -11 to 0 East and from 43 to 59 North. |
| engine | Engine wear versus size data |
| gamair | Generalized Additive Models: An Introduction With R |
| harrier | Hen Harriers Eating Grouse |
| hubble | Hubble Space Telescope Data |
| ipo | Initial Public Offering Data |
| mack | Egg data from 1992 mackerel survey |
| mackp | Prediction grid data for 1992 mackerel egg model |
| sole | Sole Eggs in the Bristol Channel |
| sperm.comp1 | Sperm competition data I |
| sperm.comp2 | Sperm competition data II |
| wine | Bordeaux Wines |

## Author(s)

Simon Wood <simon@r-project.org>

Maintainer: Simon Wood <simon@r-project.org>

## References

Wood, S.N. (2006) *Generalized Additive Models: An Introduction with R*, CRC

## See Also

mgcv

## Examples

```
library(help=gamair)
```

---

# bird *Bird distribution data from Portugal*

---

## Description

Data from the compilation of the Portuguese Atlas of Breeding Birds.

## Usage

```
data(bird)
```

## Format

A data frame with 6 columns and 25100 rows. Each row refers to one 2km by 2km square (tetrad). The columns are:

- QUADRICULAAn identifier for the 10km by 10km square that this tetrad belongs to.
- crestlarkWere crested lark (or possibly thekla lark!) found (1), not found (0) breading in this tetrad, or was the tetrad not visited (NA).
- linnetAs crestlark, but for linnet.
- xlocation of tetrad (km east of an origin).
- ylocation of tetrad (km north of an origin).

## Details

At least 6 tetrads from each 10km square were visited, to establish whether each species was breeding there, or not. Each Tetrad was visited twice for one hour each visit. These data are not definitive: at time of writing the fieldwork was not quite complete.

The data were kindly supplied by Jose Pedro Granadeiro.

## Source

The Atlas of the Portuguese Breeding Birds.

## References

Wood, S.N. (2006) Generalized Additive Models: An Introduction with R

## Examples

```
data(bird)
species <- "crestlark"
op<-par(bg="white",mfrow=c(1,1),mar=c(5,5,1,1))
ind <- bird[[species]]==0&!is.na(bird[[species]])
plot(bird$y[ind]/1000,1000-bird$x[ind]/1000,pch=19,cex=.3,col="white",
   ylab="km west",xlab="km north",cex.lab=1.4,cex.axis=1.3,type="n")
polygon(c(4000,4700,4700,4000),c(250,250,600,600),col="grey",border="black")
points(bird$y[ind]/1000,1000-bird$x[ind]/1000,pch=19,cex=.3,col="white")
ind <- bird[[species]]==1&!is.na(bird[[species]])
with(bird,points(y[ind]/1000,1000-x[ind]/1000,pch=19,cex=.3))
par(op)
```

---

# blowfly *Nicholson's Blowfly data*

---

## Description

Data on a laboratory population of Blowfies, from the classic ecological studies of Nicholson.

## Usage

```
data(blowfly)
```

## Format

A data frame with 2 columns and 180 rows. The columns are:

- popCounts (!) of the population of adult Blowflies in one of the experiments.
- dayDay of experiment.

## Details

The population counts are actually obtained by counting dead blowflies and back calculating.

## References

Nicholson, A.J. (1954a) Compensatory reactions of populations to stresses and their evolutionary significance. Australian Journal of Zoology 2, 1-8.

Nicholson, A.J. (1954b) An outline of the dynamics of animal populations. Australian Journal of Zoology 2, 9-65.

Wood, S.N. (2006) Generalized Additive Models: An Introduction with R

## Examples

```
data(blowfly)
with(blowfly,plot(day,pop,type="l"))
```

---

`brain` *Brain scan data*

---

## Description

Functional magnetic resonance imaging measurements for a human brain subject to a particular experimental stimulus. One slice of the image is provided, described as a near-axial slice through the dorsal cerebral cortex.

## Usage

```
data(brain)
```

## Format

A data frame with 5 columns and 1567 rows. Each row refers to one 'voxel' of the image. The columns are:

- X voxel position on horizontal axis.
- Y voxel position on vertical axis.
- medFPQ median of three replicate 'Fundamental Power Quotient' values at the voxel: this is the main measurement of brain activivity.
- regioncode indicating which of several regions of the brain the voxel belongs to. The regions are defined by the experimenters. `0` is the base region; `1` is the region of interest; `2` is the region activated by the experimental stimulus; `NA` denotes a voxel with no allocation.
- meanThetamean phase shift at the Voxel, over three measurments.

## Details

See the source article for fuller details.

## Source

S. Landau et al (2003) 'Tests for a difference in timing of physiological response between two brain regions measured by using functional magnetic resonance imaging'. Journal of the Royal Statistical Society, Series C, Applied Statistics, 53(1):63-82

cairo *Daily temperature data for Cairo*

## Description

The average air temperature (F) in Cairo from Jan 1st 1995.

## Usage

```
data(cairo)
```

## Format

A data frame with 6 columns and 3780 rows. The columns are:

- monthmonth of year from 1 to 12.
- day.of.monthday of month, from 1 to 31.
- yearYear, starting 1995.
- tempAverage temperature (F).
- day.of.yearDay of year from 1 to 366.
- timeNumber of days since 1st Jan 1995.

## Source

http://www.engr.udayton.edu/weather/citylistWorld.htm

## References

Wood, S.N. (2006) Generalized Additive Models: An Introduction with R

## Examples

```
data(cairo)
with(cairo,plot(time,temp,type="l"))
```

chicago *Chicago air pollution and death rate data*

## Description

Daily air pollution and death rate data for Chicago.

## Usage

```
data(chicago)
```

## Format

A data frame with 7 columns and 5114 rows. Each row refers to one day. The columns are:

- death total deaths (per day).
- pm10medianmedian particles in 2.5-10 per cubic m
- pm25medianmedian particles < 2.5 mg per cubic m (more dangerous).
- o3medianOzone in parts per billion
- so2medianMedian Sulpher dioxide measurement
- timetime in days
- tmpdtemperature in fahrenheit

## Details

See the NMMAPSdata package for fuller details. Note that there are missing values in some fields.

## Source

Roger D. Peng, Leah J. Welty and Aiden McDermott. R package NMMAPSdata.

## References

Peng, R.D. and Welty, L.J. (2004) The NMMAPSdata package. R News 4(2).

Wood, S.N. (2006) Generalized Additive Models: An Introduction with R

http://www.ihapss.jhsph.edu/data/NMMAPS/R/

chl *Chlorophyll data*

## Description

Data relating to the callibration of remote sensed satellite data. The SeaWifs satellite provides estimates of chlorophyll concentration at the ocean surface from measurements of ocean surface colour. It is of interest to attempt to use these data to predict direct bottle measurements of chl. conc.

## Usage

```
data(chl)
```

## Format

A data frame with 6 columns and 13840 rows. The columns are:

- lonlongitude
- latlatitude
- jul.dayJulian day (i.e. day of year starting at Jan 1st.)
- bathOcean depth in metres.
- chldirect chlorophyll concentration measured at given location from a bottle sample.
- chl.swchl. conc. as measured by Seawifs Satellite

## Source

http://seawifs.gsfc.nasa.gov/SEAWIFS

and the World Ocean Database.

## References

Wood, S.N. (2006) Generalized Additive Models: An Introduction with R. CRC

## Examples

```
data(chl)
with(chl,plot(chl,chl.sw))
```

co2s *Atmospheric CO2 at South Pole*

## Description

Monthly CO2 concentration in parts per million at the South Pole.

## Usage

```
data(co2s)
```

## Format

A data frame with 3 columns and 507 rows. The columns are:

- co2atmospheric CO2 concentration in parts per million
- c.monthcumulative number of months since Jan 1957
- monthmonth of year

## Source

http://cdiac.esd.ornl.gov/trends/co2/sio-spl.htm

## References

Keeling C.P. and T.P Whorf (2000) Atmospheric CO2 records from sites in the SIO air sampling network. In Trends: A Compedium of Data on Global Change. Carbon Dioxide Analyis Center, Oak Ridge National Laboratory, U.S. Department of Energy, Oak Ridge Tenn., USA

Wood, S.N. (2006) Generalized Additive Models: An Introduction with R. CRC

## Examples

```
data(co2s)
with(co2s,plot(c.month,co2,type="l",ylab=
expression(paste(CO[2]," in ppm.")),xlab="Month since Jan. 1957"))
```

coast *European coastline from -11 to 0 East and from 43 to 59 North*

## Description

The data are longitudes (degrees E) and latitudes (degrees N) defining points that can be joined up to get the European coastline in the rectangle (-11E,43N)-(0E,59N). Discontinuous sections of coast are separated by NA's.

## Usage

```
data(coast)
```

## Format

A data frame with 2 columns.

- lonLongitude in degrees East for points used to define the coast.
- latLatitude in degrees North for points used to define the coast.

## Details

`lon`, `lat` together define the co-ordinates of points that can be joined up in order to plot the coastline. The original data come from the NOAA www site given below, but have been substantially thinned, to a much lower resultion than the source.

## Author(s)

Simon Wood.

## References

http://rimmer.ngdc.noaa.gov/coast/

## Examples

```
data(coast)
# plot the entire coast .....
plot(coast$lon,coast$lat,type="l")
# or draw it clipped to  whatever the current plot is....
lines(coast$lon,coast$lat,col="blue")
```

## engine *Engine wear versus size data*

### Description

Data on engine wear against engine size for 19 Volvo car engines.

### Usage

```
data(engine)
```

### Format

A data frame with 2 columns and 19 rows. Each row refers to one engine model. The columns are:

- wearan index of engine wear rate.
- sizecylinder capacity in litres.

### Details

See the source for further details.

### Source

http://www3.bc.sympatico.ca/Volvo_Books/engine3.html

## harrier *Hen Harriers Eating Grouse*

### Description

Data on the rate at which Hen Harriers consume Grouse as a function of Grouse density.

### Usage

```
data(harrier)
```

### Format

A data frame with 2 columns and 37 rows. The columns are:

- Grouse.DensityDensity of Grouse per square kilometre.
- Consumption.RateNumber of Grouse consumed per Hen Harrier per day.

### Details

Data have been read from Figure 1 of Asseburg et al. (2005)

## Source

Asseburg, C., S. Smout, J. Matthiopoulos, C. Fernandez, S. Redpath, S. Thirgood and J. Harwood (2005) The functional response of a generalist predator. Web preprint

## References

Wood, S.N. (2006) Generalized Additive Models: An Introduction with R. CRC

## Examples

```
data(harrier)
with(harrier,plot(Grouse.Density,Consumption.Rate))
```

---

# hubble *Hubble Space Telescope Data*

---

## Description

Data on distances and velocities of 24 galaxies containing Cepheid stars, from the Hubble space telescope key project to measure the Hubble constant.

## Usage

```
data(hubble)
```

## Format

A data frame with 3 columns and 24 rows. The columns are:

- Galaxy A (factor) label identifying the galaxy.
- y The galaxy's relative velocity in kilometres per second.
- x The galaxy's distance in Mega parsecs. 1 parsec is 3.09e13 km.

## Details

Cepheids are variable stars which have a known relationship between brightness and period. Hence the distance to galaxies containing these stars can be estimated from the observed brightness of the Cepheid, relative to its absolute brightness as predicted by its period. The velocity of the galaxy can be estimated from its mean red-shift.

The data can be used to get a reasonably good idea of the age of the universe A data free alternative estimate of 6000 years is given in the reference (not the source!).

## Source

Tables 4 and 5 of Freedman et al. 2001. The Astrophysical Journal 553:47-72

## References

Freedman et al. (2001) Final results from the Hubble space telescope key project to measure the Hubble constant. The Astrophysical Journal (553), 47-72.

http://www.icr.org/pubs/imp/imp-352.htm

NUCLEAR DECAY: EVIDENCE FOR A YOUNG WORLD - IMPACT No. 352 October 2002 by D. Russell Humphreys, Ph.D.

Wood, S.N. (2006) Generalized Additive Models: An Introduction with R. CRC

---

`ipo` *Initial Public Offering Data*

---

## Description

Data on the relationship between the number of initial public offerings (of shares in a company) and other potentially important variables. It is probably necessary to lag some of the explanatory variables.

## Usage

```
data(ipo)
```

## Format

A data frame with 6 columns and 156 rows. The columns are:

- n.iponumber of initial pubilc offerings each month.
- irthe average initial return (volume weighted): this is the percentage difference between the offer proce of shares and the price after the first day of trading.
- dpthe average percentage difference between middle of the price range proposed at first filing of the IPO, and the eventual offer price.
- reg.tthe average time between filing and offer.
- ttime, in months.
- n.ipomonth of the year (1 = January).

## Source

http://schwert.ssb.rochester.edu

## References

Lowry, M. and G.W. Schwert (2002) IPO market cycles: Bubbles or sequential learning? The Journal of Finance 67(3), 1171-1198

Wood, S.N. (2006) Generalized Additive Models: An Introduction with R. CRC

## Examples

```
data(ipo)
pairs(ipo)
```

---

`mack` *Egg data from 1992 mackerel survey*

---

## Description

The data relate to the distribution of mackerel eggs and were collected as part of the 1992 mackerel survey aimed at assessing the mackerel spawning stock biomass using the daily egg production method.

## Usage

```
data(mack)
```

## Format

A data frame with 16 columns. Each row corresponds to one sample of eggs.

- egg.countThe number of stage I eggs in this sample.
- egg.densThe number of stage I eggs per square metre of sea surface, produced per day. This is calculated from `egg.count` and other information about sampling net size, and egg stage duration.
- b.depthThe sea bed depth at the sampling location.
- c.distThe distance from the sample location to the 200m contour measured in degrees as if degrees latitude equalled degrees longitude, which actually they don't.
- lonThe longitude of the sample station in degrees east.
- latThe latitude of the sample station in degrees north.
- timeThe time of day (in hours) at which the sample was taken.
- salinityThe salinity (saltiness) of the water at the sampling location.
- flowReading from the flow meter attached to the sampling net - used for calibration.
- s.depthThe depth that the sampling net started sampling from (the net is dropped to this depth and then hauled up to the surface, filtering eggs etc out of the water as it goes).
- temp.surfThe temperature at the sea surface at the sampling location.
- temp.20mThe temperature 20m down at the sampling location.
- net.areaThe area of the sampling net in square metres.
- countryA code identifying the country responsible for the boat that took this sample.
- vesselA code identifying the boat that took this sample.
- vessel.haulA code uniquely identifying this sample, given that the vessel is known.

## Details

At each of a number of stations located as defined in `lon` and `lat`, mackerel eggs were sampled by hauling a fine net up from deep below the sea surface to the sea surface. The egg count data are obtained from the resulting samples, and these have been converted to (stage I) eggs produced per metre squared per day - the egg density data. Other possibly useful predictor variable information has been recorded, along with identification information, and some information that is probably useless!

## Source

The data are effectively a combination of datasets `mackerel` and `smacker` from the `sm` library. They were originally analyzed using GAMs by:

Borchers, D.L., S.T. Buckland, I.G. Priede and S. Ahmadi (1997) "Improving the precision of the daily egg production method using generalized additive models". Can. J. Fish. Aquat. Sci. 54:2727-2742.

## Examples

```
data(mack)
# plot the egg densities against location
plot(mack$lon,mack$lat,cex=0.2+mack$egg.dens/150,col="red")
```

---

# mackp *Prediction grid data for 1992 mackerel egg model*

---

## Description

This data frame provides a regular grid of values of some predictor variables useful for modelling mackerel egg abundances. Its main purpose is to enable mackerel egg densities to be predicted over a regular spatial grid within the area covered by the 1992 mackerel egg survey (see `mack`), using a fitted generalised additive model.

## Usage

```
data(mackp)
```

## Format

A data frame with 5 columns. Each row corresponds to one spatial location within the survey area. The columns are as follows:

- lonLongitude of the gridpoint in degrees east
- latLatitude of the gridpoint in degrees north.
- b.depthThe sea bed depth at the gridpoint.
- c.distThe distance from the gridpoint to the 200m sea bed depth contour.
- salinitySalinity interpolated onto the grid (from `mack` measurements).

- temp.surfSurface temperature interpolated onto grid (from mack data).
- temp.20mTemperature at 20m interpolated from mack data.
- area.index An indexing vector that enables straightforward copying of the other variables into a matrix suitable for plotting against longitude and lattitude using image(). See the example below.

## Details

The grid is defined on a series of 1/4 degree lon-lat squares.

## References

Borchers, D.L., S.T. Buckland, I.G. Priede and S. Ahmadi (1997) "Improving the precision of the daily egg production method using generalized additive models". Can. J. Fish. Aquat. Sci. 54:2727-2742.

## Examples

```
## example of how to use `area.index' to paste gridded info.
## into a square grid (of NA's) for plotting
data(mackp)
lon<-seq(-15,-1,1/4);lat<-seq(44,58,1/4)
zz<-array(NA,57*57)
zz[mackp$area.index]<-mackp$b.depth
image(lon,lat,matrix(zz,57,57))
```

---

# sole *Sole Eggs in the Bristol Channel*

---

## Description

Data on Sole Egg densities in the Bristol Channel (West Coast of England, UK.) The data are from 5 research cruises undertaken for the purpose of measuring Sole egg densities. Samples were taken at each of a number of sampling stations, by hauling a net vertically through the water column. Sole eggs were counted and assigned to one of four developmental stages.

## Usage

```
data(sole)
```

## Format

A data frame with 7 columns and 1575 rows. The columns are:

- lalatitude of sampling station
- lolongitude of sampling station
- ttime of sampling station: actually time of midpoint of the cruise on which this sample was taken. Measured in Julian days (days since January 1st).

- eggsegg density per square metre of sea surface.
- stageto which of 4 stages the sample relates.
- a.0lower age limit for the stage (i.e. age of youngest possible egg in this sample).
- a.1upper age limit of this stage (i.e. age of oldest possible egg in sample).

## Source

Dixon (2003)

## References

Dixon, C.E. (2003) Multi-dimensional modelling of physiologically and temporally structured populations. PhD thesis. University of St Andrews

Horwood, J. (1993) The Bristol Channel Sole (solea solea (L.)): A fisheries case study. Advances in Marine Biology 29, 215-367

Horwood, J. and M. Greer Walker (1990) Determinacy of fecundity in Sole (solea solea) from the Bristol Channel. Journal of the Marine Biology Association of the United Kingdom. 70, 803-813.

Wood (2006) Generalized Additive Models: An Introduction with R. CRC

## Examples

```
data(sole)
data(coast)
par(mfrow=c(2,3))
sample.t <- unique(sole$t)
stage <- 1
for (i in 1:5)
{ egg<-sole[sole$stage==stage&sole$t==sample.t[i],]
  plot(egg$lo,egg$la,xlab="lo",ylab="la",main=paste("day",sample.t[i]),cex=egg$eggs/4,
  xlim=range(sole$lo),ylim=range(sole$la),cex.axis=1.5,cex.lab=1.5,cex.main=1.5)
  points(egg$lo,egg$la,pch=".",col=2)
  lines(coast)
}
```

---

# sperm.comp1 *Sperm competition data I*

---

## Description

Data relating sperm count to time since last inter-pair copulation and proportion of that time spent together for 15 couples living in Manchester UK.

## Usage

```
data(sperm.comp1)
```

## Format

A data frame with 4 columns and 15 rows. The columns are:

- subject An identifier for the subject/couple.
- time.ipc Time since last inter-pair copulation, in hours.
- prop.partner Proportion of `time.ipc` that the couple had spent together.
- countSperm count in millions.

## Details

The sperm counts reported are total counts in ejaculate from a single copulation, for each of 15 couples. Also recorded are the time since the couple's previous copulation, and the proportion of that time that the couple had spent together. The data are from volunteers from Manchester University and were gathered to test theories about human sperm competition. See the source article for further details.

## Source

Baker, RR and Bellis M.A. (1993) 'Human sperm competition: ejaculate adjustment by males and the function of masturbation'. Animal behaviour 46:861-885

---

`sperm.comp2` *Sperm competition data II*

---

## Description

Data relating average number of sperm ejaculated per copulation to physical characterisics of partners involved, for 24 heterosexual couples from Manchester, UK.

## Usage

```
data(sperm.comp2)
```

## Format

A data frame with 10 columns and 24 rows. The columns are:

- pair an identifier for the couple. These labels correspond to those given in `sperm.comp1`.
- n the number of copulations over which the average sperm count has been calculated.
- count the average sperm count in millions, per copulation.
- f.age age of the female, in years.
- f.height height of the female, in cm.
- f.weightweight of the female, in kg.
- m.ageage of the male, in years.
- m.heightheight of the male, in cm.
- m.weightweight of the male, in kg.
- m.volvolume of one male teste in cubic cm.

### Details

In the source article, these data are used to argue that males invest more reproductive effort in heavier females, on the basis of regression modelling. It is worth checking for outliers.

### Source

Baker, RR and Bellis M.A. (1993) 'Human sperm competition: ejaculate adjustment by males and the function of masturbation'. Animal behaviour 46:861-885

---

## stomata *Stomatal area and C02*

---

### Description

Fake data on average stomatal area for 6 trees grown under one of two C02 concentrations

### Usage

```
data(stomata)
```

### Format

A data frame with 3 columns and 24 rows. The columns are:

- areamean stomatal area.
- C02label for which C02 treatment the measurement relates to.
- treelabel for individual tree.

### Details

The context for these simulated data is given in section 6.1 of the source book.

### Source

The reference.

### References

Wood, S.N. (2006) Generalized Additive Models: An Introduction with R. CRC

wine *Bordeaux Wines*

## Description

Data on prices and growing characteristics of 25 Bordeaux wines from 1952 to 1998.

## Usage

```
data(wine)
```

## Format

A data frame with 7 columns and 47 rows. The columns are:

- yearyear of production
- priceaverage price of the wines as a percentage of the 1961 price.
- h.rainmm of rain in the harvest month.
- s.tempAverage temperature (C) over the summer preceding harvest.
- w.rainmm of rain in the winter preceding harvest.
- h.tempaverage temperature (C) at harvest.
- parkera rating of the wine quality (see source for details).

## Source

http://schwert.ssb.rochester.edu/a425/a425.htm

## References

Wood, S.N. (2006) Generalized Additive Models: An Introduction with R. CRC

## Examples

```
data(wine)
pairs(wine[,-7])
```

# Index