Package 'DATAstudio'

March 29, 2025

Version 1.2.1

Date 2025-03-16

Title The Research Data Warehouse of Miguel de Carvalho

Description Pulls together a collection of datasets from Miguel de Carvalho research articles. 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 (2024) <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://www.maths.ed.ac.uk/~mdecarv/

NeedsCompilation no

Date/Publication 2025-03-29 12:30:02 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. For a complete list of datasets and documentation, type `help.start()` and follow the link to **DATAstudio** on the Package Index.

If you use data from this package in publications, please cite the package and the references provided below.

Funding

Generative AI Lab (Univ. of Edinburgh). Royal Society of Edinburgh.

Author(s)

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

See Also

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

`alps` *Swiss Alps Temperature Data*

Description

The `alps` data 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 <- "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)
```

`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. (2024, to appear) Heavy-tailed NGG-mixture models. *Bayesian Analysis*.

`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.

Usage

```
data(california)
```

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.

Source

California State Geoportal.

References

de Carvalho, M., Huser, R., Naveau, P., and Reich, B. J. (2025, to appear). *Handbook on Statistics of Extremes*. Chapman & Hall/CRC. Boca Raton, FL.

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)
```

---

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://www.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))
```

---

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 University of Edinburgh servers. 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 link.

Examples

```
## Download data
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")
```

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)
```

`faang` *FAANG Data*

Description

Daily information on FAANG stocks.

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

To load the file use the command dataset("faang"). FAANG is an acronym for popular tech stocks, namely (Meta's) Facebook, Apple, Amazon, Netflix, and (Alphabet's) Google. The data consist of prices at close for these stocks over 2012-2024. The data were gathered from Yahoo Finance.

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)
```

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://www.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://www.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)
```

---

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.

Usage

```
hongkong
```

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.

References

de Carvalho, M., Huser, R., Naveau, P., and Reich, B. J. (2025, to appear). *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.

Usage

```
data(hurricane)
```

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).

Source

National Hurricane Center and Brian A. Fannin.

References

de Carvalho, M., Huser, R., Naveau, P., and Reich, B. J. (2025, to appear). *Handbook on Statistics of Extremes*. Chapman & Hall/CRC. Boca Raton, FL.

lisbon *Rainfall Data from Lisbon, Portugal*

Description

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

Usage

```
lisbon
```

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.

Source

IPMA (Instituto Português do Mar e da Atmosfera).

References

de Carvalho, M., Huser, R., Naveau, P., and Reich, B. J. (2025, to appear). *Handbook on Statistics of Extremes*. Chapman & Hall/CRC. Boca Raton, FL.

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, Portugal*

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).

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.

---

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. (2025, to appear). *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)
```

`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)
```

---

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.

`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.

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.

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.

sydney *Monthly sea levels for Fort Denison (Sydney)*

Description

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

Usage

```
data(sydney)
```

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).

Source

Australina Government, Bureau of Meteorology.

References

de Carvalho, M., Huser, R., Naveau, P., and Reich, B. J. (2025, to appear). *Handbook on Statistics of Extremes*. Chapman & Hall/CRC. Boca Raton, FL.

thefts *Thefts in Buenos Aires*

Description

To load the file use the command `dataset("thefts")`. 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

```
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://www.maths.ed.ac.uk/~mdecarv/decarvalho2013ash.html

Examples

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

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. (2025). 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