Package 'DATAstudio'

May 24, 2026

Version 1.2.4

Date 2026-05-24

Title The Research Data Warehouse of Miguel de Carvalho

Description Pulls together a collection of datasets from Miguel de Carvalho research articles and books. Including, for example:
- de Carvalho (2012) <doi:10.1016/j.jspi.2011.08.016>;
- de Carvalho et al (2012) <doi:10.1080/03610926.2012.709905>;
- de Carvalho et al (2012) <doi:10.1016/j.econlet.2011.09.007>);
- de Carvalho and Davison (2014) <doi:10.1080/01621459.2013.872651>;
- de Carvalho and Rua (2017) <doi:10.1016/j.ijforecast.2015.09.004>;
- de Carvalho et al (2023) <doi:10.1002/sta4.560>;
- de Carvalho et al (2022) <doi:10.1007/s13253-021-00469-9>;
- Palacios et al (2025) <doi:10.1214/24-BA1420>.

Author Miguel de Carvalho [aut, cre]

Depends R (>= 3.5)

Maintainer Miguel de Carvalho <Miguel.deCarvalho@ed.ac.uk>

License GPL (>= 3)

Repository CRAN

Suggests extremis, spearmanCI

Imports data.table, ggplot2, scales

LazyData true

URL https://webhomes.maths.ed.ac.uk/~mdecarv/

NeedsCompilation no

Date/Publication 2026-05-24 17:50:02 UTC

Contents

`DATAstudio-package` *The Research Data Warehouse of Miguel de Carvalho*

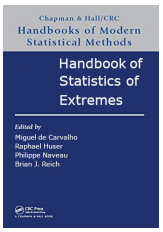
Description

DATAstudio is an add-on tool for R that pulls together a collection of datasets used in Miguel de Carvalho's research articles and books. For a complete list of datasets and documentation, type `help.start()` and follow the link to **DATAstudio** on the Package Index.

Data can be loaded using the commands `data` or `dataset`. The command `dataset` is used to retrieve datasets that are available only from GitHub (e.g., `dataset("lisbon")`).

If you use data from this package in publications, please cite the package and the references provided in the documentation. Type `citation("DATAstudio")`.

I gratefully acknowledge all contributors to the Handbook of Statistics of Extremes.



Funding

Generative AI Lab (Univ. of Edinburgh); Leverhulme Trust; Royal Society of Edinburgh.

Author(s)

Miguel de Carvalho; School of Mathematics, University of Edinburgh.

See Also

https://webhomes.maths.ed.ac.uk/~mdecarv/

AIG *AIG and Market Weekly Loss Returns (2000–2010)*

Description

Financial data on weekly *loss returns* (minus log-returns) for *American International Group* (AIG) equity and for a value-weighted US market index. The time period is from July 3rd, 2000, to June 30th, 2010.

Format

The file `AIG.RData` contains four numeric vectors:

AIGw Weekly loss returns (minus log-returns) on the AIG equity price, obtained by aggregating daily losses within each week.

xtab4 Same as `AIGw`.

Yw Weekly loss returns (minus log-returns) of a value-weighted market index over the same period, obtained by aggregating daily losses within each week.

ytab Same as `Yw`.

Details

The AIG daily returns were obtained from *Yahoo Finance* and converted to daily losses as `-log(1 + r_t)`, then aggregated weekly by summation. The market index daily returns were extracted from the file `Broker_Dealers_new.csv` used in Cai et al. (2015), and similarly converted to daily losses and aggregated weekly. Use `dataset("AIG")` to load these data from GitHub.

Source

AIG equity prices: Yahoo Finance. Market index returns: broker-dealer data from Cai et al. (2015), aggregating the NYSE and NASDAQ over the same period.

References

Cai, J.-J., Einmahl, J. H. J., de Haan, L. and Zhou, C. (2015). Estimation of the marginal expected shortfall: the mean when a related variable is extreme. *Journal of the Royal Statistical Society: Series B*, 77(2), 417–442.

Daouia, A. and Stupfler, G. (2026). Risk measures beyond quantiles. In: de Carvalho, M., Huser, R., Naveau, P., and Reich, B. J. (eds.), *Handbook on Statistics of Extremes*, Chapter 22, pp. 493–515. Chapman & Hall/CRC, Boca Raton, FL.

de Carvalho, M., Huser, R., Naveau, P., and Reich, B. J. (eds.) (2026). *Handbook on Statistics of Extremes*. Chapman & Hall/CRC, Boca Raton, FL.

alps *Swiss Alps Temperature Data*

Description

The alps data consist of daily winter temperature minima and maxima measured at 2m above ground surface at two sites in the Swiss Alps: Montana and Zermatt.

Usage

```
alps
```

Format

The alps data frame contains the following columns:

date Date of measurements.

min_montana, min_zermatt Daily minimum temperature in ºC on Montana and Zermatt.

max_montana, max_zermatt Daily maximum temperature in ºC on Montana and Zermatt.

Source

MeteoSwiss

References

Mhalla, L., de Carvalho, M., and Chavez-Demoulin, V. (2019) Regression type models for extremal dependence. *Scandinavian Journal of Statistics*, **46**, 1141-1167.

Examples

```
## visualizing the data
data(alps)
oldpar <- par(pty = 's', mfrow = c(1, 2))
plot(alps$min_montana, alps$min_zermatt, pch = 20,
     xlab = "Montana", ylab = "Zermatt", main = "Daily Minimum")
plot(alps$max_montana, alps$max_zermatt, pch = 20,
     xlab = "Montana", ylab = "Zermatt", main = "Daily Maximum")
par(oldpar)

oldpar <- par(pty = 's', mfrow = c(1, 2))
plot(alps$min_montana, alps$max_montana, pch = 20,
     xlab = "Minimum", ylab = "Maximum", main = "Montana")
abline(a = 0, b = 1, col = "red", lty = 2)
plot(alps$min_zermatt, alps$max_zermatt, pch = 20,
     xlab = "Minimum", ylab = "Maximum", main = "Zermatt")
abline(a = 0, b = 1, col = "red", lty = 2)
par(oldpar)
```

```
## Not run:
## to download the NAO daily index in Mhalla et al (2019) use
## the R package data.table to access NOAA via ftp
link <- paste0("ftp://ftp.cdc.noaa.gov/Public/gbates/teleconn/",
               "nao.reanalysis.t10trunc.1948-present.txt")
NAO.daily <- data.table::fread(link)
NAO.daily <- data.frame(NAO.daily)
colnames(NAO.daily) <- c("year", "month", "day", "NAO")

## End(Not run)
```

`beatenberg` *Beatenberg Forest Temperature Data (In Unit Fréchet Scale)*

Description

Preprocessed pairs of temperatures in unit Fréchet scale from Beatenberg forest, registered under forest cover and in the open field.

Usage

```
beatenberg
```

Format

The `beatenberg` data frame has 2839 rows and 2 columns: `x` (forest cover) and `y` (open field).

Details

Preprocessing was conducted as described in Ferrez et al (2011), and for applications of this dataset within the context of extreme value theory see de Carvalho *et al.* (2013), de Carvalho and Davison (2014) as well as Castro and de Carvalho (2017).

References

Castro, D. and de Carvalho, M. (2017) Spectral density regression for bivariate extremes. *Stochastic Environmental Research and Risk Assessment*, **31**, 1603-1613.

de Carvalho, M., Oumow, B., Segers, J., and Warchol, M. (2013) A Euclidean likelihood estimator for bivariate tail dependence. *Communications in Statistics—Theory and Methods*, **42**, 1176-1192.

de Carvalho, M. and Davison, A. C. (2014) Spectral density ratio models for multivariate extremes. *Journal of the American Statistical Association*, **109**, 764-776.

Ferrez, J., Davison, A. C., and Rebetez., M. (2011) Extreme temperature analysis under forest cover compared to an open field. *Agricultural and Forest Meteorology*, **151**, 992-1001.

Examples

```
## de Carvalho et al (2013, Fig. 5)
data(beatenberg)
attach(beatenberg)
plot(x, y, log = "xy", pch = 20, xlab = "Forest Cover", ylab = "Open Field")

## Not run:
## install package extremis if not installed
if (!require("extremis")) install.packages("extremis")

## de Carvalho et al (2013, Fig. 7)
data(beatenberg)
fit <- bev.kernel(beatenberg, tau = 0.98, nu = 163, raw = FALSE)
plot(fit)
rug(fit$w)

## End(Not run)
```

`bournemouth` *Air Pollution Measurements in Bournemouth*

Description

The `bournemouth` data frame contains daily air pollution measurements in Bournemouth UK from 2004 to 2023.

Format

A data frame with 1805 observations on 3 variables:

`Date` Calendar date in YYYYMMDD format.

`Nitric.oxide` Daily concentration of nitric oxide (NO).

`Nitrogen.dioxide` Daily concentration of nitrogen dioxide (NO_2).

Details

Use `dataset("bournemouth")` to load these data from GitHub.

References

de Carvalho, M., Huser, R., Naveau, P., and Reich, B. J. (2026). *Handbook of Statistics of Extremes*. Chapman & Hall/CRC, Boca Raton, FL.

Simpson, E. S., and Wadsworth, J. L. (2026). Conditional extremes modeling. In: *Handbook of Statistics of Extremes*, Chapter 10, pp. 199–220.

brainwave *Brainwave Data*

Description

The data contains the EEG power of two commonly-recognized types of EEG frequency bands: `Y1` for alpha and `Y2` for beta, for 30 participants and different covariates/stimulus. Column 3 to 8, represent the stimulus in the set: `x1` for "mathematics", `x2` for "relaxation", `x3` for "music", `x4` for "color", `x5` for "video", `x6` for "think and relax"). Column 9 represents the `id` of the participant, and column 10 contains the `time` in seconds.

Usage

```
brainwave
```

Format

The `brainwave` data frame has 7506 rows and 10 columns.

References

Palacios Ramirez, V., de Carvalho, M., and Gutierrez, L. (2025) Heavy-tailed NGG-mixture models. *Bayesian Analysis*, **20**, 1315-1343.

brexit *Brexit Poll Tracker*

Description

The data consist of 267 polls conducted before the June 23 2016 EU referendum, which took place in the UK.

Usage

```
brexit
```

Format

A dataframe with 272 observations on six variables.

`leave`**,** `stay`**,** `undecided` Percentage in favor of each option.

`date` Date on which the poll was conducted.

`pollster` Institution conducting the poll.

`size` Number of polled subjects.

Source

Financial Times (FT) Brexit poll tracker.

References

de Carvalho, M. and Martos, G. (2020). Brexit: Tracking and disentangling the sentiment towards leaving the EU. *International Journal of Forecasting*, **36**, 1128-1137.

Examples

```
## Leave-stay plot (de Carvalho and Martos, 2018; Fig. 1)
data(brexit)
attach(brexit)
oldpar <- par(pty = "s")
plot(leave[(leave > stay)], stay[(leave > stay)],
     xlim = c(22, 66), ylim = c(22, 66), pch = 16, col = "red",
     xlab = "Leave", ylab = "Stay")
points(leave[(stay > leave)], stay[(stay > leave)],
       pch = 16, col = "blue")
points(leave[(stay == leave)], stay[(stay == leave)],
       pch = 24)
abline(a = 0, b = 1, lwd = 3)
par(oldpar)
```

---

california *California Fire Perimeters*

Description

The `california` data frame has 16577 rows and 2 columns. The first column contains the date, the second column gives the quantity of acres consumed by the flames.

Format

This data frame contains the following columns:

`Date` A numeric vector of dates of wildfires.

`Acres` A numeric vector of thousands of acres consumed by the flames.

Details

Use `dataset("california")` to load these data from GitHub.

Source

California State Geoportal.

References

de Carvalho, M., Huser, R., Naveau, P., and Reich, B. J. (2026). *Handbook on Statistics of Extremes*. Chapman & Hall/CRC. Boca Raton, FL.

de Carvalho, M., Palacios, V., Henriques-Rodrigues, L., and Lee, M. W. (2026). Regression Models for Extreme Events. In: Handbook of Statistics of Extremes, Chapter 6, pp. 99–120.

`challenger` *Space Shuttle Challenger Data*

Description

Data on 23 flights of the space shuttle Challenger prior to the 1986 accident, wherein the shuttle blew up during takeoff.

Usage

```
challenger
```

Format

A dataframe with 23 observations on two variables, namely O-ring `temperature` (ºF) and `oring` state (`1` = failure; `0` = success).

References

de Carvalho, M. (2012) A Generalization of the Solis–Wets method. *Journal of Statistical Planning and Inference*, **142**, 633-644.

Examples

```
## Not run:
data(challenger)
ggplot(challenger, aes(x = as.factor(oring), y = temperature)) +
    geom_boxplot(fill = "steelblue", alpha = 0.3) +
    xlab("Failure") +
    ylab("Temperature (ºF)") +
    theme_minimal()

## End(Not run)
```

china_storm *China Weather Losses (EM-DAT)*

Description

The china_storm data frame has 166 rows and 47 columns. It contains storm-related disaster records for China (storms, tornadoes, tropical cyclones and hailstorms) from the EM-DAT (Emergency Events Database). The main loss variable is Total Damage, Adjusted, which can be converted to billions of USD by dividing by 10^6.

Format

This data frame contains the following columns (variable names follow EM-DAT):

DisNo. Character. EM-DAT disaster identifier.

Disaster Group Character. Disaster group (e.g., Natural).

Disaster Subgroup Character. Disaster subgroup (e.g., Meteorological).

Disaster Type Character. Disaster type (Storm).

Disaster Subtype Character. Disaster subtype (e.g., Tropical cyclone, Tornado).

Event Name Character. Named event, when available.

Country Character. Country name (China).

Region Character. Broad region (Asia).

Subregion Character. Subregion (Eastern Asia).

Location Character. Location description within China.

Start Year Integer. Event start year.

Start Month Integer. Event start month.

Start Day Integer. Event start day.

End Year Integer. Event end year.

End Month Integer. Event end month.

End Day Integer. Event end day.

Total Deaths Numeric. Total number of deaths.

No. Injured Numeric. Number of injured.

No. Affected Numeric. Number of affected.

No. Homeless Numeric. Number of homeless.

Total Affected Numeric. Total affected (injured + affected + homeless, as provided).

Magnitude Numeric. Event magnitude (when available).

Magnitude Scale Character. Magnitude scale (when available).

Latitude Numeric. Latitude (when available).

Longitude Numeric. Longitude (when available).

`Total Damage` Numeric. Total damage in thousand USD (unadjusted).

`Total Damage, Adjusted` Numeric. Total damage in thousand USD (CPI-adjusted).

`CPI` Numeric. Consumer Price Index used for adjustment.

`Admin Units` Character. Administrative-unit information (JSON-like string).

`Entry Date` Date (YYYY-MM-DD). Date the record was entered.

`Last Update` Date (YYYY-MM-DD). Date the record was last updated.

Details

Monetary variables are reported by EM-DAT in thousands of USD (000 USD), with both unadjusted and CPI-adjusted versions when available. Use `dataset("china_storm")` to load these data from GitHub. To view the data in a spreadsheet-style interface, type `View(china_storm)`.

Source

EM-DAT (Emergency Events Database), CRED / UCLouvain (2023), Brussels, Belgium.

References

Daouia, A. and Stupfler, G. (2026). Risk measures beyond quantiles. In: de Carvalho, M., Huser, R., Naveau, P., and Reich, B. J. (eds.), *Handbook on Statistics of Extremes*, Chapter 22, pp. 493–515. Chapman & Hall/CRC, Boca Raton, FL.

de Carvalho, M., Huser, R., Naveau, P., and Reich, B. J. (eds.) (2026). *Handbook on Statistics of Extremes*. Chapman & Hall/CRC, Boca Raton, FL.

EM-DAT (2023). *EM-DAT: The Emergency Events Database*. Centre for Research on the Epidemiology of Disasters (CRED), UCLouvain, Brussels, Belgium.

`claims` *Initial Claims of Unemployment*

Description

Weekly number (in thousands) of unemployment insurance claims in the US from 7 Jan 1967 until 28 Nov 2009.

Usage

```
claims
```

Format

A time series with 515 observations; the object is of class `tis` (time-indexed series).

Source

United States Department of Labor—Employment & Training Administration.

References

de Carvalho, M., Turkman, K. F. and Rua, A. (2013) Dynamic threshold modelling and the US business cycle. *Journal of the Royal Statistical Society, Ser.* C, **62**, 535-550.

See Also

https://webhomes.maths.ed.ac.uk/~mdecarv/decarvalho2013ash.html

Examples

```
## de Carvalho et al (2013; Fig 1)
data(claims)
plot(time(claims), claims, type = "l",
     xlab = "Time", ylab = "Initial Claims (in Thousands)")
```

---

cortical *Brain Shape Data*

Description

Axial brain slices gathered via magnetic resonance images (MRI) with 500 points on each outline, for 30 schizophrenia patients and 38 healthy controls.

Usage

```
cortical
```

Format

The `cortical` list has the following variables:

`age` Age, in years.

`group` Control patient (`Con`) or schizophrenia patient (`Scz`).

`sex` Male (`1`) or female (`2`).

`symm` Symmetry score obtained from raw 3D brain surface.

`x` **and** `y` Coordinates of slice from brain surface that intersects the AC (anterior commissure) and PC (posterior commissure).

`cortical\$r` 500 radii from angular polar coordinates.

Details

The data were gathered from a neuroscience study conducted at the University of British Columbia, Canada, and documented in Brignell *et al.* (2010) and Martos and de Carvalho (2018). Each brain was registered into the so-called Talairach space so that brains can be compared on the same three-dimensional referential coordinate space.

References

Brignell, C.J., Dryden, I.L., Gattone, S.A., Park, B., Leask, S., Browne, W.J., and Flynn, S. (2010) Surface shape analysis, with an application to brain surface asymmetry in schizophrenia. *Biostatistics*, **11**, 609-630.

Martos, G. and de Carvalho, M. (2018) Discrimination surfaces with application to region-specific brain asymmetry analysis. *Statistics in Medicine*, **37**, 1859-1873.

Examples

```
## Martos and de Carvalho (2018; Fig 1 a)
library(scales)
data(cortical)
m <- 500
n <- 68
plot(cortical$r[,1] * cos(2 * pi * 1:m / m),
     cortical$r[,1] * sin(2 * pi * 1:m / m) , type = "l",
     col = alpha("gray", 1 / n), xlab = "z", ylab = "x")
for(i in 2:n)
lines(cortical$r[, i] * cos(2 * pi * 1:m / m),
      cortical$r[, i] * sin(2 * pi * 1:m / m), type = "l",
      col = alpha("gray", i / n))
```

---

crypto *Crypto and Traditional Asset Returns*

Description

The `crypto` data frame contains daily returns for cryptocurrencies and traditional assets ranging from 2019 to 2023.

Format

A data frame with 977 observations on 9 variables:

`Date` Trading day.

`Bitcoin` Bitcoin return.

`Ether` Ether return.

`Gold_ETF` Gold ETF return.

`SP500` S&P 500 return.

`NASDAQ` NASDAQ return.

`Dow_Jones` Dow Jones return.

`SPUSBI` Bond index return (`SPUSBI`).

`Crypto_Index` Crypto index return.

Details

Use `dataset("crypto")` to load these data from GitHub.

References

Carl, D. L., Padoan, S. A., and Rizzelli, S. (2026). Measures of extremal dependence. In: *Handbook of Statistics of Extremes*, Chapter 8, pp. 153–174.

de Carvalho, M., Huser, R., Naveau, P., and Reich, B. J. (2026). *Handbook of Statistics of Extremes*. Chapman & Hall/CRC, Boca Raton, FL.

`cyclone_sst` *Tropical Cyclone and Sea Surface Temperature Data*

Description

The `cyclone_sst` dataset consists of point process data on tropical cyclone locations (latitude and longitude), together with information on storm intensity, status, and timing.

Usage

```
cyclone_sst
```

Format

The `cyclone_sst` data frame contains the following columns:

`latitude` Latitude (degrees).

`longitude` Longitude (degrees).

`category` Saffir–Simpson hurricane wind scale category (1–5, with 5 the most severe); 0 indicates tropical storms or tropical depressions.

`date` Date of the tropical cyclone event.

`status` Storm classification (hurricane, tropical storm, or tropical depression).

References

de Carvalho, M., Ferrer, C. and Vallejos, R. (2026). A Kolmogorov–Arnold neural model for cascading extremes. *Extremes*, to appear.

danube *Upper Danube Basin Data*

Description

River discharge data for tributaries of the Danube River.

Format

A named `list` with four components:

`data_clustered` A numeric matrix containing preprocessed discharge data for each gauging station.

`data_raw` A numeric matrix containing daily discharge observations for each gauging station.

`info` A data frame containing information on each gauging station and its catchment area.

`flow_edges` A two-column numeric matrix; each row gives the indices (in `info`) of a pair of gauging stations that are directly connected by a river segment.

Details

The matrix `data_clustered` was obtained by declustering the daily discharge data from the summer months between 1960 and 2010 contained in `data_raw`, yielding between seven and ten observations per year. Each row corresponds to one observation from the declustered time series; the *non-unique row names* indicate the year of observation. Each column corresponds to a gauging station, with column indices in `data_raw` and `data_clustered` matching row indices in `info`. See Asadi et al. (2015) for details on the preprocessing and declustering.

The `info` data frame contains the following variables for each gauging station or its associated catchment area:

`RivNames` Name of the river at the gauging station.

`Lat`, `Long` Geographic coordinates of the gauging station.

`Lat_Center`, `Long_Center` Coordinates of the center of the corresponding catchment area.

`Alt` Mean altitude of the catchment area.

`Area` Area of the catchment.

`Slope` Mean slope of the catchment.

`PlotCoordX`, `PlotCoordY` Coordinates used to arrange gauging stations when plotting a flow graph.

Use `dataset("danube")` to load these data from GitHub.

Source

Bavarian Environmental Agency, https://www.gkd.bayern.de and **graphicalExtremes**.

References

Asadi, P., Davison, A. C., Engelke, S., and Furrer, R. (2015). Extreme-value modeling of spatially dependent river discharges. *Journal of the American Statistical Association*, 110, 124–136.

de Carvalho, M., Huser, R., Naveau, P., and Reich, B. J. (2026). *Handbook of Statistics of Extremes*. Chapman & Hall/CRC, Boca Raton, FL.

Wan, P. and Janßen, A. (2026). Clustering Methods for Multivariate Extremes. In: Handbook of Statistics of Extremes, Chapter 12, pp. 243–262.

`dataset` *Load Dataset*

Description

This function loads a dataset that is not included in the package due to space constraints on CRAN, but is available online from GitHub. It works similarly to the R command `data` from the `utils` package, except that it downloads the dataset.

Usage

```
dataset(name)
```

Arguments

`name`	a string containing the dataset name.

Examples

```
## Download data
if (dataset("thefts")) {
  head(thefts)
  summary(thefts)
}
## for details on the dataset type
?thefts
```

`diabetes` *Diabetes Diagnosis Data*

Description

The `diabetes` data frame has 286 rows and 3 columns. The data were gathered from a population-based pilot survey of diabetes in Cairo, Egypt, in which postprandial blood glucose measurements were obtained from a fingerstick on 286 subjects. Based on the WHO (World Health Organization) criteria, 88 subjects were classified as diseased and 198 as healthy.

Usage

```
diabetes
```

Format

The `diabetes` data frame contains the following columns:

`marker` Postprandial blood glucose measurements (mg/dl) obtained from a fingerstick.

`status` Disease status, with `1` identifying subjects diagnosed with diabetes.

`age` Age in years.

References

Inácio de Carvalho, V., de Carvalho, M. and Branscum, A. (2017) Nonparametric Bayesian covariate-adjusted estimation of the Youden index. *Biometrics*, **73**, 1279-1288.

Inácio de Carvalho, V., Jara, A., Hanson, T. E. and de Carvalho, M. (2013) Bayesian nonparametric ROC regression modeling. *Bayesian Analysis*, **8**, 623-646.

Examples

```
data(diabetes)
plot(diabetes, pch = 20, main = "Diabetes Data")
```

`earthquake_tsunami` *Earthquake-Tsunami Data*

Description

The `earthquake_tsunami` dataset consists of point process data on earthquake locations (latitude and longitude) dating back to 2150 B.C., together with an indicator of whether the event was followed by a tsunami.

Usage

```
earthquake_tsunami
```

Format

The `earthquake_tsunami` data frame contains the following columns:

`tsunami` Indicator of tsunami occurrence (1 = yes, 0 = no).

`latitude` Epicentral latitude (º).

`longitude` Epicentral longitude (º).

`magnitude` Earthquake magnitude (Richter scale).

`focal` Focal depth of the earthquake (km).

References

de Carvalho, M., Ferrer, C. & Vallejos, R. (2026, to appear). A Kolmogorov–Arnold neural model for cascading extremes. *Extremes*.

`ecg200` *Electrocardiogram Data*

Description

The `ecg` data frame has 200 rows and 97 columns. The data is the result of monitoring electrical activity recorded during one heartbeat and it consists of 200 ECG signals sampled at 96 time instants, corresponding to 133 normal heartbeats and 67 myocardial infarction signals.

Usage

```
ecg200
```

Format

The `ecg200` data frame contains the following columns:

`status` : status of the patient, where `1` identifies subjects with myocardial infarction signals, and `0` identifies subjects with normal heartbeats.

`i1` **to** `i96` measurements at instants `i1` to `i96`; to my knowledge the exact unit of time is unknown and is not specified by Olszewski (2001), who gathered the data.

References

de Carvalho, M. and Martos, G. (2024). Uncovering sets of maximum dissimilarity on random process data. *Transactions on Machine Learning Research*, **5**, 1-31.

Olszewski, R. T. (2001). Generalized feature extraction for structural pattern recognition in time-series data. Carnegie Mellon University, PhD thesis.

Examples

```
## Not run:
## de Carvalho and Martos (2024, TMLR; Fig. 4)
if (!require("dplyr")) install.packages("dplyr")
if (!require("ggplot2")) install.packages("ggplot2")
if (!require("tidyr")) install.packages("tidyr")

packages <- c("dplyr", "ggplot2", "tidyr")
sapply(packages, require, character = TRUE)
longECG <- ecg200
    pivot_longer(cols = starts_with("i"), names_to = "instant",
                 values_to = "value")
    mutate(instant = as.integer(sub("i", "", instant)))
```

```
# create scatter plot of pooled data
ggplot(longECG, aes(x = instant, y = value, color = factor(status))) +
    geom_point(size = 1, alpha = 0.3) +
    labs(color = "Status") +
    scale_color_manual(values = c("0" = "red", "1" = "blue"),
                       labels = c("0" = "Non-diseased", "1" = "Diseased")) +
    xlab("Time") +
    ylab("ECG Signal") +
    theme_minimal()

## End(Not run)
```

`epilepsy` *Epilepsy EEG Data*

Description

Electroencephalogram (EEG) recordings from a patient experiencing a temporal lobe epileptic seizure.

Format

A numeric matrix with $50\,000$ rows and 19 columns, containing EEG recordings from 19 channels sampled at 100 Hz.

Details

The data contain EEG recordings from 19 channels of a female patient suffering from a temporal lobe epileptic seizure, monitored by a neurologist at the Epilepsy Center of the University of Michigan. The EEG signals were sampled at 100Hz (100 observations per second) over a duration of 500 seconds, yielding a total of $50\,000$ time points. The seizure onset occurs after 350 seconds (i.e., at time $35\,000$).

The data are organized as a matrix of dimension $50\,000 \times 19$, where columns correspond to EEG channels and rows correspond to recordings at times $1, \ldots, 50\,000$.

References

de Carvalho, M., Huser, R., Naveau, P., and Reich, B. J. (2026). *Handbook of Statistics of Extremes*. Chapman & Hall/CRC, Boca Raton, FL.

Redondo, P. V., Guerrero, M. B., Huser, R., and Ombao, H. (2026). Statistics of extremes for neuroscience. In: *Handbook of Statistics of Extremes*, Chapter 30, pp. 675–690.

`eurorain` *European Rainfall Monthly Maxima*

Description

The `eurorain` data frame contains 278850 observations of monthly maximum hourly rainfall (mm) and relevant covariates. Data are observed over a regular spatial grid which encompasses the British Isles, and parts of France, Belgium, and the Netherlands.

Format

This data frame contains the following columns:

`times` The year and month of each observation.

`Y` A 66 by 4225 matrix of monthly maximum hourly rainfall (mm) values.

`X` A 66 by 4225 by 17 array of relavant covariates.

`cov_names` Shorthand names for the covariates. Aligns with the last dimension of `X`.

`coords` A 4225 by 2 matrix of (longitude, latitude) coordinates.

Details

Response data `Y` are monthly maxima of hourly precipitation values (mm) for a regular spatial grid encompassing the British Isles, as well as parts of France, Belgium, and the Netherlands. These data were obtained from the ERA5-reanalysis on single levels. The grid-boxes were originally arranged on a regular

$$65 \times 65$$

latitude/longitude grid, with spatial resolution 0.25 degrees. The observation period encompasses only the summer months (June, July, August) and the years 2001-2022, inclusive. This leaves 66 observations of the monthly maximum hourly rainfall per grid-cell, which are stored in `Y`, a 66 by 4225 matrix, with the rows corresponding to observations and the columns corresponding to sampling locations. The variable `coords` is a 4552 by 2 matrix of (longitude, latitude) coordinates for the sampling locations. `times` is a vector of the year-month for each observations.

We have 17 covariates in `X` for each space-time observation of `Y`, and so `X` is a 66 by 4225 by 17 array. The covariates include the monthly mean and maximum of the following six dynamic meteorological variables: air temperature at a 2m altitude (t2m; K), mean sea level pressure (msl; Pa), surface level pressure (sp; Pa), total ozone in a column extending from the surface of the Earth to the atmosphere (tco3; kg/m^2), eastward and northward components of wind speed at a 10m altitude (u10 and v10; m/s^2). We also have five static covariates that do not change with time: anisotropy (anor; unitless), slope (slor; unitless), angle (isor; radians), and standard deviation (sdor; unitless) of the orography within a grid-cell, and a land-sea mask (lsm; unitless) which measures the proportion of land contained within a grid-box. The ordering of the covariates in the last dimension of `X` is determined by the vector `cov_names`‘.

References

Copernicus Climate Change Service, Climate Data Store (2023): ERA5 hourly data on single levels from 1940 to present. Copernicus Climate Change Service (C3S) Climate Data Store (CDS). DOI: 10.24381/cds.adbb2d47 (Accessed on 16-02-2026).

de Carvalho, M., Huser, R., Naveau, P., and Reich, B. J. (2026). *Handbook of Statistics of Extremes*. Chapman & Hall/CRC, Boca Raton, FL.

Richards, J. and Huser, R. (2026). Extreme Quantile Regression with Deep Learning. In: *Handbook of Statistics of Extremes*, Chapter 21, pp. 471–494.

`faang` *FAANG Data*

Description

Daily information on FAANG stocks; FAANG is an acronym for popular tech stocks, namely (Meta's) Facebook, Apple, Amazon, Netflix, and (Alphabet's) Google.

Format

The `faang` object is a list with five elements, each containing a matrix with columns corresponding to the opening, highest, lowest, and closing prices, as well as trading volume and adjusted closing price.

Details

The data consist of prices at close for these stocks over 2012-2024, and were gathered from Yahoo Finance. Use `dataset("faang")` to load these data from GitHub.

References

de Carvalho, M. and Palacios Ramirez, K. (2025) Semiparametric Bayesian modeling of nonstationary joint extremes: How do big tech's extreme losses behave? *Journal of the Royal Statistical Society, Ser.* C, **74**, 447-465.

Examples

```
## Not run:
dataset("faang")

## End(Not run)
```

fire *Danish Fire Insurance Claims Database*

Description

The Danish Fire Insurance Claims Database includes 2167 industrial fire losses gathered from the Copenhagen Reinsurance Company over the period 1980-1990.

Usage

```
fire
```

Format

A dataframe with 2167 observations on five variables, namely:

`Positions` Date.

`building` Loss to buildings.

`content` Loss to content.

`profits` Loss to profits.

`total` Total loss.

References

de Carvalho, M. and Marques, F. (2012) Jackknife Euclidean likelihood-based inference for Spearman's rho. *North American Actuarial Journal*, **16**, 487-492.

Examples

```
data(fire)
attach(fire)
plot(building, contents, pch = 20, xlim = c(0, 95), ylim = c(0, 133),
     xlab = "Loss of Building", ylab = "Loss of Contents",
     main = "Danish Fire Insurance Claims")

## Not run:
## Confidence intervals for Spearman rho; install the package
## spearmanCI, if not installed
if (!require("spearmanCI")) install.packages("spearmanCI")
spearmanCI(building, contents)

## End(Not run)
```

flights *Flight Delay Data*

Description

A dataset containing daily total delays of major US airlines. The raw data were obtained from the US Bureau of Transportation Statistics and subsequently preprocessed.

Format

A named `list` with three components:

`airports` A data frame containing information on US airports.

`delays` A numeric array containing daily aggregated delays at the airports in the dataset.

`flightCounts` A numeric array containing yearly numbers of flights between airports in the dataset.

Details

The component `flightCounts` is a three-dimensional array containing the number of flights between each pair of airports, aggregated on a yearly basis. Each entry gives the total number of flights between a departure airport (row) and a destination airport (column) in a given year (third dimension). This array does not contain any `NA`s; airports with no flights in a given year are represented by zeros.

The component `delays` is a three-dimensional array containing daily total positive delays (in minutes) of incoming and outgoing flights. Each column corresponds to an airport and each row to a day. The third dimension has length two, with `"arrivals"` containing delays of incoming flights and `"departures"` containing delays of outgoing flights. Zeros indicate that flights occurred but none were delayed; `NA`s indicate that no flights occurred on that day.

The component `airports` is a data frame containing information on US airports. Missing entries are indicated by `NA`.

`IATA` Three-letter IATA airport code.

`Name` Name of the airport.

`City` Primary city served by the airport.

`Country` Country or territory where the airport is located.

`ICAO` Four-letter ICAO airport code.

`Latitude` Latitude of the airport (decimal degrees).

`Longitude` Longitude of the airport (decimal degrees).

`Altitude` Altitude of the airport (feet).

`Timezone` Timezone offset from UTC (hours).

`DST` Daylight saving time used at the airport.

`Timezone2` Name of the timezone of the airport.

Use `dataset("flights")` to load these data from GitHub.

Source

Reproduced with permission from the **graphicalExtremes** package.

Raw delay data were obtained from the US Bureau of Transportation Statistics.

Airport metadata were obtained from: https://openflights.org/data.

References

de Carvalho, M., Huser, R., Naveau, P., and Reich, B. J. (2026). *Handbook of Statistics of Extremes*. Chapman & Hall/CRC, Boca Raton, FL.

Engelke, S., Hentschel, M., Lalancette, M., and Röttger, F. (2026). Graphical models for multivariate extremes. In: Handbook of Statistics of Extremes, Chapter 13, pp. 263–290.

Henzi, A., Engelke, S., and Reich, B. J. (2022). Graphical modeling for extremes. *Journal of the American Statistical Association*, **117**, 116–131.

Examples

```
require(DATAstudio)
if (dataset("flights")) {
  # Total number of flights in the dataset:
  totalFlightCounts <- apply(flights$flightCounts, c(1, 2), sum)

  # Number of flights in selected years:
  flightCounts_10_11 <- apply(flights$flightCounts[, , c("2010", "2011")],
                              c(1, 2), sum)
}
```

---

fort *Daily Precipitation Data from Fort Collins*

Description

The `fort` data frame contains daily precipitation measurements from Fort Collins (Colorado, US) over 1900–1999.

Format

A data frame with 36 524 daily observations and 5 variables:

`tobs` Day-of-year index (1–366).

`month` Month of the year (1–12).

`day` Day of the month.

`year` Calendar year.

`prec` Daily precipitation amount (inches).

Details

The variable `tobs` indexes the day within the year and ranges from 1 to 366, allowing for leap years. Use `dataset("fort")` to load these data from GitHub. The dataset is also part of the **extRemes** package.

References

de Carvalho, M., Huser, R., Naveau, P., and Reich, B. J. (2026). *Handbook of Statistics of Extremes*. Chapman & Hall/CRC, Boca Raton, FL.

Majumder, R., Shaby, B. A., and Reich, B. J. (2026). Bayesian methods for extreme value analysis. In: *Handbook of Statistics of Extremes*, Chapter 4, pp. 57–78.

`GDP` *GDP of the US Economy*

Description

US GDP (Gross Domestic Product) ranging from from 1950 (Q1) to 2009 (Q4).

Usage

```
GDP
```

Format

A time series with 268 observations on two variables. The object is of class `ts`.

Source

de Carvalho, M., Rodrigues, P. and Rua, A. (2012) Tracking the US business cycle with a singular spectrum analysis. *Economics Letters*, **114**, 32-35.

References

de Carvalho, M. and Rua, A. (2017) Real-time nowcasting the US output gap: Singular spectrum analysis at work. *International Journal of Forecasting*, **33**, 185-198.

See Also

https://webhomes.maths.ed.ac.uk/~mdecarv/decarvalho2012dsh.html

Examples

```
data(GDP)
plot(GDP, ylab = "Gross Domestic Product")

## Not run:
if (!require("ASSA")) install.packages("ASSA")
data(GDP)
fit <- bssa(log(GDP[, 1]))
plot(fit)
print(fit)

## End(Not run)
```

GDPIP *A Real-time Vintage of GDP and IP for the US Economy*

Description

US GDP (Gross Domestic Product) and IP (Industrial Production) ranging from from 1947 (Q1) to 2013 (Q4); the data correspond to a real-time vintage.

Usage

```
GDPIP
```

Format

A bivariate time series with 268 observations on two variables: `GDP` and `IP`. The object is of class `mts`.

Source

Federal Reserve Bank of Philadelphia.

References

de Carvalho, M. and Rua, A. (2017). Real-time nowcasting the US output gap: Singular spectrum analysis at work. *International Journal of Forecasting*, **33**, 185-198.

See Also

https://webhomes.maths.ed.ac.uk/~mdecarv/decarvalho2017sh.html

Examples

```
data(GDPIP)
plot(GDPIP)

## Plotting GDP against IP (de Carvalho and Rua, 2017; Fig. 4)
data(GDPIP)
oldpar <- par(mar = c(5, 4, 4, 5) + .1)
plot(GDPIP[, 1], type = "l",
     xlab = "Time", ylab = "Gross Domestic Product (GDP)",
     lwd = 3, col = "red", cex.lab = 1.4, cex.axis = 1.4)
par(new = TRUE)
plot(GDPIP[, 2], type = "l", xaxt = "n", yaxt = "n",
     xlab = "", ylab = "", lwd = 3, col = "blue", cex.axis = 1.4)
axis(4)
mtext("Industrial Production (IP)", side = 4, line = 3, cex = 1.4)
legend("topleft", col = c("red", "blue"),
       lty = 1, lwd = 3, legend = c("GDP", "IP"))
par(oldpar)

## Not run:
    ## Tracking the US Business Cycle (de Carvalho et al, 2017; Fig. 6)
    ## Install the package ASSA, if not installed
    if (!require("ASSA")) install.packages("ASSA")
    data(GDPIP)
    fit <- bmssa(log(GDPIP))
    plot(fit)
    print(fit)

## End(Not run)
```

---

heatwaves *Heatwaves Data*

Description

The heatwaves object is a list containing data sets and spatial objects related to heatwave analyses.

Format

A list with the following components:

era5_maxtemp_1950_2023 A data frame containing annual maxima of the N-day rolling mean of spatially averaged daily maximum temperature (with N varying by region), derived from the ERA5 reanalysis for 1950–2023.

era5_southwesternusa_midjuly2023 ERA5 spatial field for the South-Western US heatwave in mid-July 2023.

era5_lowlandschina_midjuly2023 ERA5 spatial field for the Lowlands China heatwave in mid-July 2023.

era5_southerneurope_midjuly2023 ERA5 spatial field for the Southern Europe heatwave in mid-July 2023.

phalodi_maxtemp A data frame containing annual maxima of daily maximum temperatures recorded in Bikaner and Jodhpur for the period 1944–2021, obtained from the GHCN dataset.

north_hemisphere_polygons A spatial object describing the geographical boundaries of selected regions in the Northern Hemisphere.

Details

The ERA5-based products rely on region definitions as described in Zachariah et al (2023). The N-day aggregation window varies by region and is chosen to reflect the temporal scale of persistent heat extremes. This dataset requires installation of the package **dash**. Use `dataset("heatwaves")` to load these data from GitHub.

References

Davison, A., and Miralles, O. (2026). Modeling univariate extremes—why and how. In: *Handbook of Statistics of Extremes*, Chapter 2, pp. 11–35.

de Carvalho, M., Huser, R., Naveau, P., and Reich, B. J. (2026). *Handbook of Statistics of Extremes*. Chapman & Hall/CRC, Boca Raton, FL.

Zachariah, M., Philip, S., Pinto, I., Vahlberg, M., Singh, R., Otto, F., Barnes, C., & Kimutai, J. (2023). Extreme heat in North America, Europe and China in July 2023 made much more likely by climate change. Imperial College London.

---

hongkong *Daily Maximum Temperature in Hong Kong*

Description

Daily Maximum Temperature Data from Hong Kong International Airport, Hong Kong, from January 1884 to October 2023.

Format

The hongkong data frame has 48517 observations and 2 columns:

`date` Year-month-day.

`value` Daily maximum temperature (in degrees Celsius).

Details

Data on daily maximum temperatures with no missing values, with a total of 48517 observations. Use `dataset("hongkong")` to load these data from GitHub.

References

Carcaiso, V., de Carvalho, M., Prosdocimi, I. and Antoniano-Villalobos, I. (2026). Bayesian mixture models for heterogeneous extremes. arXiv:2509.15359.

de Carvalho, M., Huser, R., Naveau, P., and Reich, B. J. (2026). *Handbook on Statistics of Extremes*. Chapman & Hall/CRC. Boca Raton, FL.

`hurricane` *Hurricane Tracking Data*

Description

Geographical coordinates, wind speed, and atmospheric pressure information for hurricanes from 1970 to 2011.

Format

The `hurricane` data frame has 43122 rows and 8 columns:

`Year` : Hurricane's year (ranging from 1971 to 2011).

`Number` : Year-specific hurricane identifier.

`Name` : Official name of the hurricane.

`ISO_Time` : Recorded observation time.

`Latitude` : Recorded latitude of the measurement.

`Longitude` : Recorded longitude of the measurement.

`Wind` : Wind speed (in knots)

`Pressure` : Atmospheric pressure (millibars).

Details

Use `dataset("hurricane")` to load these data from GitHub.

Source

National Hurricane Center and Brian A. Fannin.

References

de Carvalho, M., Huser, R., Naveau, P., and Reich, B. J. (2026). *Handbook on Statistics of Extremes*. Chapman & Hall/CRC. Boca Raton, FL.

kfrench *Fama–French Industry Portfolio Returns*

Description

The `kfrench` data frame contains daily returns for 30 Fama–French industry portfolios from Jan 1970 to December 2023.

Format

A data frame with 13 599 observations on 31 variables:

`time` Trading day in YYYYMMDD format.

Industry portfolios Thirty daily Fama–French industry portfolio return series, in percent.

Details

Use `dataset("kfrench")` to load these data from GitHub.

References

Cooley, D., Sabourin, A., and Wixson, T. (2026). Principal component analysis for multivariate extremes. In: *Handbook of Statistics of Extremes*, Chapter 11, pp. 221–242.

de Carvalho, M., Huser, R., Naveau, P., and Reich, B. J. (2026). *Handbook of Statistics of Extremes*. Chapman & Hall/CRC, Boca Raton, FL.

landslide *Earthquake-Induced Landslide Dataset*

Description

The `landslide` dataset contains data related with multiple-landslides following the May 2008 Wenchuan earthquake in Sichuan, China.

Format

The `landslide` dataset contains the following columns:

`presence` Binary indicator of landslide occurrence within the grid cell (1 = landslide present, 0 = no landslide).

`area_grid` Total area of the spatial grid cell.

`area_slide` Total area of landslide material mapped within the grid cell.

`count` Number of individual landslide events recorded within the grid cell.

`slope_avg` Mean slope angle within the grid cell.

`slope_stdd` Standard deviation of slope within the grid cell, representing local terrain variability.

`relief` Local terrain relief, defined as the elevation difference within the grid cell.

`TWI_avg` Mean topographic wetness index (TWI) within the grid cell, indicating potential soil moisture accumulation.

`TWI_stddev` Standard deviation of the topographic wetness index within the grid cell.

`VRM_avg` Mean vector ruggedness measure (VRM), quantifying surface roughness and terrain complexity.

`VRM_stddev` Standard deviation of the vector ruggedness measure within the grid cell.

`planCurv_a` Mean plan curvature, describing horizontal curvature of the terrain surface.

`planCurv_s` Standard deviation of plan curvature within the grid cell.

`pga_avg` Mean peak ground acceleration, representing average seismic shaking intensity.

`pga_stddev` Standard deviation of peak ground acceleration within the grid cell.

`distStream` Mean distance from the grid cell to the nearest stream or drainage network.

`distStre_s` Standard deviation of distance to streams within the grid cell.

`POINT_X` x-coordinate of the centroid of the grid cell (longitude or easting, depending on the coordinate system).

`POINT_Y` Y-coordinate of the centroid of the grid cell (latitude or northing, depending on the coordinate system).

`litho` Lithological classification indicating the dominant rock or soil type within the grid cell.

`profCurv_a` Mean profile curvature, describing vertical curvature of the terrain along the slope direction.

`profCurv_s` Standard deviation of profile curvature within the grid cell.

Use `dataset("landslide")` to load these data from GitHub.

Details

Use `dataset("landslide")` to load these data from GitHub.

References

de Carvalho, M., Huser, R., Naveau, P., and Reich, B. J. (2026). *Handbook on Statistics of Extremes*. Chapman & Hall/CRC. Boca Raton, FL.

Yadav, R., Lombardo, L., and Huser, R. (2026). Statistics of Extremes for Landslides and Earthquakes. In: *Handbook of Statistics of Extremes*, Chapter 27, pp. 611–632.

lisbon *Rainfall Data from Lisbon, Portugal*

Description

Daily rainfall data from Lisbon, Portugal, from December 1863 to June 2018.

Format

The `lisbon` data frame has 56503 observations and 2 columns:

`yearmonth` : year-month-day.

`prec` : total precipitation (mm).

Details

Prior to 1941, precipitation was measured for the 0-24 hour period; from 1941 onwards, precipitation was recorded from 9am to 9am the following day. Use `dataset("lisbon")` to load these data from GitHub.

Source

IPMA (Instituto Português do Mar e da Atmosfera).

References

Carcaiso, V., De Carvalho, M., Prosdocimi, I. and Antoniano-Villalobos, I. (2026). Bayesian mixture models for heterogeneous extremes. arXiv:2509.15359.

de Carvalho, M., and Carcaiso, V. (2026). Learning about extreme value distributions from data. In: *Handbook of Statistics of Extremes*, Chapter 3, pp. 37–56.

de Carvalho, M., Huser, R., Naveau, P., and Reich, B. J. (2026). *Handbook on Statistics of Extremes*. Chapman & Hall/CRC. Boca Raton, FL.

logreturns *Daily log-returns for international stock indices*

Description

Daily log-returns of six equity indices: NKX (Japan), KOSPI (Korea), HSI (Hong Kong), CAC (France), AEX (Netherlands), and NDQ (USA), covering Asia, Europe, and the USA. The time horizon ranges from Jan 1983 to September 2020.

Format

A named `list` with seven components:

`Date` Observation date.

`aex` AEX (Netherlands).

`cac` CAC (France).

`hsi` HSI (Hong Kong).

`kospi` KOSPI (Korea).

`ndq` NDQ (USA).

`nkx` NKX (Japan).

Each component contains the corresponding daily log-returns.

Details

Use `dataset("logreturns")` to load these data from GitHub.

Source

Stooq (https://stooq.com/db/h/).

References

Allouche, M., Girard, S., and Gobet, E. (2026). On the simulation of extreme events with neural networks. *Handbook of Statistics of Extremes*, pp. 447–468. Chapman & Hall/CRC, Boca Raton, FL.

de Carvalho, M., Huser, R., Naveau, P., and Reich, B. J. (2026). *Handbook of Statistics of Extremes*. Chapman & Hall/CRC, Boca Raton, FL.

`loss` *Loss and ALAE Insurance Data*

Description

Insurance indemnity payments and allocated loss adjustment expenses from an insurance company.

Format

The `loss` data frame contains the following variables:

`loss` Indemnity payment amount.

`alae` Allocated loss adjustment expense.

`limit` Policy limit.

`censored` Indicator of right-censoring due to the policy limit.

Details

The data were collected from Frees and Valdez (1998). Use `dataset("loss")` to load these data from GitHub.

References

de Carvalho, M., Huser, R., Naveau, P., and Reich, B. J. (2026). Handbook of Statistics of Extremes. Chapman & Hall/CRC, Boca Raton, FL.

Frees, E. and Valdez, E. (1998). Understanding relationships using copulas. North American Actuarial Journal, 2, 1–25.

Albrecher, H. and Beirlant, J. (2026). Statistics of Extremes for the Insurance Industry. In: Handbook of Statistics of Extremes, Chapter 29, pp. 655–673.

Belzile, L. R. and Nešlehová, J. G. (2026). Statistics of Extremes for Incomplete Data, with Application to Lifetime and Liability Claim Modeling. In: Handbook of Statistics of Extremes, Chapter 31, pp. 691–708.

`lse` *Selected Stocks from the London Stock Exchange*

Description

Prices at close from 26 selected stocks from the London stock exchange from 1989 to 2016.

Usage

```
lse
```

Format

The `lse` data frame has 6894 rows and 27 columns.

References

de Carvalho, M., Rubio, R., and Huser (2023). Similarity-based clustering for patterns of extreme values. Stat, **12**, e560.

lungcancer *Lung Cancer Diagnosis*

Description

The lungcancer data frame has 241 rows and 3 columns. The data were gathered gathered from a case-control study, conducted at the Mayo Clinic in Rochester (Minnesota), which included 140 controls and 101 lung cancer cases; only woman have been enrolled in the study.

Usage

```
lungcancer
```

Format

This data frame contains the following columns:

`marker` : square root of sEGFR levels (soluble isoform of the epidermal growth factor receptor).

`status` : disease status, with `1` identifying lung cancer cases and `0` identifying controls.

`pre` : premonopausal indicator, with `1` identifying premonopausal women.

`age` : age in years.

References

Inácio de Carvalho, V., Jara, A. and de Carvalho, M. (2015) Bayesian nonparametric approaches for ROC curve inference. In: *Nonparametric Bayesian Methods in Biostatistics and Bioinformatics*. Eds R. Mitra and P. Mueller. Cham: Springer.

madeira *Rainfall Data from Madeira*

Description

Rainfall data from Madeira, Portugal, from January 1973 to June 2018.

Usage

```
madeira
```

Format

The madeira data frame has 544 observations and 8 columns:

yearmonth Year and month.

prec Total monthly precipitation (0.01 inches).

amo Atlantic multi-decadal oscillation.

nino34 El Niño–Southern Oscillation (ENSO), expressed by the NINO3.4 index.

np North Pacific Index (NPI).

pdo Pacific Decadal Oscillation (PDO).

soi Southern Oscillation Index (SOI).

nao North Atlantic Oscillation (NAO).

Details

After eliminating the dry events (i.e., zero precipitation) and the missing precipitation data (two observations) one is left with a total of 532 observations, and that is the version of the data analyzed in de Carvalho et al (2022, 2026).

Source

National Oceanic and Atmospheric Administration.

References

de Carvalho, M., Pereira, S., Pereira, S., and de Zea Bermudez, P. (2022). An extreme value Bayesian lasso for the conditional left and right tails. *Journal of Agricultural, Biological and Environmental Statistics*, **27**, 222–239.

de Carvalho, M., Huser, R., Naveau, P., and Reich, B. J. (2026). *Handbook on Statistics of Extremes*. Chapman & Hall/CRC. Boca Raton, FL.

de Carvalho, M., Palacios, V., Henriques-Rodrigues, L., and Lee, M. W. (2026). Regression models for extreme events. In: *Handbook of Statistics of Extremes*, Chapter 6, pp. 99–120.

---

marketsUS *NASDAQ and NYSE Indices*

Description

Daily quotations at close of the NASDAQ and NYSE stock market indices from February 1971 till November 2021.

Usage

```
marketsUS
```

Format

The marketsUS data frame has 12562 rows and 3 columns: date and quotation at close of the nasdaq and nyse indices.

References

de Carvalho, M., Huser, R., Naveau, P., and Reich, B. J. (2026). *Handbook on Statistics of Extremes*. Chapman & Hall/CRC. Boca Raton, FL.

de Carvalho, M., Kumukova, A., and dos Reis, G. (2022) Regression-type analysis for multivariate extreme values. Extremes, **25**, 595-622.

Examples

```
## Not run:
## de Carvalho et al (2022; Fig 5.1)
data(marketsUS)
packages <- c("ggplot2", "scales")
sapply(packages, require, character.only = TRUE)
ggplot(data = marketsUS, aes(x = date, y = value, color = Indices)) +
    geom_line(aes(y = nasdaq, col = "NASDAQ"), alpha = 0.5,
              position = position_dodge(0.8), size = 1.1) +
    geom_line(aes(y = nyse, col = "NYSE"), alpha = 0.5,
              position = position_dodge(0.8), size = 1.1) +
    scale_y_continuous(breaks = seq(2000, 14000, by = 2000)) +
    scale_x_date(labels = date_format("%Y"),
                 breaks = as.Date(c("1971-01-01", "1978-01-01",
                                    "1985-01-01", "1992-01-01",
                                    "1999-01-01", "2006-01-01",
                                    "2013-01-01", "2020-01-01"))) +
    scale_color_manual(values = c("red", "blue")) +
    labs(y = "Value (in USD)", x = "Time (in Years)") +
    theme_minimal()

## End(Not run)
```

---

maxtemps *Maximum Temperatures in the Netherlands*

Description

Daily maximum temperatures measured at 18 inland stations in the Netherlands from 1990 to 2019, together with derived summaries, spatial features, and model-related objects.

Format

The dataset loads the following objects:

all.pairs Matrix of all possible station pairs for the 18 stations; each row represents one pair.

`area.maxima` Simulated area maxima arising from 30000 simulations from the fitted max-stable process; see details.

coast_sp Coastline polygon imported from **rnaturalearth** for advanced plots.

fit Fitted max-stable process; output of fitmaxstab; see details.

inland.grid 4712 inland locations with approximate grid distance 2.5 km alongside geographical information and classification into four regions; each location is at least 15 km away from the coastline; altitude information stems from the ASTER Global Digital Elevation Model V2 and was accessed via **geonames**.

lakes_sp Physical boundary of lakes imported from **rnaturalearth** for advanced plots.

maxima14days Data frame of summer 14 day-block maxima of daily maximum temperatures measured in 0.1 degree Celsius for 18 stations from 1990 to 2019 alongside geographical information about the respective stations; each summer consists of six blocks.

NL_sp Country boundary for the Netherlands imported from **rnaturalearth** for advanced plots.

stations Data frame of 18 stations in the Netherland alongside geographical information.

summer.temperature Same as data frame temperature, but restricted to 84 summer days (six blocks of 14 days).

temperature Data frame of original KNMI daily maximum temperatures measured in 0.1 degree Celsius for 18 stations from 1990 to 2019.

values.start Starting values for parameters for the optimization in fitmaxstab; see details.

Details

Use `dataset("maxtemps")` to load these data from GitHub.

The data contains the object fit describing the fitted max-stable process from Strokorb and Oesting (2026). Complementing the route taken in Oesting and Strokorb (2022), where marginals and dependence structure were estimated in two steps using the GEV independence likelihood and the M-estimator approach, the present fit arose from a one-step approach estimating marginal and dependence parameters jointly using the pairwise composite likelihood approach implemented in **SpatialExtremes**.

Thus, the object fit is an output of the function fitmaxstab, where the values from values.start were used as starting values for the parameters, see Example.R in https://github.com/strokorb/max-stable-spatial-inference.

The object area.maxima constains 30000 simulations of areal maxima arising from the fitted max-stable process in the object fit, corresponding to 5000 summers of data. The three areas S1, S2 and S3, over which maxima have been taken in space, consist of those inland grid points that are labelled as such in inland.grid in the variable region.

See Strokorb and Oesting (2026) for further details.

Source

KNMI daily climate data: https://www.knmiprojects.nl/projects/globe.

U.S./Japan ASTER Science Team. ASTER Global Digital Elevation Model, V2.doi:10.5067/ASTER/ASTGTM.002.

References

de Carvalho, M., Huser, R., Naveau, P., and Reich, B. J. (2026). *Handbook of Statistics of Extremes*. Chapman & Hall/CRC, Boca Raton, FL.

Massicotte, P. and South, A. (2023). *rnaturalearth: World Map Data from Natural Earth*. R package version 1.0.1.

Oesting, M., and Strokorb, K. (2022). A comparative tour through the simulation algorithms for max-stable processes. *Statistal Science*, **37**(1), pp. 42–63.

Ribatet, M. (2022). *SpatialExtremes: Modelling Spatial Extremes*. R package version 2.1-0.

Rowlingson, B. (2019). *geonames: Interface to the "Geonames" Spatial Query Web Service*. R package version 0.999.

Strokorb, K., and Oesting, M. (2026). Max-stable processes for spatial extremes. In: *Handbook of Statistics of Extremes*, Chapter 15, pp. 321–348.

`merval` *MERVAL Stock Market Data*

Description

Raw interval data series corresponding to weekly minimum and maximum values of the MERVAL index (Argentina stock market) ranging from January 1 2016 to September 30 2020 (along with prices at open and prices at close).

Usage

```
merval
```

Format

A dataframe with 353 observations and 5 columns: `dates`, `low`, `high`, `open`, and `close`.

Source

Yahoo Finance.

References

de Carvalho, M. and Martos, G. (2022). Modeling interval trendlines: Symbolic singular spectrum analysis for interval time series. *Journal of Forecasting*, **41**, 167-180.

Examples

```
data(merval)
attach(merval)
head(merval, 3)
oldpar <- par(pty = 's')
plot(low, high, pch = 20)
abline(a = 0, b = 1, lty = 2, col = "gray")
par(oldpar)
```

`metsynd` *Metabolic Syndrome Data*

Description

The `metsynd` data includes Gamma-Glutamyl Transferase (GGT) levels and curves of arterial oxygen saturation, for samples of women suffering from metabolic syndrome and women without metabolic syndrome; the data were gathered from a population-based survey conducted in Galicia (NW Spain), and it includes 35 women suffering from metabolic syndrome and 80 women without metabolic syndrome.

Usage

```
metsynd
```

Format

The data consist of a list with the following elements:

- `y0` GGT levels for women without metabolic syndrome.
- `y1` GGT levels for women suffering from metabolic syndrome.
- `X0` Curves of arterial oxygen saturation (%) for women without metabolic syndrome (`X0\$data`, `X0\$time`).
- `X1` Curves of arterial oxygen saturation (%) for women suffering from metabolic syndrome (`X1\$data`, `X1\$time`).

Details

The curves of arterial oxygen saturation are included in the matrices `X0$data` and `X1$data`, with each row representing a patient, and with columns representing ordered measurements over time. Here `X0$time` and `X1$time` represents the time (in hours) at which measurements were made, i.e., every 20 seconds during three hours of sleep. Further details on these data can be found in the references below.

References

Inácio de Carvalho, V., de Carvalho, M., Alonzo, T. A., González-Manteiga, W. (2016) Functional covariate-adjusted partial area under the specificity-ROC curve regression with an application to metabolic syndrome case study. *Annals of Applied Statistics*, **10**, 1472-1495

Examples

```
data(metsynd)
library(scales)
attach(metsynd)

## Inacio de Carvalho et al (2016; Fig 1)
```

```
oldpar <- par(mfrow = c(1,2))
n0 <- length(y0)
n1 <- length(y1)
t <- X1$time
plot(t, X1$data[1, ], type = "l", lwd = 3, ylim = c(70, 100),
     xlab = "Time (in hours)", ylab = "Arterial oxygen saturation (%)",
     main = "Metabolic syndrome")
for (i in 2:n1)
  lines(t, X1$data[i, ], type = "l", lwd = 3, col = alpha("black", i / n1))
plot(t, X0$data[1, ], type = "l", lwd = 3, col = "gray", ylim = c(70, 100),
     xlab = "Time (in hours)", ylab = "Arterial oxygen saturation (%)",
     main = "No metabolic syndrome")
for (i in 1:n0)
  lines(t, X0$data[i, ], type = "l", lwd = 3, col = alpha("gray", i / n0))
par(oldpar)
```

`netherlands` *Summer Maximum Temperatures in the Netherlands*

Description

Daily summer maximum temperatures (in degrees Celsius) observed from 1995 onward at 68 locations in the Netherlands.

Format

The data consist of the following separate objects:

`locations` A numeric matrix of longitude and latitude coordinates for each site.

`max.temps.summer` A numeric matrix of daily summer maximum temperatures, with rows corresponding to locations and columns to observation days.

`day` An integer vector giving the day of the month.

`month` A character vector giving the month (June–August).

`year` An integer vector giving the year of observation.

Details

Use `dataset("netherlands")` to load these data from GitHub.

References

de Carvalho, M., Huser, R., Naveau, P., and Reich, B. J. (2026). *Handbook of Statistics of Extremes*. Chapman & Hall/CRC, Boca Raton, FL.

Simpson, E. S., and Wadsworth, J. L. (2026). Conditional extremes modeling. In: *Handbook of Statistics of Extremes*, Chapter 10, pp. 199–220.

pandemics *Major Disease Outbreaks Throughout History*

Description

The dataset contais information on major disease outbreaks and their estimated mortality.

Format

The dataset contains the following components:

`original_pandemic_data` A data frame with 72 observations on 9 variables reproducing the original table reported in Cirillo and Taleb (2020), containing historical information on major pandemics.

`pandemic_deaths_year` A data frame with 63 observations on 9 variables, providing uniformly annualised estimates of total deaths attributable to pandemics for each year, together with the corresponding proportion of deaths relative to the global population in that year.

Details

Use `dataset("pandemics")` to load these data from GitHub.

References

Cirillo, P., & Taleb, N. N. (2020). Tail risk of contagious diseases. *Nature Physics*, **16**, 606-–613.

Davison, A., and Miralles, O. (2026). Modeling univariate extremes—why and how. In: *Handbook of Statistics of Extremes*, Chapter 2, pp. 11–35.

de Carvalho, M., Huser, R., Naveau, P., and Reich, B. J. (2026). *Handbook of Statistics of Extremes*. Chapman & Hall/CRC, Boca Raton, FL.

passengers *International Airline Traffic Data*

Description

Monthly number of passengers (in thousands) in a group of several international airline companies from January 1949-December 1960.

Usage

```
passengers
```

Format

A time series with 144 observations; the object is of class `ts`.

References

Brown, R.G. (1963) *Smoothing, Forecasting and Prediction of Discrete Time Series*. New Jersey: Prentice-Hall.

Rodrigues, P. C. and de Carvalho, M. (2013) Spectral modeling of time series with missing data. *Applied Mathematical Modelling*, **37**, 4676-4684.

---

pnw *Daily Maximum Temperature Extremes in the Pacific Northwest (PNW)*

Description

The pnw dataset contains summer (JJA) temperature-extremes information for 441 stations in the Pacific Northwest, covering 1950–2021. It includes 10-day block maxima, fitted generalized Pareto parameters at a 95% threshold, probability integral transform (PIT) copula values, station metadata, and location-wise GEV fits to seasonal maxima.

Format

The dataset is a list containing the following elements:

`PNW_JJA_10day` A numeric matrix of dimension 441×459 giving JJA 10-day block maxima (459 10-day maxima over 1950–2021) at 441 locations.

`params_GPD` A numeric matrix of dimension 441×2 containing location-wise generalized Pareto distribution (GPD) parameter estimates fitted above the location-specific 95th percentile threshold.

`U` A numeric matrix of dimension 441×459 containing PIT-based copula values derived from `PNW_JJA_10day` using `params_GPD`.

stationDF_PNW A data frame with 441 rows and 5 columns giving station metadata: GHCN station ID, longitude, latitude, elevation, and state (WA, OR, CA, or ID).

`Loc` A numeric matrix of dimension 441×71 of location-parameter estimates from location-wise GEV fits to 71 seasonal maxima (1950–2021).

`Scale` A numeric matrix of dimension 441×71 of scale-parameter estimates from location-wise GEV fits to 71 seasonal maxima (1950–2021).

`Shape` A numeric matrix of dimension 441×71 of shape-parameter estimates from location-wise GEV fits to 71 seasonal maxima (1950–2021).

Details

Use `dataset("pnw")` to load these data from GitHub.

Source

Global Historical Climatology Network-Daily (GHCN-D) database

References

de Carvalho, M., Huser, R., Naveau, P., and Reich, B. J. (2026). *Handbook on Statistics of Extremes*. Chapman & Hall/CRC. Boca Raton, FL.

Zhang, L., Rohrbeck, C., and Opitz, T. (2026). Subasymptotic models for spatial extremes. In *Handbook on Statistics of Extremes*, Chapter 17, pp. 377–400. Chapman & Hall/CRC, Boca Raton, FL.

Menne, M. J., Durre, I., Vose, R. S., Gleason, B. E., & Houston, T. G. (2012). An overview of the global historical climatology network-daily database. *Journal of Atmospheric and Oceanic Technology*, **29**, 897-910.

---

psa *Prostate Cancer Diagnosis Data*

Description

Longitudinal measurements of two Prostate Specific Antigen (PSA)-based biomarkers for 71 prostate cancer cases and 70 controls.

Usage

```
psa
```

Format

The `psa` data frame has 683 rows and 6 columns:

`id` patient id.

`marker1` total PSA.

`marker2` ratio of free total PSA.

`status` disease status of each subject, with `1` identifying subjects diagnosed with prostate cancer.

`age` age in years.

`t` time prior to diagnosis.

Details

The data were gathered from the Beta-Carotone and Retinol Efficacy Trial (CARET)—a lung cancer prevention trial, conducted at the Fred Hutchison Cancer Research Center. Further details on this study can be found in de Carvalho *et al.* (2020).

References

de Carvalho, M., Barney, B. and Page, G. L. (2020) Affinity-based measures of biomarker performance evaluation. *Statistical Methods in Medical Research*, **20**, 837-853.

rain_germany *Daily Precipitation in Germany*

Description

Daily precipitation recorded at 199 meteorological stations in Germany from 1923 to 2023.

Format

The dataset includes the following objects:

alldata A numeric matrix with 72051 observations on 199 variables containing daily precipitation measurements (in mm); columns correspond to stations.

metadata A data frame containing station identifiers, coordinates (longitude, latitude), elevation, station name, federal state, data availability period, and additional metadata.

Details

Water-level data are derived from measurements recorded several times per day and averaged to daily values. Use `dataset("seine")` to load these data from GitHub.

Source

German Weather Service (Deutscher Wetterdienst, DWD): https://opendata.dwd.de/climate_environment/CDC/observations_germany/climate/daily/more_precip/historical/.

References

Chavez-Demoulin, V., and Mhalla, L. (2026). Causality and extremes. In: *Handbook of Statistics of Extremes*, Chapter 19, pp. 425–446.

de Carvalho, M., Huser, R., Naveau, P., and Reich, B. J. (2026). *Handbook of Statistics of Extremes*. Chapman & Hall/CRC, Boca Raton, FL.

santiago *Santiago Temperature Data*

Description

The data consist of average daily air temperatures, measured in degrees Fahrenheit and rounded to the nearest integer, recorded in Santiago (Chile) from April 1990 to March 2017.

Usage

```
santiago
```

Format

A dataframe with 10126 observations on one variable.

Source

NOAA's National Centers for Environmental Information (NCEI).

References

Galasso, B., Zemel, Y., and de Carvalho, M. (2022). Bayesian semiparametric modelling of phase-varying point processes. *Electronic Journal of Statistics*, **16**, 2518-2549.

`seine` *Seine River Water Levels*

Description

Daily average water levels (in cm) at five locations on the Seine river (Paris, Meaux, Melun, Nemours, Sens). The series span 1 October 2005 to 8 April 2019 and comprise 3408 daily observations per station.

Format

The dataset includes the following objects:

`data_seine` A data frame with 3408 observations on 6 variables: `Date` and water levels at Paris, Meaux, Melun, Nemours, and Sens.

`hs_dat` A numeric matrix with 3408 rows and 100 columns (derived values at spatial grid locations).

`Locations` A data frame giving the `Longitude`, `Latitude`, and grid `Node` identifier for each spatial location.

Details

Water-level data are derived from measurements recorded several times per day and averaged to daily values. Use `dataset("seine")` to load these data from GitHub.

References

Asenova, S. Mazo, G. and Segers, J. (2021). Inference on extremal dependence in the domain of attraction of a structured Hüsler–Reiss distribution motivated by a Markov tree with latent variables. *Extremes*, **24**, 461–500.

Chavez-Demoulin, V., and Mhalla, L. (2026). Causality and extremes. In: *Handbook of Statistics of Extremes*, Chapter 19, pp. 425–446.

de Carvalho, M., Huser, R., Naveau, P., and Reich, B. J. (2026). *Handbook of Statistics of Extremes*. Chapman & Hall/CRC, Boca Raton, FL.

sp500 *Standard & Poor's 500*

Description

Daily S&P 500 index at close from 1988 till 2007.

Usage

```
sp500
```

Format

The `sp500` data frame has 5043 rows and 2 columns: `date` and price at `close`.

References

de Carvalho, M. (2016) Statistics of extremes: Challenges and opportunities. In: *Handbook of EVT and its Applications to Finance and Insurance*. Eds F. Longin. Hoboken: Wiley.

sp500a *Standard & Poor's 500 and Sector Indices*

Description

Daily S&P 500 index at close from 2002 till 2024 along with sector indices.

Format

A data frame with 5552 observations on 13 variables:

`Date` Trading day.

`S.P.Market` S&P 500 index level.

`S.P.Communication.Services` Communication Services sector index.

`S.P.Technology` Information Technology sector index.

`S.P.Industrial` Industrials sector index.

`S.P.Materials` Materials sector index.

`S.P.Consumer.Discretionary` Consumer Discretionary sector index.

`S.P.Financial` Financials sector index.

`S.P.Health.Care` Health Care sector index.

`S.P.Consumer.Staples` Consumer Staples sector index.

`S.P.Utilities` Utilities sector index.

`S.P.Real.Estate` Real Estate sector index.

`S.P.Energy` Energy sector index.

Details

Use `dataset("sp500a")` to load these data from GitHub.

References

Davison, A., and Miralles, O. (2026). Modeling univariate extremes—why and how. In: *Handbook of Statistics of Extremes*, Chapter 2, pp. 11–35.

de Carvalho, M., Huser, R., Naveau, P., and Reich, B. J. (2026). *Handbook of Statistics of Extremes*. Chapman & Hall/CRC, Boca Raton, FL.

`streamflow` *Annual Streamflow Maxima at HCDN Stations*

Description

The `streamflow` dataset contains annual streamflow maxima (1950–2021) at 702 HCDN stations in continental US, together with station attributes and hydrologic region labels.

Format

Data consists of the the following components:

`s` A data frame with 702 rows and columns `LONG_GAGE` and `LAT_GAGE` giving station longitude and latitude.

`Y` A matrix of dimension 702×72 of annual streamflow maxima (cubic m/s), with rows corresponding to stations and columns to years 1950–2021.

`drain` Drainage area for each station.

`HUC02` A factor identifying the hydrologic unit (25 regions) for each station.

`ID` Station identifier.

`nsites` Number of sites (i.e. 702).

`nyears` Number of years (i.e. 72).

`year` A numeric vector of length 72 giving the years 1950–2021.

Details

Use `dataset("streamflow")` to load these data from GitHub.

References

de Carvalho, M., Huser, R., Naveau, P., and Reich, B. J. (2026). *Handbook of Statistics of Extremes*. Chapman & Hall/CRC, Boca Raton, FL.

Majumder, R., Shaby, B. A., and Reich, B. J. (2026). Bayesian methods for extreme value analysis. In: *Handbook of Statistics of Extremes*, Chapter 4, pp. 57–78.

sydney *Monthly Sea Levels for Fort Denison*

Description

The sydney data frame contains monthly sea level measurements for Fort Denison (Sydney) from 1914 to 2023.

Format

This data frame contains has 1317 rows and 8 columns:

Mth Month of observation (1–12).

Year Year of observation.

Gaps Number of missing observations.

Good Number of valid observations.

Minimum Minimum sea level (m).

Maximum Maximum sea level (m).

Mean Mean sea level (m).

St.Devn Standard deviation of sea level (m).

Details

Use `dataset("sydney")` to load these data from GitHub.

Source

Australina Government, Bureau of Meteorology.

References

de Carvalho, M., and Carcaiso, V. (2026). Learning about extreme value distributions from data. In: *Handbook of Statistics of Extremes*, Chapter 3, pp. 37–56.

de Carvalho, M., Huser, R., Naveau, P., and Reich, B. J. (2026). *Handbook on Statistics of Extremes*. Chapman & Hall/CRC. Boca Raton, FL.

thefts *Thefts in Buenos Aires*

Description

Use `dataset("thefts")` to load these data from GitHub. The data consist of locations (latitude and longitude) of thefts in Buenos Aires from September 2019 to December 2020. For further details see de Carvalho and Martos (2024).

References

de Carvalho, M. and Martos, G. (2024). Uncovering sets of maximum dissimilarity on random process data. *Transactions on Machine Learning Research*, **5**, 1-31.

Examples

```
if (dataset("thefts")) {
  summary(thefts)
  head(thefts)
}
```

tmt *Trail Making Test*

Description

Completion times in seconds for TMT (Trail Making Test), part A, for 245 patients with Parkinson's disease, along with corresponding diagnostic on cognitive impairment.

Usage

```
tmt
```

Format

The `tmt` data frame has 245 rows and 2 columns:

`marker` completion times (in seconds)

`status` disease status of each subject, with `1`, `2`, and `3` respectively denoting patients diagnosed as unimpaired, mild cognitive impairment, and dementia.

References

Inácio de Carvalho, V., de Carvalho, M., and Branscum, A. (2018) Bayesian bootstrap inference for the ROC surface. *Stat*, **7**, e211.

unemployment *US Unemployment Rate*

Description

US monthly unemployment rate from January 1967 to November 2009; the 515 monthly observations are seasonally adjusted.

Usage

```
unemployment
```

Format

A time series with 515 observations; the object is of class `ts`.

Source

Bureau of Labor Statistics.

References

de Carvalho, M., Turkman, K. F. and Rua, A. (2013) Dynamic threshold modelling and the US business cycle. *Journal of the Royal Statistical Society, Ser.* C, **62**, 535-550.

See Also

https://webhomes.maths.ed.ac.uk/~mdecarv/decarvalho2013ash.html

Examples

```
## de Carvalho et al (2013; Fig. 1)
data(unemployment)
plot(unemployment, xlab = "Time", ylab = "Unemployment Rate")
```

us_torn *US Tornado Losses (NOAA Severe Weather Database)*

Description

The `us_torn` data frame has 70037 rows and 29 columns. It contains tornado event records from the NOAA Severe Weather Database, including monetary loss information. In our chapter we focus on the re-insurance perspective by considering losses in excess of 15 million USD over the period since 2000, yielding a sample of size $n = 243$.

Format

This data frame contains the following columns:

`om` Integer. NOAA event identifier.

`yr, mo, dy` Integers. Year, month and day of the event.

`date` Character. Event date (as provided in the source file).

`time` Integer/character. Event time (as provided).

`tz` Character. Time zone code.

`st` Character. US state abbreviation.

`stf` Integer. State FIPS code.

`stn` Integer. Station/zone identifier (as provided).

`mag` Integer. Tornado magnitude/scale (as provided).

`inj` Integer. Number of injuries.

`fat` Integer. Number of fatalities.

`loss` Numeric. Property loss amount with a unit change over time; see Details.

`closs` Numeric. Crop loss (as provided).

`slat, slon` Numeric. Starting latitude/longitude.

`elat, elon` Numeric. Ending latitude/longitude.

`len` Numeric. Path length.

`wid` Numeric. Path width.

`ns, sn, sg` Integer/character. Additional source fields (as provided).

`f1, f2, f3, f4, fc` Integers. County/forecast zone identifiers (as provided).

Details

Unit harmonization for the loss variable. The original loss column `loss` (column N in the raw file) is expressed in *million USD* up to year 2016, and in *USD* from year 2017 onward. In the provided file, this corresponds to rows 1–61217 (million USD) and rows 61218–70037 (USD). To convert all losses to USD, we apply:

```
us_torn_data <- us_torn$loss
us_torn_data[1:61217] <- us_torn_data[1:61217] * 10^6
```

Restriction to the period since 2000 and conversion to billion USD. In the provided file ordering, the period since 2000 corresponds to rows 41143–70037. We compute:

```
us_torn_data_2000 <- us_torn_data[41143:70037]
us_torn_data_B <- us_torn_data_2000 / 10^9
```

Re-insurance perspective (losses in excess of 15 million USD). Finally, we retain losses above 15 million USD (i.e., 0.015 billion USD):

```
data <- us_torn_data_B[ which(us_torn_data_B > (15/10^3)) ]
```

This produces a sample of size $n = 243$ used in Chapter 22 of the *Handbook of Statistics of Extremes*. Use `dataset("us_torn")` to load these data from GitHub.

Source

NOAA (National Oceanic and Atmospheric Administration), Severe Weather Database.

References

Daouia, A. and Stupfler, G. (2026). Risk measures beyond quantiles. In: de Carvalho, M., Huser, R., Naveau, P., and Reich, B. J. (eds.), *Handbook on Statistics of Extremes*, Chapter 22, pp. 493–515. Chapman & Hall/CRC, Boca Raton, FL.

de Carvalho, M., Huser, R., Naveau, P., and Reich, B. J. (eds.) (2026). *Handbook on Statistics of Extremes*. Chapman & Hall/CRC, Boca Raton, FL.

NOAA Storm Prediction Center, Severe Weather Database (see https://www.spc.noaa.gov/wcm/#data).

---

venice *Venice Sea Levels*

Description

The `venice` data frame contains 3293 observations on 8 variables recording high sea levels in Venice.

Format

This data frame contains the following columns:

`time` A numeric time index.

`data` Sea level (in cm) relative to the reference datum.

`threshold` Threshold level (80cm) used to define exceedances.

`day` Day of the month.

`month` Month of the year.

`year` Calendar year.

`hour` Hour of the recorded observation.

`MOSE` Indicator variable taking value 1 if the MOSE barrier system was in operation, and 0 otherwise.

Details

The values from 1887 to 1981 are reported in Pirazzoli (1982), whereas the observations for 1982–2023 are obtained from the official tables of tides exceeding 80cm published by the City of Venice. The MOSE (Modulo Sperimentale Elettromeccanico) is a system of mobile submerged gates designed to protect the city from high tides (*acqua alta*). Use `dataset("venice")` to load these data from GitHub.

References

Città di Venezia. (2017, May 4). Archivio storico: Livello di marea a Venezia. Comune di Venezia. City of Venice tide archive.

Davison, A., and Miralles, O. (2026). Modeling univariate extremes—why and how. In: *Handbook of Statistics of Extremes*, Chapter 2, pp. 11–35.

de Carvalho, M., Huser, R., Naveau, P., and Reich, B. J. (2026). *Handbook of Statistics of Extremes*. Chapman & Hall/CRC, Boca Raton, FL.

Pirazzoli, P. A. (1982). Maree estreme a Venezia (periodo 1872-1981). Acqua e Aria, 10, 1023–1039.

`waveheights` *Wave Heights and Locations*

Description

Wave-height observations at 100 locations in the French coast, together with their spatial coordinates.

Format

The dataset contains the following objects:

`hs_dat` A data frame with 1895 observations of wave-height measurements at 100 locations.

`Locations` A data frame containing the `Latitude`, `Longitude`, and `node` of resourcecode grid for each location.

Details

Use `dataset("waveheights")` to load these data from GitHub.

References

de Carvalho, M., Huser, R., Naveau, P., and Reich, B. J. (2026). *Handbook of Statistics of Extremes*. Chapman & Hall/CRC, Boca Raton, FL.

Dombry, C., Legrand, J., and Opitz, T. (2026). Pareto processes for threshold exceedances in spatial extremes. In: *Handbook of Statistics of Extremes*, Chapter 16, pp. 349–376.

wildfire *Portugal Wildfire Data*

Description

The wildfire data from Portugal contains daily burnt area (in hectares) for wildfires in Portugal, and Canadian Forest Fire Weather Index System indices between 1980 to 2019.

Usage

```
wildfire
```

Format

wildfire is a data frame with 14609 occurances (rows) and 11 variables (columns).

The wildfire data frame contains the following columns:

Burnt_Area : daily burnt area in hectares.

DSR : Daily Severity Rating (DSR), a numeric rating of the difficulty of controlling fires.

FWI : Fire Weather Index (FWI), a numeric rating of fire intensity.

BUI : Buildup Index (BUI), a numeric rating of the total amount of fuel available for combustion.

ISI : Initial Spread Index (ISI), a numeric rating of the expected rate of fire spread.

FFMC : Fine Fuel Moisture Code (FFMC), a numeric rating of the moisture content of litter and other cured fine fuels.

DMC : Duff Moisture Code (DMC), a numeric rating of the average moisture content of loosely compacted organic layers of moderate depth.

DC : Drought Code (DC), a rating of the average moisture content of deep, compact organic layers.

day, month, year : timestamp to date for each datapoints.

Source

Instituto Dom Luiz

References

Lee, M. W., de Carvalho, M., Paulin, D., Pereira, S., Trigo, R., and da Camara, C. (2026). BLAST: A Bayesian lasso tail index regression model with an application to extreme wildfires. *Submitted*.

Examples

```
## preview of the data
data(wildfire)
head(wildfire, 10)
summary(wildfire)

## Not run:
```

```
require(ggplot2)
## visualizing the data by month
ggplot(wildfire, aes(x = month, y = Burnt_Area, color = month)) +
    geom_point(size = 3) +
    xlab("Month") +
    ylab("Burnt Area (ha)") +
    theme_minimal()

## End(Not run)
```

Index