

# Package ‘TH.data’

January 20, 2025

**Title** TH's Data Archive

**Date** 2025-01-17

**Version** 1.1-3

**Description** Contains data sets used in other packages Torsten Hothorn maintains.

**Depends** R (>= 3.5.0), survival, MASS

**Suggests** trtf, tram, rms, coin, ATR, multcomp, gridExtra, vcd, colorspace, lattice, knitr

**LazyData** yes

**VignetteBuilder** knitr

**License** GPL-3

**NeedsCompilation** no

**Author** Torsten Hothorn [aut, cre]

**Maintainer** Torsten Hothorn <Torsten.Hothorn@R-project.org>

**Repository** CRAN

**Date/Publication** 2025-01-17 18:20:02 UTC

## Contents

birds . . . . .	2
bodyfat . . . . .	3
GBSG2 . . . . .	4
geyser . . . . .	5
GlaucomaM . . . . .	6
mammoexp . . . . .	9
mn6.9 . . . . .	9
sphase . . . . .	10
Westbc . . . . .	11
wpbc . . . . .	12

<b>Index</b>	<b>15</b>
--------------	-----------

---

birds

*Habitat Suitability for Breeding Bird Communities*

---

### **Description**

Environmental variables and bird counts for identifying suitable bird habitats

### **Usage**

```
data("birds")
```

### **Format**

A data frame with 258 observations on the following 10 variables.

GST Growing stock per grid

DBH Mean diameter of the largest three trees

AOT Age of oldest tree

AFS Age of forest stand

DWC Amount of dead wood of conifers

LOG Amount of logs per grid

x\_gk grid location, x coordinate

y\_gk grid location, y coordinate

SG4 observed number of birds from structural gild 4: Requirement of regeneration (Phylloscopus trochilus, Aegithalos caudatus)

SG5 observed number of birds from structural gild 5: Requirement of regeneration combined with planted conifers (Phylloscopus collybita, Turdus merula, Sylvia atricapilla).

### **Details**

Counts of breeding bird communities collected at 258 observation plots in a northern Bavarian forest district are the response variable of interest. Along with the number of birds in two structural gilds, 6 covariates are given here and one is interested in quantifying their impact on habitat suitability.

### **Source**

Joerg Mueller (2005). Forest structures as key factor for beetle and bird communities in beech forests. PhD thesis, Munich University of Technology.

### **References**

Thomas Kneib and Joerg Mueller and Torsten Hothorn (2008), Spatial smoothing techniques for the assessment of habitat suitability, *Environmental and Ecological Statistics*, **15**(3), 343–364.

---

bodyfat	<i>Prediction of Body Fat by Skinfold Thickness, Circumferences, and Bone Breadths</i>
---------	--

---

**Description**

For 71 healthy female subjects, body fat measurements and several anthropometric measurements are available for predictive modelling of body fat.

**Usage**

```
data("bodyfat")
```

**Format**

A data frame with 71 observations on the following 10 variables.

age age in years.

DEXfat body fat measured by DXA, response variable.

waistcirc waist circumference.

hipcirc hip circumference.

elbowbreadth breadth of the elbow.

kneebreadth breadth of the knee.

anthro3a sum of logarithm of three anthropometric measurements.

anthro3b sum of logarithm of three anthropometric measurements.

anthro3c sum of logarithm of three anthropometric measurements.

anthro4 sum of logarithm of three anthropometric measurements.

**Details**

Garcia et al. (2005) report on the development of predictive regression equations for body fat content by means of common anthropometric measurements which were obtained for 71 healthy German women. In addition, the women's body composition was measured by Dual Energy X-Ray Absorptiometry (DXA). This reference method is very accurate in measuring body fat but finds little applicability in practical environments, mainly because of high costs and the methodological efforts needed. Therefore, a simple regression equation for predicting DXA measurements of body fat is of special interest for the practitioner. Backward-elimination was applied to select important variables from the available anthropometrical measurements, and Garcia (2005) report a final linear model utilizing hip circumference, knee breadth and a compound covariate which is defined as the sum of log chin skinfold, log triceps skinfold and log subscapular skinfold.

**Source**

Ada L. Garcia, Karen Wagner, Torsten Hothorn, Corinna Koebnick, Hans-Joachim F. Zunft and Ulrike Trippo (2005), Improved prediction of body fat by measuring skinfold thickness, circumferences, and bone breadths. *Obesity Research*, **13**(3), 626–634.

Peter Buehlmann and Torsten Hothorn (2007), Boosting algorithms: regularization, prediction and model fitting. *Statistical Science*, **22**(4), 477–505.

Benjamin Hofner, Andreas Mayr, Nikolay Robinzonov and Matthias Schmid (2012). Model-based Boosting in R: A Hands-on Tutorial Using the R Package mboost. *Computational Statistics*.  
doi:[10.1007/s0018001203825](https://doi.org/10.1007/s0018001203825)

Available as vignette via: `vignette(package = "mboostDevel", "mboost_tutorial")`

**Examples**

```
data("bodyfat", package = "TH.data")

### final model proposed by Garcia et al. (2005)
fmod <- lm(DEXfat ~ hipcirc + anthro3a + kneebreadth, data = bodyfat)
coef(fmod)
```

---

GBSG2

*German Breast Cancer Study Group 2*


---

**Description**

A data frame containing the observations from the GBSG2 study.

**Usage**

```
data("GBSG2")
```

**Format**

This data frame contains the observations of 686 women:

**horTh** hormonal therapy, a factor at two levels no and yes.

**age** of the patients in years.

**menostat** menopausal status, a factor at two levels pre (premenopausal) and post (postmenopausal).

**tsize** tumor size (in mm).

**tgrade** tumor grade, a ordered factor at levels I < II < III.

**pnodes** number of positive nodes.

**progrec** progesterone receptor (in fmol).

**estrec** estrogen receptor (in fmol).

**time** recurrence free survival time (in days).

**cens** censoring indicator (0- censored, 1- event).

## Source

W. Sauerbrei and P. Royston (1999). Building multivariable prognostic and diagnostic models: transformation of the predictors by using fractional polynomials. *Journal of the Royal Statistics Society Series A*, Volume **162**(1), 71–94.

## References

M. Schumacher, G. Basert, H. Bojar, K. Huebner, M. Olschewski, W. Sauerbrei, C. Schmoor, C. Beyerle, R.L.A. Neumann and H.F. Rauschecker for the German Breast Cancer Study Group (1994), Randomized  $2 \times 2$  trial evaluating hormonal treatment and the duration of chemotherapy in node-positive breast cancer patients. *Journal of Clinical Oncology*, **12**, 2086–2093.

## Examples

```
data(GBSG2)

thsum <- function(x) {
  ret <- c(median(x), quantile(x, 0.25), quantile(x,0.75))
  names(ret)[1] <- "Median"
  ret
}

t(apply(GBSG2[,c("age", "tsize", "pnodes",
               "progrec", "estrec")], 2, thsum))

table(GBSG2$menostat)
table(GBSG2$tgrade)
table(GBSG2$horTh)
```

---

geyser

*Old Faithful Geysler Data*

---

## Description

A version of the eruptions data from the ‘Old Faithful’ geyser in Yellowstone National Park, Wyoming. This version comes from Azzalini and Bowman (1990) and is of continuous measurement from August 1 to August 15, 1985.

Some nocturnal duration measurements have originally been described as ‘short’, ‘medium’ or ‘long’ and are given as interval censored observations in this version of the dataset.

## Usage

geyser

**Format**

A data frame with 299 observations on 2 variables.

duration	Surv	Eruption time in mins
waiting	numeric	Waiting time for this eruption

**Note**

Variable duration was converted to a Surv object for representing interval censored nocturnal observations.

**References**

Azzalini, A. and Bowman, A. W. (1990) A look at some data on the Old Faithful geyser. *Applied Statistics* **39**, 357–365.

Venables, W. N. and Ripley, B. D. (2002) *Modern Applied Statistics with S*. Fourth edition. Springer.

**See Also**

[faithful](#), [geyser](#).

---

GlaucomaM

*Glaucoma Database*

---

**Description**

The GlaucomaM data has 196 observations in two classes. 62 variables are derived from a confocal laser scanning image of the optic nerve head, describing its morphology. Observations are from normal and glaucomatous eyes, respectively.

**Usage**

```
data("GlaucomaM")
```

**Format**

This data frame contains the following predictors describing the morphology of the optic nerve head and a membership variable:

- ag** area global.
- at** area temporal.
- as** area superior.
- an** area nasal.
- ai** area inferior.
- eag** effective area global.

**eat** effective area temporal.  
**eas** effective area superior.  
**ean** effective area nasal.  
**eai** effective area inferior.  
**abrg** area below reference global.  
**abrt** area below reference temporal.  
**abrs** area below reference superior.  
**abrn** area below reference nasal.  
**abri** area below reference inferior.  
**hic** height in contour.  
**mhcg** mean height contour global.  
**mhct** mean height contour temporal.  
**mhcs** mean height contour superior.  
**mhcn** mean height contour nasal.  
**mhci** mean height contour inferior.  
**phcg** peak height contour.  
**phct** peak height contour temporal.  
**phcs** peak height contour superior.  
**phcn** peak height contour nasal.  
**phci** peak height contour inferior.  
**hvc** height variation contour.  
**vbsg** volume below surface global.  
**vbst** volume below surface temporal.  
**vbss** volume below surface superior.  
**vbsn** volume below surface nasal.  
**vbsi** volume below surface inferior.  
**vasg** volume above surface global.  
**vast** volume above surface temporal.  
**vass** volume above surface superior.  
**vasn** volume above surface nasal.  
**vasi** volume above surface inferior.  
**vbrg** volume below reference global.  
**vbrt** volume below reference temporal.  
**vbrs** volume below reference superior.  
**vbrn** volume below reference nasal.  
**vbri** volume below reference inferior.  
**varg** volume above reference global.

**vart** volume above reference temporal.  
**vars** volume above reference superior.  
**varn** volume above reference nasal.  
**vari** volume above reference inferior.  
**mdg** mean depth global.  
**mdt** mean depth temporal.  
**mds** mean depth superior.  
**mdn** mean depth nasal.  
**mdi** mean depth inferior.  
**tmg** third moment global.  
**tmt** third moment temporal.  
**tms** third moment superior.  
**tmn** third moment nasal.  
**tmi** third moment inferior.  
**mr** mean radius.  
**rnf** retinal nerve fiber thickness.  
**mdic** mean depth in contour.  
**emd** effective mean depth.  
**mv** mean variability.  
**Class** a factor with levels glaucoma and normal.

### Details

All variables are derived from a laser scanning image of the eye background taken by the Heidelberg Retina Tomograph. Most of the variables describe either the area or volume in certain parts of the papilla and are measured in four sectors (temporal, superior, nasal and inferior) as well as for the whole papilla (global). The global measurement is, roughly, the sum of the measurements taken in the four sector.

The observations in both groups are matched by age and sex to prevent any bias.

### Source

Torsten Hothorn and Berthold Lausen (2003), Double-Bagging: Combining classifiers by bootstrap aggregation. *Pattern Recognition*, **36**(6), 1303–1309.



---

 mammoexp

*Mammography Experience Study*


---

**Description**

Data from a questionnaire on the benefits of mammography.

**Usage**

data(mammoexp)

**Format**

A data frame with 412 observations on the following 6 variables.

**ME** Mammograph experience, an ordered factor with levels Never < Within a Year < Over a Year

**SYMPT** Agreement with the statement: ‘You do not need a mamogram unless you develop symptoms.’ A factor with levels Strongly Agree, Agree, Disagree and Strongly Disagree

**PB** Perceived benefit of mammography, the sum of five scaled responses, each on a four point scale. A low value is indicative of a woman with strong agreement with the benefits of mammography.

**HIST** Mother or Sister with a history of breast cancer; a factor with levels No and Yes.

**BSE** Answers to the question: ‘Has anyone taught you how to examine your own breasts?’ A factor with levels No and Yes.

**DECT** Answers to the question: ‘How likely is it that a mammogram could find a new case of breast cancer?’ An ordered factor with levels Not likely < Somewhat likely < Very likely.

**Source**

Hosmer and Lemeshow (2000). *Applied Logistic Regression*, 2nd edition. John Wiley & Sons Inc., New York. Section 8.1.2, page 264.

---

 mn6.9

*I.Q. and attitude towards science*


---

**Description**

Responses given by 2982 New Jersey high-school seniors on 4 questions concerning attitude towards science. Also recorded was whether students had a high or low I.Q.

**Usage**

data(mn6.9)

**Format**

A data frame with 2982 observations on the following 5 variables.

y1 Agree=1/disagree=0 to "The development of new ideas is the scientist's greatest source of satisfaction"

y2 Agree=1/disagree=0 to "Scientists and engineers should be eliminated form the military draft"

y3 Agree=1/disagree=0 to "The scientist will make his maximum contribution to society when he has freedom to work on problems that interest him"

y4 Agree=1/disagree=0 to "The monetary compensation of a Nobel Prize-winner in physics should be at least equal to that given to popular entertainers"

group I.Q. levels: 1=low, 2=high

**Source**

McCullagh, P. and Nelder, J.A. (1989, p. 239). *Generalized Linear Models*. Second Edition. Chapman & Hall/CRC.

copied from multmod package 1.0 (CRAN archive)

---

sphase

*S-phase Fraction of Tumor Cells*

---

**Description**

S-phase fraction of tumor cells in breast cancer patients.

**Usage**

```
data("sphase")
```

**Format**

This data frame contains the following columns:

**SPF** S-phase fraction

**RFS** recurrence free survival

**event** censoring indicator: FALSE means censored, TRUE is an event.

**Details**

The data have been used to address the question whether a simple cutpoint in S-phase fraction can be used to discriminate between patients with good and bad prognosis (for example in Hothorn & Lausen, 2003).

**Source**

J. Pfisterer, F. Kommoss, W. Sauerbrei, D. Menzel, M. Kiechle, E. Giese, M. Hilgarth & A. Pfeiderer (1995). DNA flow cytometry in node positive breast cancer: Prognostic value and correlation to morphological and clinical factors. *Analytical and Quantitative Cytology and Histology* **7**(6), 406–412.

**References**

Torsten Hothorn & Berthold Lausen (2003). On the Exact Distribution of Maximally Selected Rank Statistics. *Computational Statistics & Data Analysis* **43**(2), 121–137.

---

Westbc

*Breast Cancer Gene Expression*

---

**Description**

Gene expressions for 7129 genes in 49 breast cancer samples and the status of lymph node involvement.

**Usage**

```
data("Westbc")
```

**Format**

An list with two elements to be converted to class ExpressionSet (see package Biobase).

**Details**

A full description of the data can be found in West et al. (2001) and an application of boosted linear models is given by Buehlmann (2006).

**Source**

Mike West, Carrie Blanchette, Holly Dressman, Erich Huang, Seiichi Ishida, Rainer Spang, Harry Zuzan, John A. Olson Jr., Jeffrey R. Marks and Joseph R. Nevins (2001), Predicting the clinical status of human breast cancer by using gene expression profiles, *Proceedings of the National Academy of Sciences*, **98**, 11462-11467.

**References**

Peter Buehlmann (2006), Boosting for high-dimensional linear models. *The Annals of Statistics*, **34**(2), 559–583.

Peter Buehlmann and Torsten Hothorn (2007), Boosting algorithms: regularization, prediction and model fitting. *Statistical Science*, **22**(4), 477–505.

**Examples**

```
## Not run:
library("Biobase")
data("Westbc", package = "TH.data")
westbc <- new("ExpressionSet",
             phenoData = new("AnnotatedDataFrame", data = Westbc$pheno),
             assayData = assayDataNew(exprs = Westbc$assay))

## End(Not run)
```

---

wpbc

*Wisconsin Prognostic Breast Cancer Data*


---

**Description**

Each record represents follow-up data for one breast cancer case. These are consecutive patients seen by Dr. Wolberg since 1984, and include only those cases exhibiting invasive breast cancer and no evidence of distant metastases at the time of diagnosis.

**Usage**

```
data("wpbc")
```

**Format**

A data frame with 198 observations on the following 34 variables.

status a factor with levels N (nonrecur) and R (recur)  
time recurrence time (for status == "R") or disease-free time (for status == "N").  
mean\_radius radius (mean of distances from center to points on the perimeter) (mean).  
mean\_texture texture (standard deviation of gray-scale values) (mean).  
mean\_perimeter perimeter (mean).  
mean\_area area (mean).  
mean\_smoothness smoothness (local variation in radius lengths) (mean).  
mean\_compactness compactness (mean).  
mean\_concavity concavity (severity of concave portions of the contour) (mean).  
mean\_concavepoints concave points (number of concave portions of the contour) (mean).  
mean\_symmetry symmetry (mean).  
mean\_fractaldim fractal dimension (mean).  
SE\_radius radius (mean of distances from center to points on the perimeter) (SE).  
SE\_texture texture (standard deviation of gray-scale values) (SE).  
SE\_perimeter perimeter (SE).  
SE\_area area (SE).

SE\_smoothness smoothness (local variation in radius lengths) (SE).  
 SE\_compactness compactness (SE).  
 SE\_concavity concavity (severity of concave portions of the contour) (SE).  
 SE\_concavepoints concave points (number of concave portions of the contour) (SE).  
 SE\_symmetry symmetry (SE).  
 SE\_fractaldim fractal dimension (SE).  
 worst\_radius radius (mean of distances from center to points on the perimeter) (worst).  
 worst\_texture texture (standard deviation of gray-scale values) (worst).  
 worst\_perimeter perimeter (worst).  
 worst\_area area (worst).  
 worst\_smoothness smoothness (local variation in radius lengths) (worst).  
 worst\_compactness compactness (worst).  
 worst\_concavity concavity (severity of concave portions of the contour) (worst).  
 worst\_concavepoints concave points (number of concave portions of the contour) (worst).  
 worst\_symmetry symmetry (worst).  
 worst\_fractaldim fractal dimension (worst).  
 tsize diameter of the excised tumor in centimeters.  
 pnodes number of positive axillary lymph nodes observed at time of surgery.

## Details

The first 30 features are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. They describe characteristics of the cell nuclei present in the image.

There are two possible learning problems: predicting status or predicting the time to recur.

1) Predicting field 2, outcome: R = recurrent, N = non-recurrent - Dataset should first be filtered to reflect a particular endpoint; e.g., recurrences before 24 months = positive, non-recurrence beyond 24 months = negative. - 86.3 previous version of this data.

2) Predicting Time To Recur (field 3 in recurrent records) - Estimated mean error 13.9 months using Recurrence Surface Approximation.

The data are originally available from the UCI machine learning repository.

## Source

W. Nick Street, Olvi L. Mangasarian and William H. Wolberg (1995). An inductive learning approach to prognostic prediction. In A. Prieditis and S. Russell, editors, *Proceedings of the Twelfth International Conference on Machine Learning*, pages 522–530, San Francisco, Morgan Kaufmann.

Peter Buehlmann and Torsten Hothorn (2007), Boosting algorithms: regularization, prediction and model fitting. *Statistical Science*, **22**(4), 477–505.

**Examples**

```
data("wpbc", package = "TH.data")

### fit logistic regression model
coef(glm(status ~ ., data = wpbc[,colnames(wpbc) != "time"],
        family = binomial()))
```

# Index

## \* datasets

- birds, [2](#)
- bodyfat, [3](#)
- GBSG2, [4](#)
- geyser, [5](#)
- GlaucomaM, [6](#)
- mammoexp, [9](#)
- mn6.9, [9](#)
- sphase, [10](#)
- Westbc, [11](#)
- wpbc, [12](#)

- birds, [2](#)
- bodyfat, [3](#)

- faithful, [6](#)

- GBSG2, [4](#)
- geyser, [5, 6](#)
- GlaucomaM, [6](#)

- mammoexp, [9](#)
- mn6.9, [9](#)

- sphase, [10](#)

- Westbc, [11](#)
- wpbc, [12](#)