DScreen User Manual

DScreen Version 1.0 $Exp

Raymond Venneker

Rev. January 10, 2011

UNESCO-IHE, Delft, The Netherlands

Venneker, R. (2011) DScreen User Manual (Version 1.0 $Exp), UNESCO-IHE, Delft, The Netherlands.

Contents

[page intentionally left blank]

1 Introduction

1.1 Overview

The computer program DScreen is a utility for analyzing and exploring long-term annual and monthly hydrological, hydrometeorological and environmental time series, with a focus on screening the data for homogeneity and consistency. It is inspired by the older MSDOS-based DatScr program that accompanied Dahmen and Hall (1990), with substantial added features and capabilities. Although primarily developed as a classroom instruction tool, DScreen may also be useful in small self-contained study projects. The program aims to bridge the gap between operational hydrological data base analysis tools, comprehensive general purpose statistics packages, and spreadsheets that can be a nuissance to arrange efficiently and error-free in complicated analyses. It is hoped that by providing a simple, easy to use dedicated program the student or course participant can concentrate on learning the concepts of hydrological time series screening instead of organizing data and constructing algorithms. On the other hand, much of the output is presented in such a way that it also paves the way for comprehensive statistics tools, such as for example R (R Development Team, 2006), used for serious work.

1.2 Main features

DScreen works exclusively with univariate time series. Given the emphasis on screening and analysis of long-term characteristics, the data are typically contained in annual or monthly series covering periods of at least 10 to 20 years, but preferably longer. The analyses are not applicable to daily or sub-daily data. The major features and capabilities of DScreen are summarized as:

- The computation of major statistical descriptors relevant to time series;
- The transformation of data to comply with normality assumptions by parametric analysis;
- The visualization of time series, their statistical character and potential change points; and
- Analyzing time series for long-term statistical consistency, homogeneity and stationarity.

1.3 Installation

DScreen is written as a platform-independent Java application, and requires to have the Java Runtime Environment (JRE) properly installed on your machine. The program has been tested with JRE version 1.6. The latest JRE release can always be downloaded from `www.java.com`.

The DScreen distribution is packed in a single WinZip (Windows) or gzipped tar (Unix) file that contains all program files, documentation and example data. The files need to be extracted and placed in a suitable directory on your computer. Follow the instructions in the `README.txt` file to complete the installation. Mac OS X users can proceed from the Unix instructions but we have no means to verify such implementation.

1.4 Terms of Use

The DScreen program, source code and documentation may be used, reproduced and redistributed free of charge under the terms and conditions of the GNU General Public License (GPL) of the Free Software Foundation. The program makes use of the JFreeChart library which is distributed under the the GNU Lesser General Public License (LPGL). Copies of the GPL and LPGL are included with the distribution files. See also Appendix A.

2 Data Screening

2.1 Purpose and definitions

The purpose of data screening is to analyze long-term hydrological or hydrometeorological observation records for statistical stationarity, which can be subdivided in homogeneity and consistency. Engineering studies for water resources management, development, and planning require stationarity of the data when used for frequency analysis or time series modelling (Dahmen and Hall, 1990; Hipel and McLeod, 1994; Helsel and Hirsch, 2002). On the other hand, data screening also provides the tools to detect and describe trends and abrupt changes in water resources regimes.

A time series is statistically *stationary* if the location (mean, median), the scale (variance), and the autocorrelation structure (serial dependence) do not change with time. A change in the statistical properties of hydrological time series is also referred to as *non-homogeneity* if this arises due to natural or human-induced alterations in climate, the natural environment, or due to station relocation. Furthermore, *inconsistency* refers to a change in the amount of systematic error arising from changes in instrumentation and observing methods. In any event, the users need to be (made) aware how these alterations affect the data and derived information.

When analyzing data at sub-annual time intervals (e.g. monthly), environmental time series commonly exhibit *seasonality*. The mean, variance and possibly the autocorrelation structure change with the time of the year, typically in cycles with a period of 12 months. However, the mean seasonal statistical properties over many years do not change in a homogeneous and consistent time series.

Note that application of the data screening techniques is not necessarily restricted to data from a single location. For the purpose of data quality control and observing network analysis it is possible to screen the data for *relative consistency* between a station and a *reference series*. The latter can be either another nearby station or a regional average from a number of stations. Data screening of relative consistency proceeds from constructing a new series as either the difference or ratio of the values from the station and the reference series.

2.2 Generic data screening procedure

Analyzing time series requires a combination of exploratory data analysis and confirmatory statistics (Hipel and McLeod, 1994). Exploratory data analysis (EDA) is based on computing general statiscal descriptors and application of graphical techniques to obtain insight in the character of the data and to present any intrinsic relations in the data (Tukey, 1977). Confirmatory statistics provide a quantitative analysis through application of statistical hypothesis tests. Below follows a concise overview of a generic data screening procedure. Detailed descriptions of the techniques are published in Dahmen and Hall (1990) and the relevant sections in Gilbert (1987), and Hipel and McLeod (1994). The latter publication and that of Box et al. (2008) also extend the subject to further analysis and modelling of time series.

The procedure is broken down into the following broad steps (after Dahmen and Hall, 1990):

1. Collect and arrange the data into monthly or annual time series. Initial quality control of the daily or sub-daily data to identify and possibly correct errors data or deal with missing data is also required. This step is carried out before the actual data screening.

2. A rough screening of the data is carried out by collecting and verifying totals, averages, internal variability and distribution characteristics.

3. Plot the data in normal or cummulative run-sequence graphs. A well constructed graphical representation is a powerful tool to identify the data characteristics (Cleveland, 1992; Helsel

and Hirsch, 2002, Chapter 16). A trained analyst will immediately see possible trends, change points or other suspicicious features.

4. Test the time series for normality. It is often (but not always) possible to apply a data transform to make the series resemble a normal distribution. Results of statistical tests that assume normally distributed data should be interpreted with care if normality is not achieved.

5. Test the time series for persistence and serial correlation in order to obtain an impression of its autocorrelation structure. If necessary, autocorrelation in a time series can be reduced by pre-whitenining the data.

6. Test the time series for absence of trends. A non-parametric test is preferred here as a robust technique in the absence of knowledge about the underlying statistical distribution, or if normality is not achieved.

7. Identify the most likely change-points and subdivide the data into sub-series that can be tested for stability (i.e. similarity) of variance and mean, using an F-test and t-test, respectively.

2.3 Statistical hypothesis testing

The common way to arrive at a conclusion with respect to a characteristic of the data is to test and accept or reject a hypothesis. This requires a statement of a null hypothesis (H_0) and, consequently, the alternative hypothesis (H_1). For example,

H_0 : The data series does not have a trend,
H_1 : There is a trend in the series.

The idea now is to provide a criterion for which the null hypothesis would be rejected (and therefore the alternative be accepted). The possibility of falsely rejecting H_0 is related to the uncertainty. As such, the conclusion is made for a certain significance level α, which defines the sensitivity of the test. A significance level of $\alpha = 0.05$ means that H_0 is rejected for 5% of the time when it is actually true. The choice for α is somewhat arbitrary. Often, $\alpha = 0.05$, and sometimes $\alpha = 0.01$ are used.

Based on the type of test, a test statistic for the specific hypothesis test is computed. Depending on the statistical distribution of the test statistic, a critical region for that distribution can be defined that contains those values of the test statistic for which H_0 would be rejected. The critical region depends on the choosen significance of the test and the degrees of freedom (closely related to the number of data points) for the series.

Depending on the computation method, there are two ways complete the test:

1. Compare the test statistic with the limiting values of the critical region. Test statistic values above the upper limit or below the lower limit of the critical region result in rejecting H_0.

2. Compare a p-value for the test statistic against the selected significance level. The p-value is the probability that the (absolute) value of the test statistic will be exceeded. If the p-value is smaller than the choosen significance (e.g. 0.05), H_0 is rejected.

For details on statistical test methodology and interpretation the reader is referred to Helsel and Hirsch (2002, Chapter 4), general textbooks on statistics, or to specialized treatments of the subject such as e.g. Sheskin (2007).

3 Using DScreen

3.1 Starting up

When the program is started, a window appears, which contains the menu, a panel for entering start and end times of the (sub)series and a mostly blank text area where all textual output is to be produced. The first step is to use the File-Open menu option to load a time series data file to work with. After the file is read in memomory, the name of the file will appear in the window title and in the output area. The menu is now fully enabled and the data screening can begin.

The (sub)series time bounds can now be set to fit your requirements. For reference, the default time bounds are displayed to the right of the text fields. Depending on whether an annual or monthly time series is used, the text fields show the calendar year (`yyyy`), or the year and the month (`yyyy-mm`). When changing the time fields, they must be entered in the same format, otherwise they will not be recognized. Start and end times are inclusive.

The majority of the analyses are carried on the "series", which is the period of time starting with the first value of "sub-series 1" and ending with the last value of "sub-series 2". These times can never be outside those of the data file. In setting sub-series time bounds for tests that use split-samples, there may be a gap or overlap (although the latter is not recommended) between the sub-series.

When reporting p-values, DScreen appends a significance level indication. A significance level of less than 0.01 is marked by ***. Significance levels between 0.05 and 0.01 are marked by ** and levels between 0.1 and 0.05 are marked by *. If the significance is larger than 0.1, no indication is given. This is common practice in statistical analysis packages, but note that the levels of significance may differ between packages.

3.2 The Data menu

The Data menu has a number of options for the statistical characterization of the time series data. The analyses are carried out for the entire series with results produced in the text area. Output contains a one-line description of the analysis and the start, end and length of the data series. The Data menu provides the following items.

Statistics

This gives the major statistical descriptors of the series, i.e.

`Mean`	The arithmetic mean of the data
`SD`	The standard deviation
`Skew`	The skewness
`SE`	The standard error
`CV`	The coefficient of variation

The standard error represents the standard deviation of the mean. The coefficient of variation is simply the ratio of the standard deviation and the mean.

Summary

This option computes Tukey's five-number summary, characterizing the distribution of the data (Tukey, 1977). The output consists of:

`Min`	The minimum value
`Q1`	The first quartile

`Median`	The median or second quartile
`Q3`	The third quartile
`Max`	The maximum value

Note that there are a number of different methods in use to compute quantiles (Hyndman and Fan, 1995). Therefore, the results are generally slightly different between statistical computation packages.

Apparent trend

The apparent trend is computed as a linear least-squares regression line fit between the time points and the corresponding data values. The output contains the following items for the regression constant (`Const`) and the regression slope (`Slope`) coefficients:

`Estimate`	The values of the coefficients
`SE`	The standard errors of the coefficients
`t`	The t-statistic values
`Pr(>|t|)`	The probabilities of exceeding the absolute t-statistic value

The constant value is the fit for the first point in the series, not the y-axis intercept at time zero (the Gregorian calendar year zero does not exist anyhow). The regression equation to fit a value for any year or month i then becomes $\hat{y} = a + b(t_i - t_1)$, in which a is the regression constant and b is the regression slope. The t-value exceedance probabilities are the p-values for significance of the coefficients, assuming that the resisuals follow a normal distribution. A p-value larger than the choosen significance level indicates that the coefficient is not different from zero (and could have been omitted from the equation).

Additional diagnostic output to judge the correlation between the variables consists of:

`Residual SE`	The standard error of the residuals of the regression fit
`Regression DF`	The degrees of freedom in the regression
`R-squared`	The coefficient of determination, i.e. r^2
`Adj R-squared`	The adjusted $\bar{r}^2$ for the degrees of freedom

The R-squared is a biased estimate of the population correlation, while the adjusted R-squared is always smaller than the non-adjusted R-squared. Note that a significant correlation does not necessarily imply the presence of a trend in time series data (see also Helsel and Hirsch, 2002, Chapter 9). This is better assessed by a nonparametric trend test, as described in Section 3.4.

Autocorrelation

This option determines the correlation for a number of time lags in the series, providing:

`Lag`	The time interval lag number, starting from zero
`ACF`	The autocorrelation function result
`CL`	The positive 95% confidence limit for a moving average (MA) process

The ACF result is the Pearson correlation coefficient (indicated by r) between the lagged series values. By definition, for lag zero $r(0) \equiv 1$.

Transform

This option enables to evaluate and set a small number of often-used transformations to make the data resemble a normal distribution. The Evaluate choice produces the results of the logarithm of the likelihood function (`LLF`) for the following transforms, indicated by their respective λ (`Lambda`) power coefficient (Box and Cox, 1964):

`None`	The data are not transformed ($\lambda = +1.0$)
`Sqrt y`	Transformation by the square root ($\lambda = +0.5$)
`Log y`	Base-e logarithmic transformation; the base does not really matter ($\lambda = 0$)
`1/Sqrt y`	Transformation by the reciprocal square root ($\lambda = -0.5$)
`1/y`	Transformation by the reciprocal value ($\lambda = -1.0$)

Evaluate only analyzes the data transforms. The actual transform can be applied from the choices in the sub-menu. The best choice is the transform which maximizes the LLF value. Whether or not the transformed data are actually normal needs to be tested (see Section 3.4).

Note that the data are transformed only once, no matter how often you choose the same transform option. Selecting None resets the transformed data to the original values of the data file. The Evaluate option uses the current series. When you change the time bounds, the active transform applies to the new time window but it may not necessarily be the best choice.

3.3 The Plot menu

Graphics output is displayed in a new window for each plot. Using a right mouse-click brings up a small popup menu with options to zoom, rescale axes, save and print the graphs. The left mouse-button can be used to set a zoom region. The mouse-wheel can also be used to zoom in or out.

Series

This option shows a run-sequence graph of the current series. Besides the data curve, additional graphical information includes a thin line indicating the mean value, which is located inside a shaded 95% confidence band to indicate the uncertainty interval of the mean. Two thin marker lines are drawn at ± 1 SD of the mean to provide an impression of the variation in the data points. Furthermore, the apparent trend in the series (as detailed in Section 3.2) is shown as a green line.

The Continuous line graph is usually better when the number of data values is large. A Stepped line graph may better visualize the data if there are less than 30 or 40 data values, or when the data represent time step totals.

Correlogram

Produces a plot of the autocorrelation coefficient values against the lag number as computed by the autocorrelation function (see Section 3.2). A shaded marker region shows the 95% confidence level band around zero for assessing the stationarity, or randomness, of the series.

Change point

Three options can be selected to visualize the location of probable change points in the form of a run-sequence graph.

[Document Buishand]

[Document SNHT]

The Pettitt option plot the probability for a change point versus time, based on the non-parametric method of Pettitt (1979). A 0.95 probability marker line is also drawn to aid in identifying possible change points. Points with probability values larger than 0.95 can be considered potential change points. Further verification is obtained from tests for stability of variance and/or mean.

Box plots

Displays a graph of three box plots (also referred to as box-and-whisker plots) providing visual information about the nature of the series and sub-series data distribution. The box plots are conventionally constructed (Tukey, 1977; McGill et al., 1978; Helsel and Hirsch, 2002). The medians are shown as a black line inside the boxes, which extend from the first to the third quartile (Q_1 to Q_3) showing the Interquartile Range ($IQR = Q_3 - Q_1$) containing 50% of the data. The lower whiskers extend to $Q_1 - 1.5\,IQR$, or to the lower extreme. The upper whiskers extend to $Q_3 + 1.5\,IQR$, or to the upper extreme. For normally distributed data, the whiskers plus the box show 99.3% of the data. Note that various alternatives are in use for defining the whisker extents. Some box plot representations also depict (outlier) points above or below the whisker lines as dots or circles. These points are not shown by DScreen.

3.4 The Test menu

The Test menu contains several methods for common time series analysis. For each option, the text output contains the timing details of the series or sub-series and a short description of the null hypothesis. This is followed by the test results as described below.

Normality

Carries out a Shapiro-Wilk test for normality (Shapiro and Wilk, 1965). The output contains two values, i.e.

`w`	The W-statistic value
`p`	The p-value

Normal distribution of the data is commonly required by parametric tests.

Autocorrelation

These options test aspects of the autocorrelation structure of the data series. The Persistence option tests for absence of serial correlation between consecutive values in the series, i.e. lag-1 autocorrelation. The output is arranged for interval comparison and consists of:

`r(1)`	The lag-1 autocorrelation coefficient
`UCL`	The upper confidence limit
`LCL`	The lower confidence limit

Four sets of critical level limits are arranged in tabular format for predefined significance levels (`Alpha`) of 0.1, 0.05, 0.02 and 0.01, respectively.

The Ljung-Box test is an overall or portmanteau test of independence, or randomness, between data values for a group of lags. More precisely, this option tests whether any autocorrelation coefficient from lag-1 up to the lag that equals the degrees of freedom is significantly different from zero. The test output includes:

`DF`	The degrees of freedom, number of lags
`Q`	The Lung-Box test statistic
`p`	The p-value

The degrees of freedom are taken as 15 for annual and 25 for monthly series, but never more than one-fourth of the number of data in the series.

If the Pre-white versions of the tests are selected the data are pre-whitened. This will effectively remove or substantially reduce any autocorrelation in the series for the majority of cases. Absence of persistence is a requirement for trend tests. The test results after pre-whitening are shown in a similar fashion as those of the natural data.

Secular trend

This item provides a choice between two nonparametric tests for long-term, or secular, trend. Spearman's rank-correlation method is claimed to have nearly equal power for both linear and nonlinear trends (WMO, 1966). The output of the Spearman option contains:

`S`	The (Spearman) S-statistic value
`rho`	The Spearman rank-correlation coefficient (ρ)
`DF`	The degrees of freedom (see below)
`t`	The t-statistic value (see below)
`p`	The p-value

Note that `DF` and `t` are *only* given when the data contain ties, which is also indicated by an additional output message (see Section 4.8 for details).

The Mann-Kendall method is considered even more robust and is among the most widely used test for detecting trends in environmental data (Helsel et al., 2006). The output of the Mann-Kendall option contains:

`S`	The (Kendall) S-statistic value
`tau`	The Kendall rank-correlation coefficient (τ)
`SD`	The standard deviation of S (see below)
`z`	The standardized normal deviate (z-score, see below)
`p`	The p-value

Note that `SD` and `z` are *only* given when the data contain ties, which is also indicated by an additional output message (see Section 4.8 for details).

In addition, both trend tests can be applied after pre-whitening the data to reduce autocorrelation. The test procudure and output presentation is the same as for using the natural data.

Stability

There are two options to analyze stability of variance and mean, respectively. These tests use the partial records as defined by the sub-series time bounds.

Stability of variance among the two sub-series is carried out as an F-test, which generates the following results:

`DF1`	The degrees of freedom for sub-series 1
`DF2`	The degrees of freedom for sub-series 2
`SD1`	The standard deviation of sub-series 1
`SD2`	The standard deviation of sub-series 2
`F`	The F-statistic value
`p`	The p-value

A t-test is used for assessing stability of the mean between the two sub-series. The output contains the following information:

`DF`	The degrees of freedom
`M1`	The mean of sub-series 1

`M2`	The mean of sub-series 2
`t`	The t-statistic value
`p`	The p-value

4 Technical Reference

4.1 General

The standard deviation SD is computed by a two-pass algorithm for the variance with compensation for numerical roundoff error (e.g. Press et al., 1992). The standard error and the coefficient of variation are defined by

$$\begin{aligned} \text{SE} &= \text{SD}/\sqrt{n} \\ \text{CV} &= \text{SD}/\text{Mean} \end{aligned}$$

where n is the number of data values. The standard error is used to compute confidence levels for the mean. For example, a 95% uncertainty interval is computed as the mean ± 1.96 SE.

Quantile values Q_p for $0 < p < 1$ are computed from the ascending-order sorted data values using

$$Q_p = y_{\lfloor k \rfloor} + (k - \lfloor k \rfloor)\left(y_{\lfloor k \rfloor + 1} - y_{\lfloor k \rfloor}\right)$$

where $\lfloor k \rfloor$ defines an index in the sorted data array from taking the floor (i.e., the lower integer) value of $k = np + \frac{1}{2}$. This is a long-established method in hydrology due to Hazen (1914) (implemented as type 5 method in the `quantile` function of R).

4.2 Apparent trend line computation

The computation is based on fitting a straight line through the data points. The equation of the trend line is

$$y = a + bx + \epsilon$$

where a and b are the regression constant and slope coefficients, respectively, x is the time axis coordinate, i.e. $x = t_i - t_1$, for $i = 1, \ldots, n$, and ϵ is an error term.

It is important to realize, however, that the equation should not be mistaken as a linear model to predict individual y-values from the points in time. Here, we try to show a potential long-term, gradual shift in the location (the mean) of the data series. In other words, an apparent trend in the y-values.

The equation is obtained from minimizing the sum of the squares of the vertical distances between the data points and the line (called residuals). The method is known as linear least squares fitting or linear regression. In the simple form of one independent variable x and one dependent variable y, we can define the following sums of squares:

$$\begin{aligned} S_{xx} &= \sum_{i=1}^{n} (x_i - \bar{x})^2 \\ S_{yy} &= \sum_{i=1}^{n} (y_i - \bar{y})^2 \\ S_{xy} &= \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) \end{aligned}$$

The regression coefficients are readily estimated from

$$\begin{aligned} b &= \frac{S_{xy}}{S_{xx}} \\ a &= \bar{y} - b\bar{x} \end{aligned}$$

The square of the correlation coefficient (the coefficient of determination) gives an overall quality of the fit. The r-squared is computed by

$$r^2 = \frac{S_{xy}^2}{S_{xx} S_{yy}}$$

which measures the proportion of the variation in S_{yy} that is explained by the regression.

Statistical packages and spreadsheet regression functions (such as LINEST in Excel) compute several other quantities to assess the quality. The degrees of freedom is equal to the number of data pairs minus the number of coefficients k (here, $k = 2$). Thus,

$$\mathrm{DF} = n - k$$

The standard error of the regression, i.e. the standard deviation of the mean of the residuals between the fitted and observed y's is

$$\mathrm{SE} = \sqrt{\frac{S_{yy} - bS_{xy}}{\mathrm{DF}}}$$

Similarly, the standard errors of the coefficients are

$$\begin{aligned} \mathrm{SE}(a) &= \mathrm{SE}\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}} \\ \mathrm{SE}(b) &= \frac{\mathrm{SE}}{\sqrt{S_{xx}}} \end{aligned}$$

Additionally, it is possible to define t-statistics that can be used in a test to determine the significance of the coefficients, viz.

$$\begin{aligned} t(a) &= a/\mathrm{SE}(a) \\ t(b) &= b/\mathrm{SE}(b) \end{aligned}$$

which can be compared against a t-distribution to obtain the p-values that denote the probability that a next trial yields a larger $|t|$. If the p-value is less than a predefined significance level (e.g. 0.05), the coefficient is significantly different from zero, and therefore cannot be neglected.

In a generic least squares regression, an adjusted $\bar{r}^2$ can be used to penalize for excess coefficients that do not add explanatory power to the regression, viz.

$$\bar{r}^2 = 1 - \frac{n-1}{n-k}(1 - r^2)$$

The $\bar{r}^2$ is always less than the r^2, and may become negative for very poor correlations.

4.3 Autocorrelation

The autocorrelation coefficient is computed as the ratio $r_h = c_h/c_0$, where c_h is the autocovariance at lag h, viz.

$$c_h = \frac{1}{n}\sum_{i+1}^{n-h}(y_i - \bar{y})(y_{i+h} - \bar{y})$$

and c_0 is the variance of the data, obtained by setting $h = 0$ (Box et al., 2008; Hipel and McLeod, 1994). The maximum number of lags used for DScreen output is determined as $10 \log_{10} n$ (cf. the relevant functions in R). The confidence bounds on the autocorrelation for a 95% significance are computed by $\pm 1.96\,\mathrm{SE}(r_h)$, where the standard error is obtained from

$$\mathrm{SE}(r_h) = \begin{cases} \dfrac{1}{\sqrt{n}} & \text{for } h = 0 \\ \left[\dfrac{1 + 2\sum_{i=1}^{h} r_i^2}{n}\right]^{1/2} & \text{for } h > 0 \end{cases}$$

The equations compute the standard error for a moving average (MA) process. When testing for randomness, i.e. the absence of persistence, the 95% confidence levels are given by $\pm 1.96/\sqrt{n}$. These levels are also displayed in the correlogram plot (see Section 3.3).

4.4 Data transformation

A small set of typically applied transformations in time series analysis are the following functions

$$z = \sqrt{y}, \quad z = \log y, \quad z = 1/\sqrt{y}, \quad z = 1/y$$

where y are the data values and z are the transformed result. A generalized scheme for transforming data was presented by Box and Cox (1964) as

$$z = \begin{cases} \dfrac{y^\lambda - 1}{\lambda} & \text{for } \lambda \neq 0 \\ \ln y & \text{for } \lambda = 0 \end{cases}$$

where λ is a real-valued coefficient that takes values from a small range around zero. For the predefined transforms above $\lambda = +0.5, 0, -0.5, -1$, respectively. The case for $\lambda = +1$ is treated as no transformation. For $\lambda = 0$, DScreen uses the natural logarithm, the base of the logarithm does not change the character of the data. When the data contain zero or negative values, a constant is added to all values in the untransformed series before applying the function. The logarithm of the likelihood function (LLF) is computed as

$$f(y, \lambda) = -\frac{n}{2} \ln \sum_{i=1}^{n} \frac{(z_i - \bar{z})^2}{n} + (\lambda - 1) \sum_{i=1}^{n} \ln y_i$$

where n is the number of data values.

4.5 Test for normality

The Shapiro-Wilk test (Shapiro and Wilk, 1965) is an established and powerful method to assess the null hypothesis that a sample is drawn from a normal population. First, the data values y are sorted in ascending order. The ith smallest value in the data is called the ith order statistic, denoted by $y_{(i)}$. The test statistic is computed as

$$W = \frac{\left(\sum_{i=1}^{n} a_i y_{(i)}\right)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$$

with the coefficients a_i given by

$$(a_1, \ldots, a_n) = \frac{m^\top \mathbf{V}^{-1}}{(m^\top \mathbf{V}^{-1} \mathbf{V}^{-1} m)^{1/2}}$$

where $m = (m_1, \ldots, m_n)^\top$ is the transposed vector of the expected values of the order statistics of identically distributed random variables sampled from a normal distribution and $\mathbf{V}$ is the covariance matrix of the order statistics.

The p-value for W is obtained from a normal distribution, and can be compared against the chosen significance level. The DScreen implementation of the Shapiro-Wilk test is adopted from algorithm R94 (Royston, 1995).

4.6 Autocorrelation tests

The test for absence of persistence proceeds from comparison of the lag-1 autocorrelation coefficient r_1, described in Section 4.3, against the critical region defined by the confidence limits given by

$$\mathrm{CL} = \pm z_{1-\alpha/2} \frac{1}{\sqrt{n}}$$

where $z_{1-\alpha/2}$ is the normal distribution quantile for significance level α. For two-tailed confidence levels $z_{1-\alpha/2} = \{1.645, 1.960, 2.326, 2.576\}$ for $\alpha = \{0.1, 0.05, 0.02, 0.01\}$, yielding confidence intervals of 90, 95, 98 and 99%, respectively.

The null hypothesis that there is no persistence, essentially $H_0 : r_1 = 0$, is accepted if the relation

$$-z_{1-\alpha/2} \frac{1}{\sqrt{n}} < r_1 < +z_{1-\alpha/2} \frac{1}{\sqrt{n}}$$

is true. The $\pm z_{1-0.05/2} = 1.96$ yields the 95% confidence interval that is visualized in the correlogram plot output (Section 3.3)[1].

The Ljung-Box portmanteau test can be used to accept or reject the null hypothesis that the data are independent (or random), i.e. that none of the correlation coefficients within a region of L lags are significantly different from zero. The test statistic is computed as

$$Q = n(n+2) \sum_{k=1}^{L} r_k^2/(n-k)$$

The Q-statistic is χ^2 distributed on L degrees of freedom, from which the p-value is obtained. Following the recommendations by Hipel and McLeod (1994), L ranges between 15 (annual data) and 25 (monthly data), but is capped at $n/4$.

Within the present scope of time series analysis testing (lag-1) persistence is commonly sufficient. Most often, the Ljung-Box test is applied to residuals in time series modelling applications. It may however be useful in quality control applications when screening data for relative consistency using difference series between stations. In such cases, the analist wants to assert that the residuals between station and reference series are random.

4.7 Pre-whitening

Autocorrelation in the data can be substantially reduced by pre-whitening. This is carried out by a simple recursive process removing lag-1 autocorrelation from consecutive observations, viz.

$$y'_i = \begin{cases} (1 - r_1) y_i & \text{for } i = 1 \\ y_i - r_1 y_{i-1} & \text{for } i = 2, \ldots, n \end{cases}$$

[1] Note that the computation of the confidence interval differs from the equations by Anderson (1942, cited in Dahmen and Hall, 1990, Eqn (4.2)), which produce asymmetric upper and lower bounds. The present calculation is in agreement with the modern literature.

where r_1 is the lag-1 autocorrelation coefficient of the natural, stationary data series. The pre-whitening process results in a non-stationary series y' with otherwise the same statistical properties as the natural data series y.

4.8 Nonparametric trend tests

Two established methods can be used to test if there is a *monotonic*, that is a continously increasing or decreasing, relation between the time and observations. Both methods are nonparametric, because only the rank-order of the variables is taken into account, not their values. These tests do not depend on the statististical distribution of the data and are therefore resistant against possible outliers in the data (see for details e.g. Helsel and Hirsch, 2002; Hipel and McLeod, 1994). However, the data are required to be stationary, that is, autocorrelation—or persistence—must be absent or should be removed by pre-whitening. The null hypothesis for the trend tests is stated as H_0 : there is no trend.

Spearman trend test This test is uses Spearman's rank-order correlation coefficient ρ, which similarly to Pearson's r^2 varies between ± 1. The ρ is based on sorting the data in ascending order. The smallest value has rank 1, the largest value has rank n. In our case, the time is already ranked from $r(x = t_1) = 1$ to $r(x = t_n) = n$, so we only need to sort the observations for ranks $r(y) = 1, \ldots, n$. The Spearman rank correlation coefficient is defined as

$$\rho = 1 - \frac{6S}{n(n^2 - 1)}$$

where the Spearman S-statistic is given by the sum of the squared rank differences, i.e.

$$S = \sum_{i=1}^{n} [r(x)_i - r(y)_i]^2$$

Note that S is a positively-valued integer. In fact, the ρ is a special case of the Pearson correlation coefficient (explained in Section 4.2), which is actually a more convenient way used by DScreen to compute ρ. The p-values for Spearman's ρ are computed to a high degree of accuracy from S, using the algorithm of Best and Robberts (1975).

In the case that *ties* are present, that is two or more equal y-values, a tied rank is assigned to these data points from averaging the rank numbers of the ties in ascending order. As a consequence, the S-statistic becomes approximate and will take a decimal value. The test is then carried out using a t-statistic, which is computed from ρ by

$$t = \frac{\rho\sqrt{n-2}}{\sqrt{1-\rho^2}}$$

The t-statistic approximately follows a Student t-distribution with $n - 2$ degrees of freedom, from which the p-value is computed.

Mann-Kendall test Another rank-based measure of the association between two variables is Kendall's tau, which also varies between ± 1 and is particularly applicable for data exhibiting strong skewness (Helsel and Hirsch, 2002). Kendall's tau is defined as

$$\tau = \frac{S}{n(n-1)/2}$$

Here, S denotes the Kendall S-statistic obtained from comparison between all possible (x,y) *pairs* of data. The denominator $n(n-1)/2$ equals the total number of pairs for a given series length n. Kendall's S-statistic is defined as

$$S = n_c - n_d$$

where n_c is the number of concordant pairs and n_d is the number of discordant pairs. For all $i = 1, \ldots, n-1$ and $j = i+1, \ldots, n$, a pair is concordant if $y_i < y_j$ for $x_i < x_j$ or $y_i > y_j$ for $x_i > x_j$, and a pair is discordant if $y_i > y_j$ for $x_i < x_j$ or $y_i < y_j$ for $x_i > x_j$. A direct computation to obtain S, often seen in the literature (e.g. Gilbert, 1987; Hipel and McLeod, 1994, and references therein), is the following scheme

$$S = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \operatorname{sgn}(x_i - x_j) \operatorname{sgn}(y_i - y_j)$$

where the sign function is defined (for any variable u) as

$$\operatorname{sgn}(u) = \begin{cases} -1 & \text{if } u < 0 \\ 0 & \text{if } u = 0 \\ +1 & \text{if } u > 0 \end{cases}$$

It is to be noted that the Kendall S is always integer-valued and can be either positive, negative, or zero. For the case of a time series in which where no ties occur, a p-value can be accurately computed using the algorithm of Best and Gipps (1974).

If ties are present in the data, a correction must be made. The Kendall tau is then computed as

$$\tau = \frac{S}{\sqrt{n_x}\sqrt{n_y}}$$

with

$$n_{(x,y)} = n(n-1)/2 - \sum_{i=1}^{n} t_{i(x,y)} i(i-1)$$

where t_i is the number of ties of extent i in either the x or y data. For time series, ties can be present only in the y data. The p-value is now estimated from a normal distribution, for which a standardized deviate, or z-score, is obtained as

$$z = \begin{cases} \dfrac{S+1}{\sqrt{\operatorname{Var}(S)}} & \text{if } S < 0 \\ 0 & \text{if } S = 0 \\ \dfrac{S-1}{\sqrt{\operatorname{Var}(S)}} & \text{if } S > 0 \end{cases}$$

where the tie-corrected variance in S is computed from

$$\operatorname{Var}(S) = \frac{n(n-1)(2n+5) - \sum_{i=1}^{n} t_i\, i(i-1)(2i+5)}{18}$$

(Helsel and Hirsch, 2002; Hipel and McLeod, 1994).

4.9 Tests for stability of variance and mean

The tests for stability assess the equality of variance and mean between split-samples (i.e. sub-series). Strictly speaking, the sub-series should be approximately equal in length and the time periods should not overlap (Dahmen and Hall, 1990). Variance is tested using an F-test, mean is tested using a t-test. If the variance is not stable, that is the sub-series have a significantly different variance, testing the mean is no longer meaningful (Dahmen and Hall, 1990). Both tests require that the sub-series data values resemble a normal distribution.

Stability of variance The null hypothesis is commonly formulated as H_0 : the variances of the sub-series are equal. The F-statistic for the test is given by the ratio of of the sub-series variances, denoted by s_1^2 and s_2^2, respectively, viz.

$$F = \frac{s_1^2}{s_2^2}$$

Defining the degrees of freedom in each sub-series as $\mathrm{DF}_1 = n_1 - 1$ and $\mathrm{DF}_2 = n_2 - 1$, the p-value is obtained from an F-distribution. Idiosyncratic cases in the F-test of variance as described in Dahmen and Hall (1990, pp21-22) are handled properly, with results matching those of the R `var.test` function result.

Stability of mean The test for the mean is meaningful only if the null hypothesis for the variance is accepted. In that case, the t-statistic to test the stability of the mean, under the null hypothesis H_0 : the means are equal, is obtained as

$$t = (\bar{y}_1 - \bar{y}_2) \left[\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} \left(\frac{1}{n_1} + \frac{1}{n_2} \right) \right]^{-1/2}$$

where the $\bar{y}_i$ are the sub-series mean values and the n_i are the number of data points in the sub-series. The sub-series degrees of freedom are $\mathrm{DF}_1 = n_1 - 1$ and $\mathrm{DF}_2 = n_2 - 1$, respectively. The t-test proceeds by calculating the p-value from a Student t-distribution.

The F and t test implementations in DScreen follow a common approach where the p-values are obtained from an incomplete beta function, which based on the algorithms described in Press et al. (1992).

4.10 Change point detection

In order to guide the selection of the series and sub-series time bounds, there are several change point detection menthods. All methods use a graphical representation of the test results, indicating whether or not there is a probable change point.

[Document Buishand]

[Document SNHT]

The non-parametric detection method by Pettitt (1979) is often useful. For every point in time of the series, a probability that it constitutes a change-point is calculated by comparing the distribution of the data values before and after that time. This results in a new time series of probabilities that has the same length as the source data series. Formally, the method is a test for the null hypothesis H_0 : no change occurs anywhere in the series, against the alternative hypothesis H_1 : a change occurs somewhere in the series. The result is best represented graphically for visual interpretation.

A statistic U_t is computed for each time point $1 \leq t \leq n$ from the differences in value between all possible pairs as

$$U_t = \sum_{i=1}^{t} \sum_{j=t+1}^{n} \operatorname{sgn}(y_i - y_j)$$

where

$$\operatorname{sgn}(y_i - y_j) = \begin{cases} -1 & \text{if } y_i < y_j \\ 0 & \text{if } y_i = y_j \\ +1 & \text{if } y_i > y_j \end{cases}$$

To find out if there is a change-point, an overall p-value for the whole series is obtained from

$$p = \exp \frac{-6(\max |U_t|)^2}{n^2 + n^3}$$

In practice, it is convenient to evaluate the probability for each point being a change-point as

$$\Pr(t_{\text{cp}}) = 1 - \exp \frac{-6U_t^2}{n^2 + n^3}$$

and plot the results in a run-sequence graph for visual comparison against a predefined confidence level, e.g. 95%.

5 Example Data Files

5.1 Annual series

`Bangkok_ann.txt` and `Bangkok_ann-tie.txt` Annual rainfall (mm) for 1952–1985 in Bangkok collected by the Meteorological Department, Thailand (source: Dahmen and Hall, 1990). The second file contains an artificial tie by adding 1 mm to the value for 1980.

`Colorado_ann.txt` Annual maximum discharge ($\mathrm{m^3\,s^{-1}}$) of the Colorado River at Black Canyon, USA, for 1878–1929 (source: Gumbel, 1954).

`Derwent_ann.txt` Annual mean discharge ($\mathrm{m^3\,s^{-1}}$) of the Derwent River at the Yorkshire Bridge for 1906–1961. Catchment area 126 $\mathrm{km^2}$. Data collected by the National Rivers Authority, U.K. (source: Dahmen and Hall, 1990).

`Global_temp_ann.txt` Global average annual mean land surface temperature (°C) compiled from station observations for 1880–2009. Data obtained from NASA Goddard Institute of Space Studies, `http://data.giss.nasa.gov/gistemp`.

`Green_sl_ann.txt` Annual suspended sediment load (tons) in Green River, Kentucky, USA, for 1952–1972 (source: Helsel et al., 2006, example MK4b).

`Pahang_ann.txt` Annual mean discharge ($\mathrm{m^3\,s^{-1}}$) of Pahang River at Lubok Paku, Malaysia for 1988–2007. Basin area 25,600 $\mathrm{km^2}$. Source: Department of Irrigation and Drainage Malaysia. Parts of the original records have been pre-processed by seasonal mean infilling to handle missing data.

`Problem_ann.txt` Annual precipitation (mm) at a problem station. Location undisclosed, time period assigned arbitrary (source: Dahmen and Hall, 1990).

`YR-TNH_ann.txt` Annual mean discharge ($\mathrm{m^3\,s^{-1}}$) of the upper Yellow River at Tangnaihai, China for 1991–2005. Basin area 125,000 $\mathrm{km^2}$. Source: Yellow River Conservancy Commission, Ministry of Water Resources of the People's Republic of China.

5.2 Monthly series

`Nile_mon.txt` Monthly discharge ($m^3\,s^{-1}$) of the Nile River at Aswan, Egypt, from May 1870 until April 1984. Basin area $3.3 \times 10^6\,km^2$. The value for April 1892 has been manually corrected (data contributed by E.R. Dahmen).

`RPB_CO2_mon.txt` Monthly CO_2 concentrations (ppm) collected at Ragged Point, Barbados, from January 1988 until December 2007 (Conway et al., 2010). Data obtained from NOAA Earth System Research Laboratory, `http://www.esrl.noaa.gov`.

6 Data File Format

The time series data for DScreen are ASCII text files that can be read and produced by spreadsheets, text editors or hydrological and statistical analysis packages. The formatting is such that the files can also be used as data frames in R and S-plus.

A hash character ('`#`') is used to indicate the start of a comment up to the end of a line. Comments can (actually, must) be used to provide information about the station location, type of data, units of measurement, etc. They can also serve to note specific details or identify corrections for individual data points.

The data are entered as time-value pairs, one pair per line. The first column gives the time, the second the value. The data series must be presented in chronological order with equal time interval spacing. Missing data are not allowed in DScreen.

The time is specified following ISO 8601 format (ISO, 2004), either as a year only or as year and month, i.e. `yyyy` or `yyyy-mm` (without whitespace), where yyyy is the four-digit calendar year and mm is the two-digit calendar month (01 to 12). In cases where the times are unknown, which does not really make sense, arbitrary "years" (e.g. starting from 1001) can be used.

The data values can be supplied in integer, decimal or scientific (e.g. 1.23E+4 or 32.1e-3) notation. Time and values are separated by one or more space or tab characters (no commas or other punctuation characters).

References

Alexandersson, H. (1986) A homogeneity test applied to precipitation data, *International Journal of Climatology*, 6:661–675.

Best, D.J., and Gipps, P.G. (1974) Algorithm AS 71: The upper tail probabilities of Kendall's tau, *Applied Statistics*, 23:98–100.

Best, D.J., and Roberts, D.E. (1975) Algorithm AS 89: The upper tail probabilities of Spearman's rho, *Applied Statistics*, 24:377–379.

Box, G.E.P., and Cox, G.M. (1964) An analysis of transformations, *Journal of the Royal Statistical Society*, B26:211–252.

Box, G.E.P., Jenkins, G.M., and Reinsel, G.C. (2008) Time Series Analysis: Forecasting and Control, 4th ed., John Wiley & Sons, Hoboken, NJ.

Buishand, T.A. (1982) Some methods for testing the homogeneity of rainfall records, *Journal of Hydrology*, 58:11–27.

Cleveland, W.S. (1992) The Elements of Graphing Data, Rev. Ed., AT&T Bell Laboratories, Murray Hill, NJ.

Conway, T.J., P.M. Lang, and K.A. Masarie (2010), Atmospheric Carbon Dioxide Dry Air Mole Fractions from the NOAA ESRL Carbon Cycle Cooperative Global Air Sampling Network, 1968-2009, Version: 2010-09-08, Path: `ftp://ftp.cmdl.noaa.gov/ccg/co2/flask/month/`.

Dahmen, E.R., and Hall, M.J. (1990) Screening of Hydrological Data: Tests for Stationarity and Relative Consistency, ILRI Publication 49, International Institute for Land Reclamation and Improvement, Wageningen, Netherlands. Available online at `http://www.alterra.wur.nl/UK/publications/ILRI-publications/Downloadable/`.

Gilbert, R.O. (1987) Statistical Methods for Environmental Pollution Monitoring, Van Nostrand Reinhold, New York, NY.

Gumbel, E.J. (1954) Statistical Theory of Extreme Values and Some Practical Applications: A Series of Lectures, U.S. National Bureau of Standards, Applied Mathematics Series, 33, U.S. Govt. Printing Office, Washington DC.

Hazen, A. (1914) Discussion on: Flood flows by W.E. Fuller, *Transactions of the American Society of Civil Engineers*, 77:628.

Helsel, D.R., and R. M. Hirsch (2002) Statistical Methods in Water Resources, Techniques of Water Resources Investigations, Book 4, chapter A3. U.S. Geological Survey. Available online at `http://pubs.usgs.gov/twri/twri4a3/`.

Helsel, D.R., Mueller, D.K., and Slack, J.R. (2006) Computer Program for the Kendall Family of Trend Tests, U.S. Geological Survey Scientific Investigations Report 2005-5275, Available online at `http://www.usgs.gov/pubprod`.

Hipel, K.W., and McLeod, A.I. (1994) Time Series Modelling of Water Resources and Environmental Systems, Elsevier, Amserdam.

Hyndman, R.J., and Fan, Y. (1996) Sample quartiles in statistical packages, *American Statistician*, 50:361–365.

ISO (2004) International Standard ISO 8601. Data Elements and Interchange Formats – Information Interchange – Representation of Dates and Times, 3rd ed., ISO 8601:2004(E), International Organization for Standardization, Geneva.

McGill, R., Tukey, J.W., and Larsen, W.A. (1978) Variations of box plots, *The American Statistician*, 32:12-16.

Pettitt, A.N. (1997) A non-parametric approach to the change-point problem, *Applied Statistics*, C28:126–135.

Press, W.H., Teukolsky, S.A., Vetterling, W.T., and Flannery, B.P. (1992) Numerical Recipes in C. The Art of Scientific Computing, 2nd ed., Cambridge Univ. Press.

R Development Core Team (2006), R: A language and environment for statistical computing, R Foundation for Statistical Computing, Vienna, Austria (URL `http://www.R-project.org`).

Royston, P. (1995) Remark AS R94. A remark on algorithm AS 181: The W-test for normality, *Applied Statistics*, 44:547–551.

Shapiro, S.S., and Wilk, M.B. (1965) An analysis of variance test for normality (complete samples), *Biometrika*, 52:591–611.

Sheskin, D.J. (2007) Handbook of Parametric and Nonparametric Statistical Procedures, 4th ed., Chapman & Hall/CRC, Boca Raton, FL.

Tukey, J.W. (1977) Exploratory Data Analysis, Addison-Wesley, Reading, Mass.

WMO (1966) Climatic Change, Edited by W.J. Mitchell, WMO No. 195 TP 100, World Meteorological Organization, Geneva.

Wijngaard, J.B., Klein Tank, A.M.G., and Können, G.P. (2003) Homogeneity of 20th century European daily temperature and precipitation series, *International Journal of Climatology*, 23:679–692.

A Notices

The following notice applies to the DScreen program distribution

```
Copyright (C) 2010,2011 by UNESCO-IHE Institute for Water Education.

This program is free software: you can redistribute it and/or modify
it under the terms of the GNU General Public License as published by
the Free Software Foundation, either version 3 of the License, or
(at your option) any later version.

This program is distributed in the hope that it will be useful,
but WITHOUT ANY WARRANTY; without even the implied warranty of
MERCHANTABILITY or FITNESS FOR A PARTICULAR PURPOSE.  See the
GNU General Public License for more details.

You should have received a copy of the GNU General Public License
along with this program.  If not, see <http://www.gnu.org/licenses/>.
```

The following notice applies to third party software contained in the DScreen distribution:

```
This software incorporates JFreeChart, (C)opyright 2000-2009 by
Object Refinery Limited and Contributors.

This library is free software; you can redistribute it and/or
modify it under the terms of the GNU Lesser General Public
License as published by the Free Software Foundation; either
version 2.1 of the License, or (at your option) any later version.

This library is distributed in the hope that it will be useful,
but WITHOUT ANY WARRANTY; without even the implied warranty of
MERCHANTABILITY or FITNESS FOR A PARTICULAR PURPOSE.  See the GNU
Lesser General Public License for more details.

You should have received a copy of the GNU Lesser General Public
License along with this library; if not, write to the Free Software
Foundation, Inc., 59 Temple Place, Suite 330, Boston, MA  02111-1307  USA
```

The JFreeChart source code and class library are available from `http://www.jfree.org`.

The implementations of algorithms AS 89, R94 and dependencies are based on transliterations of FORTRAN source code obtained from the Statlib—Applied Statistics Algorithms repository at `http://lib.stat.cmu.edu/apstat/`.