Functions: CEPHaStat

Monday 29th November, 2021

Version 3.2

Date Monday 29th November, 2021

Description CEPHaStat is a collection of R functions produced by Working Group 4 of the UKRI GCRF project CEPHaS for use in the analysis of survey and experimental data on conservation agriculture. It is intended ultimately that these be aggregated into an R package. Note that this document contains all driver functions, called to complete particular tasks, but does not yet cover every function written for subtasks and called by the drivers.

License GNU GPL v 3.0

Requirements Libraries: `MASS, Hmisc, saem, RColorBrewer`

Maintainer All comments and queries to CEPHaS Working Group 4,
via murray.lark@nottingham.ac.uk, chawezi.miti@nottingham.ac.uk, or jchimungu@luanar.ac.mw

Contents

1 Summary statistics and plots

1.1 `cor.sig` Correlation and Significance

Description

Takes a dataframe of two or more variables and outputs their correlations, and their p-values.

Usage `cor.sig (x, roc, rop)`

Arguments

`x`	a data frame of numeric values with at least two columns and five rows
`rop`	the number of decimal places for the p-values. The default value is 3
`roc`	the number of decimal places for the correlation coefficients. The default value is 3

Value

A matrix with correlation coefficients above the diagonal and corresponding p-values below the diagonal.

Author(s) CM

References

Dalgaard. P. (2008). Introductory statistics with R. Springer Science and Business Media

1.2 `histoplot` Plot a histogram with a box and whisker plot

Description

A function to produce a histogram and boxplot of a set of data.

Usage `histoplot(x,varname)`

Arguments

`x`	a vector of numeric values
`varname`	the name of the variable (optional), character so in quotes e.g. `"Clay content"`. If not used then the variable is called `x` on plots.

Value

A plot is produced with a histogram of the data, and, over this, a boxplot. The upper and lower limits of the box are the third and first quartiles of the data. The line in the middle of the box is the median value. The whiskers show the range of data values which fall within the inner fences (Tukey, 1977) of the data. The inner fences are at 1.5 times the interquartile range (quartile 3 minus quartile 1) above quartile 3 and below quartile 1 respectively.

Note

The function `na.drop` is used to remove any missing values from the data before the analysis is undertaken. Quantiles are obtained using the default option (method 7) of the R base function `quantile`.

Author(s) RML

References

Tukey, J.W. 1977. *Exploratory data analysis*. Addison-Wesley

1.3 `kurt` Coefficient of kurtosis

Description

A function to calculate an estimate of the coefficient of kurtosis from a set of data.

Usage `kurt(x)`

Arguments

`x`	a vector of numeric values

Value

The reduced coefficient of kurtosis. This is obtained as the ratio of the estimated fourth moment of the data to the fourth power of the sample standard deviation. Three is then subtracted from the result to give a quantity with an expected value of zero in the case of a normal random variable:

$$\kappa_{\mathrm{r}} \;=\; \frac{m_4}{s^4} - 3 \;=\; \frac{\sum_{i=1}^{n} \{x_i - \bar{x}\}^4 / (n-1)}{s^4} - 3,$$

where n is the number of data, s is the sample standard deviation, $\bar{x}$ is the sample mean and x_i denotes the value of the i^{th} observation.

Note

The function `na.drop` is used to remove any missing values from the data before the analysis is undertaken.

Author(s) RML

References

1.4 `ocskew` Octile skewness

Description

A function to calculate an estimate of the octile skewness from a set of data.

Usage `ocskew(x)`

Arguments

`x`	a vector of numeric values

Value

The octile skewness, due to Brys *et al.* (2003). This measures the asymmetry of a data set by comparing the differences of the first and seventh octile of the data from their median. This value is normalized by the difference between the seventh and first octile, so if O_i denotes the i^{th} octile of a data set (the value such that proportion $i/8$ of the observations are smaller), then :

$$\gamma_{\text{Oc}} = \frac{(O_7 - O_4) - (O_4 - O_1)}{O_7 - O_1},$$

Note that O_4 is the median value. Variables with a distribution from Tukey's g-family (Hoaglin *et al.,* 1985) with the conventional coefficient of skewness in the range [-1,1] have an octile skewness in the range [-0.2,0.2] (Rawlins *et al.,* 2005).

Note

The function `na.drop` is used to remove any missing values from the data before the analysis is undertaken. Quantiles are obtained using the default option (method 7) of the R base function `quantile`.

Author(s) RML

References

Brys, G., Hubert, M., & Struyf, A. 2003. A comparison of some new measures of skewness. In: *Developments in robust statistics*, eds R. Dutter, P. Filzmoser, U. Gather & P.J. Rousseeuw. Physica-Verlag Heidelberg pp 98–113.

Hoaglin D.C., Mosteller F. & Tukey J.W. 1985. *Exploring data tables, trends, and shapes*. Wiley, New York.

Rawlins, B. G., Lark, R. M., O'Donnell, K. E., Tye, A. & Lister, T. R. 2005. The assessment of point and diffuse soil pollution from an urban geochemical survey of Sheffield, England. *Soil Use and Management*, **21**, 353–362.

1.5 `outliers` Identify probable outliers

Description

A function to produce summary plots of a set of data.

Usage `outliers(x,trim)`

Arguments

`x`	a vector of numeric values
`trim`	Logical variable. Optional. See below for explanation.

Value

This function will count the number of probable outliers (Tukey, 1977) in the data. Probable outliers are values that fall outside Tukey's (1977) 'outer fences'. The outer fences are at 3 times the interquartile range (quartile 3 minus quartile 1) above quartile 3 and below quartile 1 respectively. The function will print the values and the indices (positions in the input vector) of any probable outliers by this criterion. If `trim` is set to `F`, or is not set, then the output of the function is the index of any outlying observations. If `trim` is set to `T` then the output is a vector with the data removed.

Note

The function `na.drop` is used to remove any missing values from the data before the analysis is undertaken. Quantiles are obtained using the default option (method 7) of the R base function `quantile`.

Author(s) RML

References

Tukey, J.W. 1977. *Exploratory data analysis*. Addison-Wesley

1.6 `qqnorm.line` Plot empirical and normal quantiles, with line added

Description

A function to produce summary plots of a set of data.

Usage `qqnorm.line(x,col)`

Arguments

`x`	a vector of numeric values
`col`	Colour of the line on the plot. Optional.

Value

This function produces a QQ plot, a plot of the theoretical quantiles (standard normal variable) against the corresponding quantiles of the data. If the data follow a normal distribution then the plot is expected to be a straight line, sitting on the line drawn.

Author(s) CM

References

Tukey, J.W. 1977. *Exploratory data analysis*. Addison-Wesley

1.7 `skew` Coefficient of skewness

Description

A function to calculate an estimate of the coefficient of skewness from a set of data.

Usage skew(x)

Arguments

`x`	a vector of numeric values

Value

The coefficient of skewness. This is obtained as the ratio of the estimated third moment of the data by the third power of the sample standard deviation:

$$\gamma = \frac{m_3}{s^3} = \frac{\sum_{i=1}^{n} \{x_i - \bar{x}\}^3 / (n-1)}{s^3},$$

where n is the number of data, s is the sample standard deviation, $\bar{x}$ is the sample mean and x_i denotes the value of the i^{th} observation.

Note

The function `na.drop` is used to remove any missing values from the data before the analysis is undertaken.

Author(s) RML

References

1.8 summa Summary statistics

Description

A function to calculate summary statistics of a set of data.

Usage `summa(x,sigf)`

Arguments

`x`	a vector of numeric values
`sigf`	the number of significant figures to report (optional)

Value

A matrix containing the mean value, median value first and third quartiles, sample variance, sample standard deviation, coefficient of skewness, octile skewness, coefficient of kurtosis and the number of probable outliers in a data set. Probable outliers are defined as values falling outside Tukey's (1977) outer fences (see function `outliers`).

Note

The function `na.drop` is used to remove any missing values from the data before the analysis is undertaken. Quantiles are obtained using the default option (method 7) of the R base function `quantile`. See functions `skew`, `kurt` and `ocskew` for definitions of these statistics.

Author(s) RML

References

Tukey, J.W. 1977. *Exploratory data analysis*. Addison-Wesley

1.9 `summaplot` Summary plot

Description

A function to produce summary plots of a set of data.

Usage `summaplot(x,varname)`

Arguments

`x`	a vector of numeric values
`varname`	the name of the variable (optional), character so in quotes e.g. `"Clay content"`. If not used then the variable is called `x` on plots.

Value

A plot is produced with a histogram of the data, and, over this, a boxplot. The upper and lower limits of the box are the third and first quartiles of the data. The line in the middle of the box is the median value. The whiskers show the range of data values which fall within the inner fences (Tukey, 1977) of the data. The inner fences are at 1.5 times the interquartile range (quartile 3 minus quartile 1) above quartile 3 and below quartile 1 respectively. The second window shows the QQ plot, a plot of the theoretical quantiles (standard normal variable) against the corresponding quantiles of the data. If the data follow a normal distribution then the plot is expected to be a straight line, sitting on the line drawn. Points shown in red correspond to Tukey's (1977) probable outliers (see notes for function `outliers`).

Note

The function `na.drop` is used to remove any missing values from the data before the analysis is undertaken. Quantiles are obtained using the default option (method 7) of the R base function `quantile`.

Author(s) RML

References

Tukey, J.W. 1977. *Exploratory data analysis*. Addison-Wesley

2 Geostatistics

2.1 `okgridvar` Compute kriging variance at the centre of a square grid cell

Description

A function to produce summary plots of a set of data.

Usage `okgridvar(space,modtyp,c0,c1,a1)`

Arguments

`space`	real value, the interval between nodes on a square sampling grid
`modtyp`	character string `"Sph"` or `"Exp"`. See below for details.
`c0`	the nugget variance of the specified variogram model.
`c1`	the correlated variance of the specified variogram model.
`a1`	the distance parameter of the specified variogram, in the same units as `space`

Value

The returned value is the ordinary kriging variance for a prediction at the centre of a grid cell (i.e. at maximum distance from any sample point) when kriging from the nearest 36 observations on a square grid. This assumes that the variogram is either spherical (`modtyp="Sph"`) or exponential (`modtyp="Exp"` with nugget variance, correlated variance and distance parameter taking values `c0`, `c1` and `a1` respectively. This is based on the ossfim algorithm of McBratney *et al.* (1981).

Note

In future this function will be modified to allow the number of neighbours to be set, and a wider range of variogram models.

Author(s) RML

References

McBratney, A.B., R. Webster 1981. The design of optimal sampling schemes for local estimation and mapping of regionalized variables: 2 program and examples. *Computers and Geosciences* **7**, 335–365.

2.2 `ossfim` Compute kriging variance at the centre of a square grid cell: wrapper for `okgridvar` for use with `geoR` model objects

Description

A function to produce summary plots of a set of data.

Usage `ossfim(mod,spmin,spmax,L,wrout)`

Arguments

`mod`
`spmin`
`spmax`
`L`
`wrout` optional logical variable. If TRUE then the function will return a matrix with two columns: grid spacing and corre

Value

The function creates a plot of kriging variance against grid spacing, and indicates the spacing which achieves the target L if this is in the range [$spmin$,$spmax$]. If `wrout=TRUE` then a matrix with two columns, grid spacing and corresponding kriging variances, is returned. This is based on the ossfim algorithm of McBratney *et al.* (1981).

Note

In future this function will be modified to allow the number of neighbours to be set, and a wider range of variogram models.

Author(s) RML

References

McBratney, A.B., R. Webster 1981. The design of optimal sampling schemes for local estimation and mapping of regionalized variables: 2 program and examples. *Computers and Geosciences* **7**, 335–365.

2.3 `sv` Return the value of the variogram

Description

A function to return the value of the variogram for a specified lag distance, model type and set of parameters.

Usage `sv(lag,modtype,c0,c1,a1)`

Arguments

`lag`	the lag distance, a scalar value in the same units as `a1`.
`modtyp`	the variogram model type, either `Sph` (spherical) or `Exp` (exponential)
`c0`	the nugget variance
`c1`	the correlated variance component, sometimes called the partial sill
`a1`	the distance parameter of the variogram model

Value

The value of the variogram is returned for the specified lag distance, given the variogram parameters and model type also included in the function's arguments.

Note

In future this function will be modified to allow a wider range of variogram models.

Author(s) RML

References

3 Soil physics

3.1 `VanGenuchten.fit.single` Fit a Van Genuchten water release curve to a single set of observations

Description

A function to fit the parameters of a Van Genuchten water release curve to an unreplicated set of observations. In CEPHaStat we assume by default that tensions are given in units of kPa. At a tension of h the volumetric water content is $\theta(h)$ where

$$\theta(h) \;=\; \theta_{\mathrm{r}} \;+\; (\theta_{\mathrm{s}} - \theta_{\mathrm{r}})\,\frac{1}{\{1+(\alpha h)^n\}^m},$$

where θ_{s} and θ_{r} are respectively the volumetric water content at saturation and the residual water content, α is related to the reciprocal of air-entry tension, and n is a parameter related to the pore size distribution. Here we do not fit m as a separate parameter but set it to $m = 1 - n^{-1}$.

Usage `(VanGenuchten.fit.single(data.df,init.vals,roundoff=4))`

Arguments

`data.df`	data.df data frame with tensions (kPa) in variable `h` and volumetric water content in variable `theta`
`init.vals`	init.vals a vector with starting point for parameters, in order `thr,ths,alp,nscal` corresponding to $\theta_{\mathrm{r}},\ \theta_{\mathrm{s}},\ \alpha,\ n$
`roundoff`	the number of decimal places for the returned values (defaults to 4)

Value

A list with a single item `Coefficient.estimates` which contains the estimated parameters in order `thr,ths,alp,nscal` corresponding to $\theta_{\mathrm{r}},\ \theta_{\mathrm{s}},\ \alpha,\ n$

Note

This function uses `nls`, an `R` function to fit a non-linear model. The initial values should be chosen with care

Author(s) CM, RML

References

3.2 `VanGenuchten.fit.group` Fit a single Van Genuchten water release curve to a replicated set of observations

Description

A function to fit the parameters of a single Van Genuchten water release curve to a replicated set of observations. In CEPHaStat we assume by default that tensions are given in units of kPa. At a tension of h the volumetric water content is $\theta(h)$ where

$$\theta(h) \;=\; \theta_{\mathrm{r}} \;+\; (\theta_{\mathrm{s}} - \theta_{\mathrm{r}})\,\frac{1}{\{1+(\alpha h)^n\}^m},$$

where θ_{s} and θ_{r} are respectively the volumetric water content at saturation and the residual water content, α is related to the reciprocal of air-entry tension, and n is a parameter related to the pore size distribution. Here we do not fit m as a separate parameter but set it to $m = 1 - n^{-1}$.

Usage `(VanGenuchten.fit.group(data.df,init.vals,g.name))`

Arguments

`data.df`	data.df data frame with tensions (kPa) in variable `h` and volumetric water content in variable `theta` a factor `g.name` which groups all observations from the same soil sample.
`init.vals`	init.vals a vector with starting point for parameters, in order `thr,ths,alp,nscal` corresponding to $\theta_{\mathrm{r}},\ \theta_{\mathrm{s}},\ \alpha,\ n$
`g.name`	A factor which groups together the observations from a single soil (or an averaged set of values from a single plot).

Value

A list with an item `Coefficient.estimates` which contains the estimated parameters in order `thr,ths,alp,nscal` corresponding to $\theta_{\mathrm{r}},\ \theta_{\mathrm{s}},\ \alpha,\ n$ and an item `Standard.errors` which contains the standard errors of the estimates.

Note

This function uses `saemix`, an `R` implementation of the Stochastic Approximation EM algorithm to fit a non-linear model. The initial values should be chosen with care, and it may be advisable to use `VanGenuchten.fit.single` to find initial values. As set up the algorithm will use importance sampling, and graphical outputs are suppressed.

Author(s) RML, CM

References

Comets, E., Lavenu, A., Lavielle, M. 2017. Parameter estimation in nonlinear mixed effect models using saemix, an R implementation of the SAEM algorithm. *Journal of Statistical Software*, **80**, 1–41.

3.3 `VanGenuchten.fit.compare` Fit and compare Van Genuchten water release curves for different sets of replicated observations

Description

A function to fit the parameters of a separate Van Genuchten water release curve to two or more sets of replicated set of observations, which may be defined, for example, by experimental treatment or soil type. Two models can be obtained and compared, a 'null' model in which some or all parameters are common across the groups, and an alternative model in which more parameters are fitted separately for each group. The coefficients for each group in the alternative model (i.e. with most parameters differing between the groups) are exported as are the log-likelihood ratio for the comparison between the alternative and the null, and a P-value for the null hypothesis that the separately-fitted parameters do not differ between the groups. models. In CEPHaStat we assume by default that tensions are given in units of kPa. At a tension of h the volumetric water content is $\theta(h)$ where

$$\theta(h) \;=\; \theta_{\mathrm{r}} \;+\; (\theta_{\mathrm{s}} - \theta_{\mathrm{r}})\,\frac{1}{\{1+(\alpha h)^n\}^m},$$

where θ_{s} and θ_{r} are respectively the volumetric water content at saturation and the residual water content, α is related to the reciprocal of air-entry tension, and n is a parameter related to the pore size distribution. Here we do not fit m as a separate parameter but set it to $m = 1 - n^{-1}$.

Usage `(VanGenuchten.fit.compare(data.df,init.vals,g.name,par.var.null,par.var.full,cov.name))`

Arguments

`data.df`	data.df data frame with tensions (kPa) in variable `h` and volumetric water content in variable `theta` a factor `g.name` which groups all observations from the same soil sample.
`init.vals`	init.vals a vector with starting point for parameters, in order `thr,ths,alp,nscal` corresponding to $\theta_{\mathrm{r}},\ \theta_{\mathrm{s}},\ \alpha,\ n$
`g.name`	A factor which groups together the observations from a single soil (or an averaged set of values for a plot) from a single plot).
`par.var.null`	a vector coding which parameters (same order as init.vals) are common over groups (0) and which vary (1) between the groups in the null model
`par.var.full`	a vector coding which parameters (same order as init.vals) are common over groups (0) and which vary (1) between the groups in the alternative model. Note that the two models must be nested, so any parameters estimated separately in the null model must also be estimated separately in the alternative.
`cov.name`	A factor which groups together the observations by group for which separate models are to be fitted.

Value

A list with an item `Coefficient.estimates` which contains the estimated parameters in order `thr,ths,alp,nscal` corresponding to $\theta_{\mathrm{r}},\ \theta_{\mathrm{s}},\ \alpha,\ n$ by group; an item `Comparison` which shows which parameters were common and which estimated separately in the null and alternative models, an item `Inference` which presents the log-likelihood ratio statistic for a comparison of the null and alternative models, and a P-value to test the null hypothesis

that the additional parameters estimated separately in the alternative model do not differ between the groups.

Note

This function uses `saemix`, an `R` implementation of the Stochastic Approximation EM algorithm to fit a non-linear model. The initial values should be chosen with care, and it may be advisable to use `VanGenuchten.fit.single` to find initial values. As set up the algorithm will use importance sampling, and graphical outputs are suppressed.

Author(s) RML, CM

References

Comets, E., Lavenu, A., Lavielle, M. 2017. Parameter estimation in nonlinear mixed effect models using saemix, an R implementation of the SAEM algorithm. *Journal of Statistical Software*, **80**, 1–41.

3.4 plot.wrc.data Plot points from experimental water retention curves, optionally with fitted Van Genuchten model(s)

Description

This function will make a plot of data which consist of measurements of the water release curve (volumetric water content against tension with the latter on a log scale). If provided, coefficients for the Van Genuchten function are used to draw the corresponding model on the plot. If data (and models) for more than one sets of measurements (e.g. different treatments) are provided then these are plotted separately on the graph.

Usage `(plot.wrc.data(data.df,coeffs,hvar="h",thetavar="theta", xlab="Tension /kPa",maxth=-1, groups="none",main=""))`

Arguments

`data.df`	data.df data frame with tensions (kPa) in variable `hvar` and volumetric water content in variable `thetavar` a factor `g.name` which groups all observations from the same soil sample.
`coeffs`	optional set(s) of Van Genuchten parameters for models to be drawn on the plot. In the format of the item `Coefficient.estimates` exported by the `VanGenuchten.fit...` functions. This can be one set or multiple, in which case `data.df` must contain a factor for the points in the different groups with name set by `groups =` as described below
`hvar`	optional name of variable in `data.df` which contains tensions (kPa), `"h"` by default
`thetavar`	optional name of variable in `data.df` which contains volumetric water content, `"theta"` by default
`xlab`	optional label for ordinate of plot. By default `"Tension /kPa"`
`maxth`	optional maximum volumetric water content for plot, by default this is adjusted to the data and any model(s)
`groups`	optional name of factor determining groups in the data (e.g. treatments) to be plotted in distinct colours
`main`	optional character string as a main title for the plot

Value

No output, a plot is produced in the active graphics window

Note

Author(s) RML, CM

References

3.5 `SQ.indices` Compute soil quality indices given Van Genuchten parameters

Description

Computes a set of quality indices from Van Genuchten parameters and (optionally) bulk density. The indices and their interpretation is based on papers of Dexter (2004) and Reynolds *et al*. (2007).

Usage `(SQ.indices((coeffs,bulk_density,t_m=4.9,t_FC=9.8, t_PWP=1471)`

Arguments

`coeffs`	Van Genuchten parameters (one or more sets) in the format of the item `Coefficient.estimates` exported by the `VanGenuchten.fit...` functions
`bulk_density`	optional vector of bulk density values (g cm^{-3}), one per parameter set. If not provided bulk density is computed from the value(s) of θ_{S} (see notes)
`t_m=4.9`	optional value of tension (kPa) at which all macropores are empty (see notes)
`t_FC=9.8`	optional value of tension (kPa) at field capacity (see notes)
`t_PWP=1471`	optional value of tension (kPa) at permanent wilting point (see notes)

Value

A list with two elements. The first, `Indices`, contains values for Dexter's S (Dexter, 2004), total porosity, macroporosity, relative water capacity, plant available water capacity and air capacity. The second element, `Interpretation`, gives the interpretation of each of these values following Dexter (2004) and Reynolds *et al*. (2007) as discussed in more detail below. In each element there are as many rows as sets of parameters in `coeffs`.

Note

Dexter's S is the modulus (absolute value) of the gradient of the water release curve at its inflection point (i.e. where the slope stops increasing with increased tension), interpreted in terms of the microstructure of the soil, which is better-defined, with a wider range of pore sizes, when S is large. Dexter (2004) gives an interpretation of values of S which this function reproduces.

S	
$S > 0.035$	Good microstructural quality
$0.02 < S \leq 0.035$	Poor microstructural quality
$S \leq 0.02$	Very poor microstructural quality

Note that S is defined with respect to the water release curve for *gravimetric* water content, unlike in conventional usage. For this reason a bulk density value is required to rescale the

volumetric water content computed from the Van Genuchten parameters. If bulk density is not measured directly, then it is computed from the value of θ_{S}, assuming a particle density of 2.65 g cm^{-3}: – this may underestimate particle density for very ferruginous soils, and overestimate it for soils with a large organic content (Landon, 1984).

Total porosity is equal to θ_{S}, i.e. the volumetric water content of the saturated soil.

Macroporosity is the difference between total porosity and porosity at a tension when it is assumed that only micropores are filled (matrix porosity). Reynolds *et al*. (2007) suggest three values, by default we use the middle one (4.9 kPa), but this can be changed by setting the optional parameter `t_m` to the selected value. The interpretation is based on Reynolds *et al*. (2007), if macroporosity is $\leq$0.04 (volumetric) then the soil is assumed to be degraded by compaction or consolidation. Otherwise, for medium to fine textured soils, it is regarded as undegraded.

The relative water capacity, RWC, is defined as the ratio of the volumetric water content at field capacity to the total porosity. This is interpreted as optimal for microbial activity in the interval $0.6 < \mathrm{RWC} \leq 0.7$, too dry below the range and too wet above (Reynolds *et al*., 2007).

Field capacity by default is assumed to be the volumetric water content at a tension of 9.8 kPa, but this can be changed by adjusting the optional argument `t_FC`. Permanent wilting point by default is assumed to be the volumetric water content at a tension of 1471 kPa, but this can be changed by adjusting the optional argument `t_PWP`.

Plant available water capacity, PAW, is the difference between the water content at field capacity and the permanent wilting point. Following Reynolds *et al*. (2007), this is interpreted as follows:

PAW m^3 m^{-3}	
$\mathrm{PAW} > 0.2$	Ideal
$0.15 < \mathrm{PAW} \leq 0.2$	Good
$0.1 < \mathrm{PAW} \leq 1.5$	Limited
$\mathrm{PAW} \leq 0.1$	Poor

The air capacity, AC, of the soil is the difference between the total porosity and the field capacity. Following Reynolds *et al*. (2007) these values are interpreted as follows

AC m^3 m^{-3}	
$\mathrm{AC} > 0.15$	Aeration likely to be adequate for all soils
$0.10 < \mathrm{AC} \leq 0.15$	Aeration likely to be adequate except for fine-textured soils
$\mathrm{AC} \leq 0.1$	Crop-damaging aeration deficit likely

a value less than 0.1 (volumetric) is interpreted as likely to lead to crop-damaging aeration deficit, aeration is likely to be adequate if air capacity is between 0.1 and 0.15, except for fine-textured soils, and a value exceeding 0.15 is likely to mean adequate aeration in all soils.

Author(s) CM,RML

References

Dexter, A.R. 2004. Soil physical quality. Part I. Theory, effects of soil texture, density and organic matter, and effects on root growth. *Geoderma*, **120**, 201–214.

Landon, J.R. (ed), 1984. Booker Tropical Soil Manual (1st Edn) Longman, Harlow. page 97, section 6.5.

Reynolds, W.D., Drury, C.F., Yang, X.M., Fox, C.A., Tan, C.S., Zhang, T.Q. 2007. Land management effects on the near-surface physical quality of a clay loam soil. *Soil & Tillage Research*, **96**, 316–330.

3.6 `print.indices` Print soil quality indices from output of `SQ.indices`

Description

Prints the soil quality indices in an object produced by the function `SQ.indices` along with their interpretation. Optionally returns these values as a dataframe.

Usage `print.indices(index.values,ret=F)`

Arguments

`index.values`	a set of soil quality index values with their interpretation produced by `SQ.indices`
`ret`	Optional logical value, default FALSE. If TRUE the function returns its output as a dataframe.

Value

If `ret=T` then a dataframe with index values and their interpretations in the columns

Note

Author(s) CM,RML

References

4 Estimation for censored (log) normal variables

4.1 `mean.censor` Estimate the mean of a variable which is censored (left or right) by a known value

Description

A function to take data in a vector which include values fixed at a censoring value (left or right), and to estimate the mean, with its standard error, and standard deviation, assuming that the variable is normally distributed prior to censoring. If the data are first transformed to natural logarithms, then an unbiased backtransform of the mean, and a median-unbiased estimate with 95% confidence bounds, may also be requested.

Usage `mean.censor(y,cen,side.left,log.t)`

Arguments

`y`	Vector containing data on y. This may have all censored values replaced by `cen`, but any values above (right-censoring) or below (left-censoring `cen` are replaced in the function.
`cen`	Censoring value. If `side.left=T` then all values $\leq$`cen` are censored, and replaced by `cen`. If `side.left=F` then all values $\geq$`cen` are censored, and replaced by `cen`.
`left.side`	Optional, default = T If true, then the data are left-censored (see `cen`), If false, then the data are right-censored.
`log.t`	Optional, default = F. If set to true then, in addition to the standard outputs, then it is assumed that $y = \log(x)$ and the unbiased estimate of mean x is also returned with the median unbiased value and its 95% confidence bounds (see notes).

Value

A matrix with the following outputs when `log.t=F`.

`log_likelihood`	The maximized value of the log-likelihood
`Number_censored`	The number of values in `y` affected by censoring
`Mean`	The maximum likelihood estimate of the mean of y
`SD`	The maximum likelihood estimate of the standard deviation of y
`SE_mean`	The standard error of the estimated mean
`Lower_95%`	lower and upper confidence bounds (95%)
`Upper_95%`	for the mean of y
`Mean_uncensored_only`	The mean value of the uncensored values only.

In the even that `log.t=F`, the following additional output is generated

`Back_mean`	The unbiased back-transformed estimate of the mean on the untransformed scale
`Back_median`	The median unbiased back-transformed estimate
`Back_Lower_95%`	lower and upper confidence bounds (95%)
`Back_Upper_95%`	for the median-unbiased back-transformation.

Note

We have a set of observations, Y which consists of a subset of those which are replaced by censoring, Y_c, and the remainder in set Y' so $Y = Y_c \cup Y'$. There are $n_c = |Y_c|$ censored observations. In the left-censored case, with censoring value y_c, the likelihood, for a set of

parameters μ and σ for the normal distribution of the uncensored variable is:

$$L \;=\; \sum_{y_i \in Y'} f\left(y_i|\mu,\sigma\right) + n_c \int_{-infty}^{y_c} f\left(y_c|\mu,\sigma\right)$$

In the case of right censoring this expression takes the form

$$L \;=\; \sum_{y_i \in Y'} f\left(y_i|\mu,\sigma\right) + n_c \int_{y_c}^{\infty} f\left(y_c|\mu,\sigma\right)$$

The estimated values of the parameters are found with the `optim` function in R. The hessian matrix is evaluated at the estimates and used to evaluate their Fisher information matrix and so the covariance matrix, from which the standard error of the estimated mean is extracted.

If the data require transformation to logs to make the proposed censored normal distribution appear plausible, then the transformation is applied before the `mean.censor` function, and the censor value must, of course, also be transformed. If the optional argument `log.t` is set to TRUE, then the output includes, along with the estimated mean and standard deviation, $\widehat{\mu}$ and $\widehat{\sigma}$ on the log-scale, an unbiased estimate of the mean value on the original scale:

$$\widehat{\mu_x} \;=\; \exp\left\{\widehat{\mu} + \frac{\widehat{\sigma^2}}{2}\right\},$$

a median unbiased estimate on the original scale:

$$\mathrm{median}_x \;=\; \exp\left\{\widehat{\mu}\right\},$$

and the 95% confidence bounds for the latter.

Author(s) RML

References