XTREND:
A computer program for estimating trends in the occurrence rate of extreme weather and climate events

Manfred Mudelsee

LIM—Institute of Meteorology

University of Leipzig

Stephanstr. 3

D–04103 Leipzig

FRG

Email: mudelsee@rz.uni-leipzig.de

URL: http://www.uni-leipzig.de/~meteo/MUDELSEE

Preface

In the years following the Contribution of Working Group I to the Second Assessment Report of the Intergovernmental Panel on Climate Change (Houghton *et al.* 1996) speculations grew that the occurrence rate of extreme weather and climate events such as, for example, landfalling hurricanes or floods, may have changed in the past decades, and that this change be related to a change of global climate (e. g., Olsen *et al.* 1998, Easterling *et al.* 2000). It is further currently being realized that the nonstationary character of a climate change has to be considered when analysing extreme events and their probability density functions (PDFs) in meteorology and climatology (Olsen *et al.* 1998). Finally, it seems that the understanding is growing that, in a world of limited climate data and coarse climate models, estimated quantities without error bars or confidence intervals have little use for assessing results and evaluating consequences (Allen *et al.* 2000).

This means that, in principle, we have to estimate time-dependent PDFs when analysing extreme-climate data. Whereas this is less problematic when using climate model output, it is formidably difficult for observed data since then we cannot repeat experiments (von Storch and Zwiers 1999). The computer program documented here, XTREND, estimates only the time-dependent *occurrence rate* of extreme events for observed time series (t, x), treating the data as realizations of a non-homogeneous Poisson process. Whereas this recognizes the t-dependence, it is a simplified approach because the x-information is drastically reduced (magnitude classes for x). However, as a positive aside, also qualitative x-data, as they are often found in written documents, can be analysed, and even data without x-information. That is, XTREND is a robust method as regards the distribution of data, x.

XTREND consists of the following methodical Parts. Time interval extraction (Part 1) to analyse different parts of a time series; extreme events detection (Part 2) with robust smoothing; magnitude classification (Part 3) by hand; occurrence rate estimation (Part 4) with kernel functions; bootstrap

simulations (Part 5) to estimate confidence bands around the occurrence rate. You work interactively with XTREND (parameter adjustment, calculation, graphics) to acquire more intuition for your data. Although, using "normal" data sizes (less than, say, 1000) and modern machines, the computing time seems to be acceptable (less than a few minutes), parameter adjustment should be done carefully to avoid spurious results or, on the other hand, too long computing times. This Report helps you to achieve that. Although it explains the statistical concepts used, this is generally done with less detail, and you should consult the given references (which include some textbooks) for a deeper understanding.

The structure of the Report follows closely that of XTREND. A Glossary which includes the used variable notation is supplied after the Table of Contents. One point is emphasized already: In statistics, the term "occurrence rate" is identical to "intensity" of a Poisson process whereas in other sciences the meaning of "intensity" varies. XTREND and this Report follow statistics in usage. The strength of an event (value x) is denoted as "magnitude".

Obviously, XTREND can also be used to analyse data from fields such as Econometrics or Actuarial Sciences etc. Welcome to readers from other fields!

Comments by you on XTREND, reports about bugs etc. are welcome! If you wish to abtain a copy of XTREND please contact me via email, providing information about your machine and system.

Leipzig, December 2001, Manfred Mudelsee

Contents

Glossary and notation

BP	before present
CV	cross validation
H0	null hypothesis
H1	alternative hypothesis
MAD	median of absolute distances to median

Notation

$t_{\mathrm{obs-l}}$	observation interval, left bound
$t_{\mathrm{obs-r}}$	observation interval, right bound
$t1$	time, original data
$x1$	time series value, original data
$d1$	segment duration, original data (**segmented**)
$n1$	number of points
$t2$	time, interval extracted data
$x2$	time series value, interval extracted data (not **segmented**)
$x2\mathrm{u}$	untransformed time series value, interval extracted data (**segmented**)
$x2$	transformed time series value, interval extracted data (**segmented**)
$d2$	segment duration, interval extracted data (**segmented**)
$b2$	background value, interval extracted data (not **segmented**)
$b2\mathrm{u}$	background value, interval extracted and untransformed data (**segmented**)
$b2$	background value, interval extracted and transformed data (**segmented**)
$b_{k,-1}$	delete-one background value
$v2$	variability, interval extracted data (not **segmented**)
$v2\mathrm{u}$	variability, interval extracted and untransformed data (**segmented**)
$v2$	variability, interval extracted and transformed data (**segmented**)
$n2$	number of points
$t3$	time, extreme data
$x3$	time series value, extreme data
$n3$	number of points
$t4$	time, classified extreme data, without pseudodata
$n4$	number of points
t	time, classified extreme data, with pseudodata
n	number of points
t^*	simulated t

alpha, α	confidence level
c	magnitude class bound
$CV_1(k)$	CV function (L_1-norm), running median smoothing
$CV_2(k)$	CV function (L_2-norm), running median smoothing
$CV_\mathrm{m}(k)$	CV function (median criterion), running median smoothing
$CV(h)$	CV function, kernel bandwidth selection
ϵ	tiny value
E	average
h	kernel bandwidth
h_relmax	determines right search bound for h
h_CV	minimizes $CV(h)$
k	smoothing parameter, running median smoothing of $x2$
k_CV1	minimizes $CV_1(k)$
k_CV2	minimizes $CV_2(k)$
k_CVm	minimizes $CV_\mathrm{m}(k)$
k_rel	determines right search bound for k and k_u
k_u	smoothing parameter, running median smoothing of $x2_\mathrm{u}$
$k_\mathrm{u-CV1}$	minimizes $CV_1(k_u)$
$k_\mathrm{u-CV2}$	minimizes $CV_2(k_u)$
$k_\mathrm{u-CVm}$	minimizes $CV_\mathrm{m}(k_u)$
K_h	kernel function
l	number of magnitude classes
l_x	design points (time), estimated occurrence rate
l_y	estimated occurrence rate
$l_{y_{h,i}}$	leave-one-out occurrence rate
l_y^*	simulated l_y
$l_{y\mathrm{l}}$	lower bound, estimated occurrence rate
$l_{y\mathrm{u}}$	upper bound, estimated occurrence rate
n_grid	number of design points l_x
n_hsrch	number of search values (h)
n_min	minimum number of points
n_sim	number of bootstrap simulations
p	one-sided probability (H0 test)
$r1, r2$	probabilities
rule	pseudodata generation rule
T	studentized quantity
t_α	percentile
u	test statistic (H0 test)
z	threshold parameter (magnitude classification)
z_test	test values for z

Part 0:
Starting XTREND

XTREND is written in Fortran 90 and uses a precision similar to "double precision" of FORTRAN 77.

0.1   System requirements

- $\gtrsim$ 400 MHz processor

- a few MB free disk space

- several MB RAM (variable space in XTREND is mostly allocated dynamically), a rough formula is:

$$\mathrm{RAM} = 16\,\mathrm{B}\,[24 \cdot n1 \,+\, n_{\mathrm{grid}} \cdot (l \cdot n_{\mathrm{grid}} \,+\, n_{\mathrm{sim}}) \,+\, 5 \cdot n_{\mathrm{hsrch}}]$$

  (see Glossary for explanation of variables)

- DOS window under Windows 95, Windows 98 or Windows NT (other operating systems not tested)

0.2   Installation

Copy files `XTREND.EXE`, `XTREND.CFG` and `CVMED.PLT` to a directory of your choice. XTREND requires that the freeware graphics program Gnuplot (version 3.6 or higher) is installed on your computer with the path name set accordingly. (If you intent to use Gnuplot solely with XTREND, you may simply install the Gnuplot files in XTREND's directory. Gnuplot executable name has to be `gnuplot.exe`. You may obtain a copy of Gnuplot from the author's URL.)

0.3   Configuration file

With XTREND.CFG you pass parameter values that are not being altered during running. At first, use the typical values given in the comment lines of XTREND.CFG. How to change them is discussed in the following Parts.

This is a typical example of the content of XTREND.CFG (see Glossary for explanation of variables):

```
&cfg
        krel=0.30,
        rule='threepoint',
        hrelmax=1.0,
        nhsrch=400,
        ngrid=1024,
        alpha=0.900
/


! Typical values:
!       krel=0.25,
!       rule='twopoint  ' / 'threepoint' / 'reflection'
!       hrelmax=0.5,
!       nhsrch=400,
!       ngrid=1024,
!       alpha=0.80, 0.90, 0.95, 0.99
```

0.4   Running

By the command XTREND at the DOS prompt.

0.5   Datatype

Your data have to be in ASCII (text) format. Since information about extreme weather or climate events may be of different types, XTREND allows four datatypes:

- **ordinary**

Part 0: Starting XTREND

- **ordinary and segmented**, also called **segmented** in XTREND output

- **extreme**

- **extreme times**, also called **times** in XTREND output

0.5.1 Ordinary

An **ordinary** time series,

$$
\begin{array}{cc}
t1(1) & x1(1) \\
t1(2) & x1(2) \\
\vdots & \vdots \\
t1(n1) & x1(n1),
\end{array}
$$

is assumed to consist of not only extreme but also some background values. Example 1 is an **ordinary** time series. Background estimation and extreme events detection is carried out in Part 2.

0.5.2 Segmented

An **ordinary and segmented**, in short **segmented**, time series,

$$
\begin{array}{ccc}
t1(1) & x1(1) & d1(1) \\
t1(2) & x1(2) & d1(2) \\
\vdots & \vdots & \vdots \\
t1(n1) & x1(n1) & d1(n1),
\end{array}
$$

means that your data have not the character of point values but rather represent a segment of an archive, for example, an ice core. The larger the duration, $d1(i)$, is in relation to the spacing, $t(i)-t(i-1)$, the more important is that you supply $d1$, in particular if $d1(i)$ is not constant. See Example 2.

0.5.3 Extreme

An **extreme** time series,

$$
\begin{array}{cc}
t1(1) & x1(1) \\
t1(2) & x1(2) \\
\vdots & \vdots \\
t1(n1) & x1(n1),
\end{array}
$$

already consists of only extreme values. It might, for example, be the result of a previous data analysis using methods of Extreme Value Theory (e. g., Embrechts *et al.* 1997, Reiss and Thomas 1997).

0.5.4 Times

An **extreme times**, in short **times**, time series,

$$t1(1)$$
$$t1(2)$$
$$\vdots$$
$$t1(n1),$$

consists of only the times of extreme events. This form is used when $x1$ is unknown or no quantitative value may be assigned to an event.

0.6 Number of points

XTREND requires at least $n_{\min} = 25$ data points. With less data your results would likely be without significance.

0.7 Observation interval

The interval over which the observations were made is an important point without which your data give no complete information! For example, you may have started in AD 1871 with observing major volcanic eruptions, but the first recording is from AD 1883 (Krakatau). The observation time interval, $[t_{\mathrm{obs-l}}; t_{\mathrm{obs-r}}]$, influences the generation of pseudodata (Part 4) and, hence, the estimated occurrence rate at the upper and lower boundaries of your observed data. If the observation interval is unknown, you should experiment with reasonably chosen values and try different methods of pseudodata generation; eventually a smaller kernel bandwidth for occurrence rate estimation (Part 4) should be selected since then the boundary effect is smaller.

0.8 Numerical values of input data

Too big or too small data values might cause problems in numerical calculations. In such cases XTREND asks you to rescale accordingly. XTREND

further requires that $t1$ increases strictly monotonically—whether it means time or geological age.

0.9   Interactive working

After you have supplied datatype, path and file name of data, $n1$ and the observation interval, XTREND plots the timeseries on the screen. (In case of **segmented**, also $d1$ is plotted.) Proceeding with a keystroke, you get information field 1 and Part 1 decision tree. This illustrates how working is done with XTREND: information fields and graphics help you to adjust parameters for an effectice calculation and to carry out the analyses of your data.

0.10   Notes

0.10.1   Gaps

Obviously, gaps (time intervals without observation) in your data cause problems for interpreting results. You might wish to split your data and analyse the parts separately. In case of small gaps, it might work to analyse the data as follows: stack the observed intervals together (i. e., time shiftings), analyse the stack with XTREND, reverse the time shiftings (i. e., gaps re-appear) in the estimated occurrence rate and confidence bands. Gaps further influence interpretation of result of hypothesis test (Subsection 4.4.6).

0.10.2   Transformation

In principle, a transformation tool could have been added to XTREND (after data read): $x1$ linearly transformed, absolute value, etc. The purpose could be: firstly, to have extreme events only in one direction (positive or negative) as XTREND assumes; secondly, to allow a logarithmic x-scale. This transformation tool is avoided: for most of the envisaged geological applications, it seems not necessary because extreme negative and positive events should be treated separately. Further, transformations as the logarithmic should be carried out prior to the XTREND analysis.

0.11   Examples
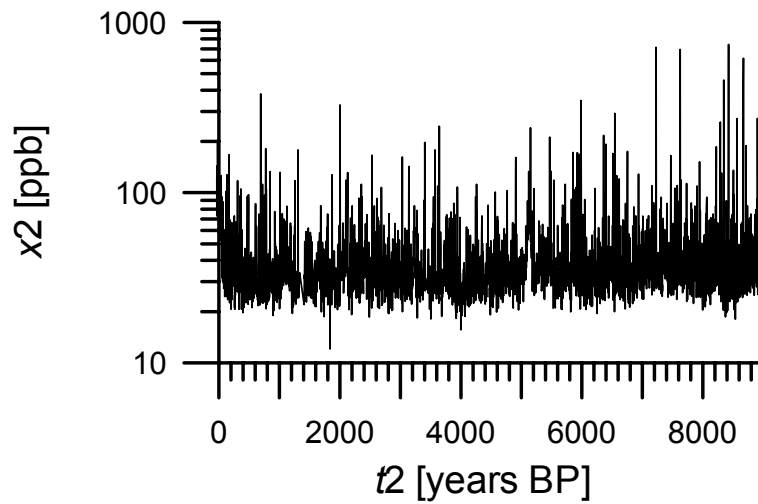
Example 1: GISP2 ice core (Greenland), SO_4^{2-}, Datatype: **ordinary**, $n1 = 3929$, observation interval: [-36.0; 9001.0] years before present (defined as AD 1950) (Zielinski *et al.* 1994). The major sulfate peaks indicate events of explosive volcanism.
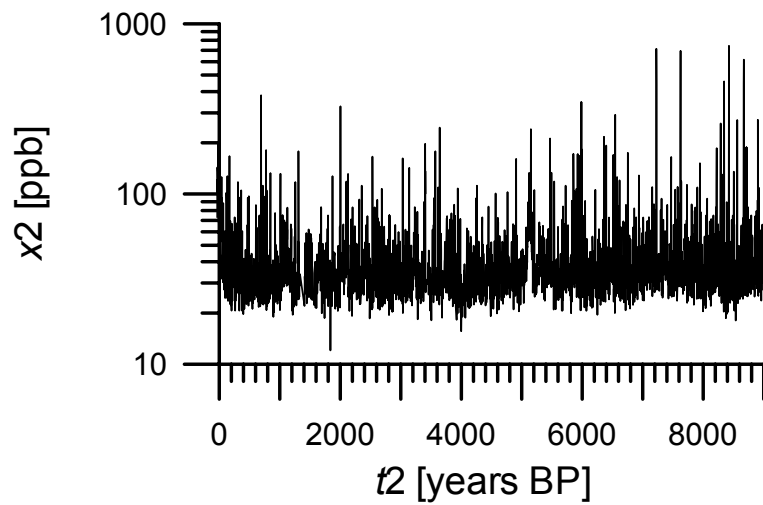
Example 2: CCB ice core (Alps), Ca^{2+}, Datatype: **segmented**, $n1 = 793$, observation interval: [0.0; 360.0] years before present (defined as AD 1996) (Wagenbach *et al.* 1996), logarithmic age model (Wagenbach 1999, pers. comm.). The major calcium peaks indicate events of Saharan dust input.
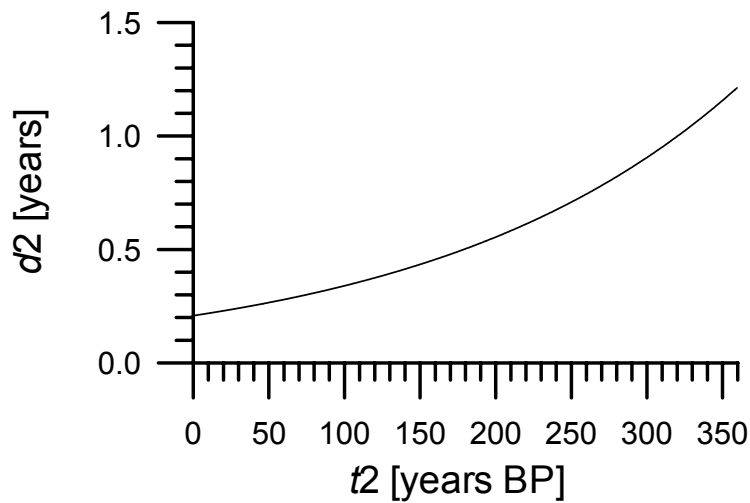
Example 2: CCB ice core, duration.

Part 1: Time interval extraction

1.1   Access

Part 1 is accessible for datatypes:

- **ordinary**

- **segmented**

- **extreme**

- **times**

1.2   Structure

1.2.1   Ordinary, extreme

1.2.2 Segmented
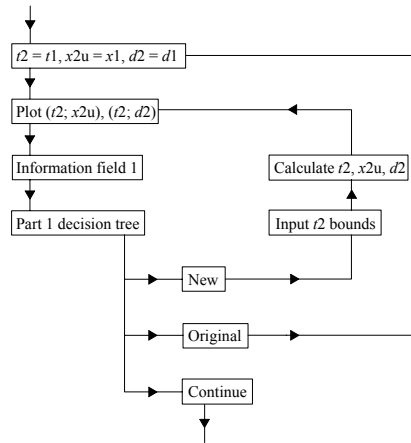
1.2.3 Times
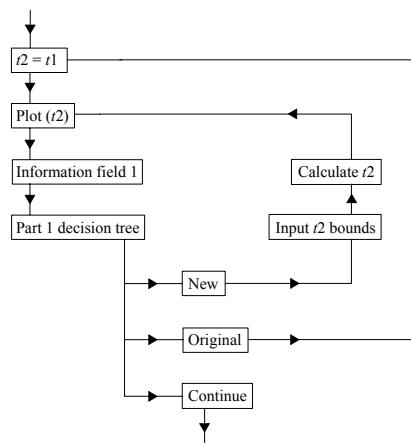
1.3 Information field 1

This information field displays:

- data file name

- datatype

- time interval—original ($t1$), value of $n1$

- time interval—observed

- (if $n1 \neq n2$) time interval—extracted ($t2$), value of $n2$

1.4 Purpose

Time interval extraction allows to analyse parts of your time series in detail using the methods of Part 2 (Extreme events detection) and Part 3 (Magnitude classification).

1.5 Notes

1.5.1 Minimum number of points

A new time interval has to contain at least $n_{\min} = 25$ data points.

1.5.2 You and XTREND

When your input is outside of the permissible range XTREND alerts you accordingly and asks for new input. You have five chances to supply a sufficient input.

1.5.3 Original time interval

For estimating the occurrence rate (Part 4), the original time interval has to be used.

Part 2: Extreme events detection

2.1 Access

Part 2 is accessible for datatypes:

- **ordinary**

- **segmented**

2.2 Structure

2.2.1 Ordinary

2.2.2   Segmented

```
                              │
                              ▼
                    ┌─────────────────────┐
                    │ Information field 1  │
                    └─────────────────────┘
                              │
                              ▼
                    ┌──────────────────────────────┐
                    │ Print k_u_cv1, k_u_cv2, k_u_cvm │───────────────┐
                    └──────────────────────────────┘                 │
                              │                                       │
                              ▼                                       │
                    ┌─────────────┐                                   │
                    │ Input k_u   │                                   │
                    └─────────────┘                                   │
                              │                                       │
                              ▼                                       │
                    ┌─────────────────────┐                          │
                    │ Plot (t2; x2u, b2u) │──────────┐               │
                    └─────────────────────┘          │               │
                              │                       │               │
                              ▼                       │               │
                    ┌─────────────────────┐          │               │
                    │ Information field 1  │          │               │
                    └─────────────────────┘          │               │
                              │                       │               │
                              ▼                       │               │
                    ┌─────────────────────┐          │               │
                    │ Information field 2  │          │               │
                    └─────────────────────┘          │               │
                              │                       │               │
                              ▼                       │               │
                    ┌─────────────────────┐          │               │
                    │ Part 2 decision tree i │        │               │
                    └─────────────────────┘          │               │
                         │                            │               │
                         ├───►┌──────────────────┐    │               │
                         │    │ Change plot setting │──┘               │
                         │    └──────────────────┘                    │
                         │                                            │
                         ├───►┌──────────┐                           │
                         │    │ Set k_u  │──────────────────────────┘
                         │    └──────────┘
                         │
                         └───►┌──────────┐
                              │ Continue │
                              └──────────┘
                                   │
                              ▼
                    ┌──────────────────────────┐
                    │ x2 = (x2u - b2u)/v2u      │───────┘
                    └──────────────────────────┘
                              │
                              ▼
                    ┌──────────────────┐
                    │ z_test = 2.0, 4.0 │
                    └──────────────────┘
                              │
                              ▼
                    ┌─────────────────────┐
                    │ Information field 1  │──────────────────┐
                    └─────────────────────┘                   │
                              │                               │
                              ▼                               │
                    ┌──────────────────────────────┐          │
                    │ Print k_cv1, k_cv2, k_cvm     │          │
                    └──────────────────────────────┘          │
                              │                               │
                              ▼                               │
                    ┌──────────┐                             │
                    │ Input k  │                             │
                    └──────────┘                             │
                              │                               │
                              ▼                               │
                    ┌────────────────────────────────────┐    │
                    │ Plot (t2; x2, b2, b2 + z_test * v2) or │──┐ │
                    │ plot (t2; x2, b2, b2 + z * v2)         │  │ │
                    └────────────────────────────────────┘    │ │
                              │                                 │ │
                              ▼                                 │ │
                    ┌─────────────────────┐                    │ │
                    │ Information field 1  │                    │ │
                    └─────────────────────┘                    │ │
                              │                                 │ │
                              ▼                                 │ │
                    ┌─────────────────────┐                    │ │
                    │ Information field 2  │                    │ │
                    └─────────────────────┘                    │ │
                              │                                 │ │
                              ▼                                 │ │
                    ┌──────────────────────┐                   │ │
                    │ Part 2 decision tree ii │                 │ │
                    └──────────────────────┘                   │ │
                         │                                      │ │
                         ├───►┌──────────────────┐             │ │
                         │    │ Change plot setting │───────────┘ │
                         │    └──────────────────┘               │
                         │                                        │
                         ├───►┌──────────────┐                   │
                         │    │ Set z_test / z │──────────────────┘
                         │    └──────────────┘
                         │
                         ├───►┌──────────┐
                         │    │ Set k    │──────────
                         │    └──────────┘
                         │
                         ├───►┌──────────┐
                         │    │ Set k_u  │──────────
                         │    └──────────┘
                         │
                         ├───►┌──────────┐
                         │    │ Part 1   │──────────
                         │    └──────────┘
                         │
                         └───►┌──────────┐
                              │ Continue │
                              └──────────┘
                                   │
                                   ▼
```
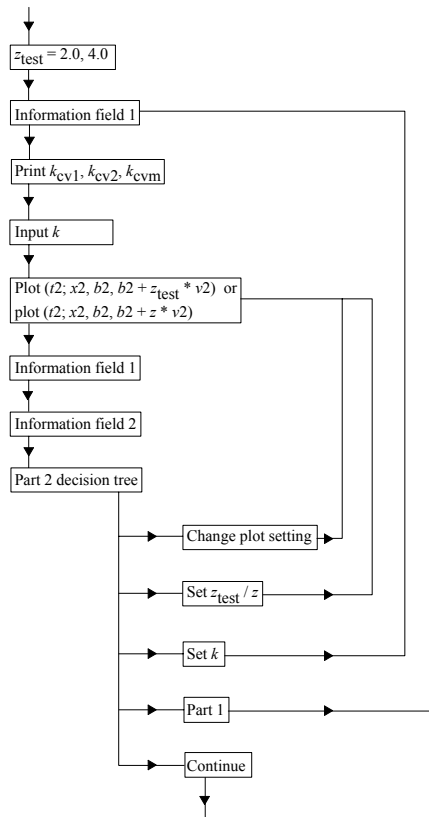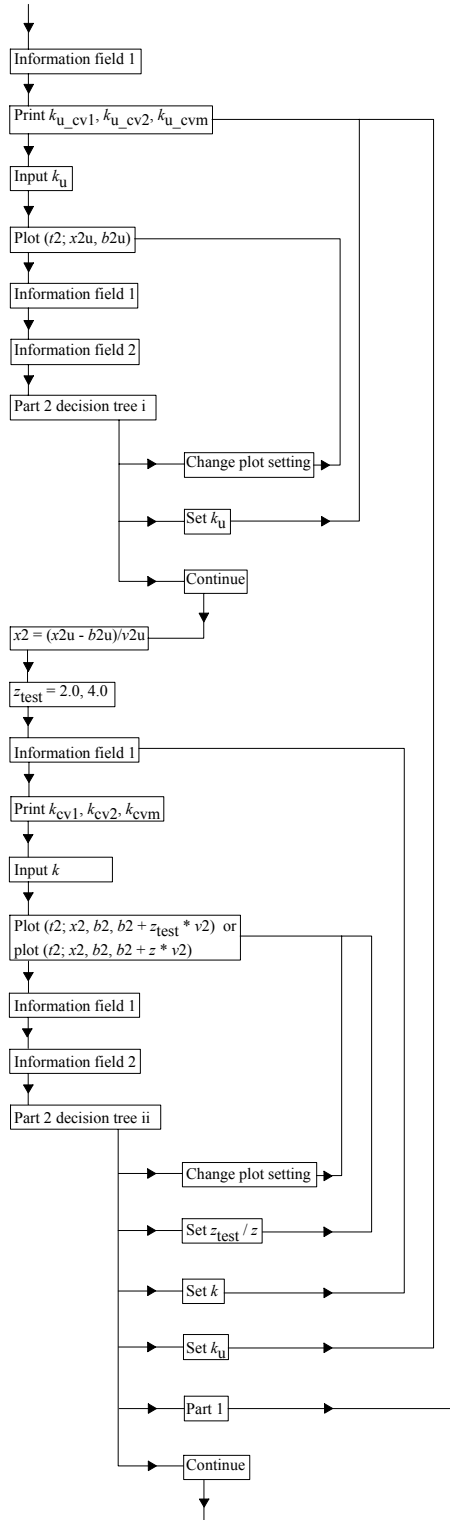
2.3   Information field 2

This information field displays:

- values of k (and, for **segmented**, k_u)

- value of z

- number $n3$ of extreme points

2.4   Plot setting

Allows you to change the plot style:

- horizontal range

- vertical range

- linear/logarithmic abscissa

2.5   Purpose

Both datatypes that have access to Part 1 assume that a background source contributes to the data. In climatology, a constant background is often unrealistic. XTREND estimates the time-dependent background, $b2$, by running median smoothing ($2k + 1$ points). This is a robust method (e. g., Tukey 1977), that means, backround estimation is not affected by outliers/extreme values which, indeed, are assumed to exist in the time series.[1]

Of particular importance is the choice of k (smoothing problem). XTREND attacks that problem by cross-validation, using three criterions: L_1-norm,

---

[1] On the contrary, the running mean would not provide a robust background estimation. Likewise, the standard deviation does not provide a robust estimation of the variability in the time series. Despite these facts, Cuomo *et al.* (2000) who used mean and standard deviation, titled their study "Robust statistical methods to discriminate extreme events in geoelectrical precursory signals: Implications with earthquake prediction".

L_2-norm and median criterion (Zheng and Yang 1998):

$$CV_1(k) = \left[\sum_{i=1}^{n2} |x(i) - b_{k,-1}(i)| \right] \Big/ n2,$$

$$CV_2(k) = \left[\sum_{i=1}^{n2} (x(i) - b_{k,-1}(i))^2 \right]^{1/2} \Big/ n2,$$

$$CV_\mathrm{m}(k) = \mathrm{median}\left\{ |x(i) - b_{k,-1}(i)| \right\},$$

where $b_{k,-1}$ is the delete-one background estimate. The cross-validation functions measure the average ability of $b_{k,-1}(i)$ to predict the observation $x(i)$. k_{CV1} minimizes $CV_1(k)$, analogously k_{CV2} and k_{CVm}.

To detect an extreme value against the background, $b2$, the running MAD (median of absolute distances to the median), $v2$, is calculated as a robust measure of variability[2] ($2k+1$ window points) (Tukey 1977). A threshold, z, has to be selected. The extreme times, $t3$, are then defined and the extreme values, $x3$, calculated as follows:

$$\text{if} \quad \frac{x2(i) - b2(i)}{v2(i)} > z > 0 \quad \text{(positive extreme event)}$$

$$\text{or} \quad \frac{x2(i) - b2(i)}{v2(i)} < z < 0 \quad \text{(negative extreme event)}$$

$$\text{then} \quad x3(j) = \frac{x2(i) - b2(i)}{v2(i)} \quad \text{and}$$

$$t3(j) = t2(i).$$

A typical value for z is 3.5 (Hampel 1985).

2.6 Notes

2.6.1 Transformation for segmented data

Index "u" in $x2\mathrm{u}$, $b2\mathrm{u}$, $v2\mathrm{u}$ etc. (**segmented**) means "untransformed". Mudelsee (1999) has shown that in case of non-constant $d2$, the transformation

$$x2(i) = [x2\mathrm{u}(i) - b2\mathrm{u}(i)] \cdot d2(i)$$

has to be applied to correct for the "dilution" of extreme values by background values (the degree of dilution depending on $d2$).

---

[2]Cf. the last footnote.

2.6.2   Double heap

The algorithm of Härdle and Steiger (1995) employs a double heap order of window data which is updated as the window moves. XTREND performs sorting to generate the initial double heap which is evidently not the most efficient method (although the increase in computing time is negligible). It might constitute an interesting research problem how to efficiently generate a double heap order from unsorted data.

2.6.3   Median smoothing parameter selection

XTREND writes the data of cross validation functions $CV_1(k)$, $CV_2(k)$ and $CV_\mathrm{m}(k)$, $k = 1, \ldots, n1 \cdot k_\mathrm{rel}$ into file `CVMED.DAT`. You may wish to inspect these curves when being prompted to select k or, for **segmented**, k_u. For that, open a second DOS window and run

<div align="center">

`gnuplot CVMED.PLT`

</div>

(the plot file is included in the XTREND files). The three criterions used are aimed to provide three looks at the trend/variability characteristics of the time series. It is conjectured that robust measures CV_1 and CV_m are more useful than CV_2 for data "contaminated" with extreme values. However, note that autocorrelation in the time series can seriously influence the CV functions (Simonoff 1996). For climatic and meteorological time series positive autocorrelation (persistence) is to be expected (Mudelsee (in press)) which means that you should use somewhat higher smoothing values. Be therefore advised to experiment with different smoothing values (e. g., local minima of CV functions) (Marron 1987, 1988), use a value of k_rel large enough (say, 0.5), and also to use your knowledge about the data: at which timescales do the recorded climatic processes act etc.

2.6.4   Threshold

z is rounded from your input to the nearest half integer (also negative). As regards z selection, Hampel's rule $z = 3.5$ is only a guide. Obviously, a larger value for z makes it more likely to (correctly) reject background values—but also more likely to miss extreme values, and vice versa. XTREND allows to plot $b2 + z_\mathrm{test} \cdot v2$ for up to three values of z_test simultaneously to help you to develop an intuition where the right threshold for your data lies.

2.7 Examples
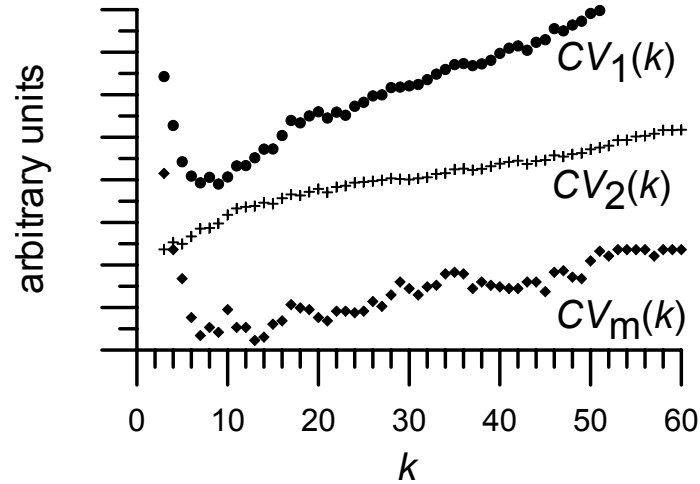
Example 1: GISP2 ice core, CV functions (running median smoothing). CV_1 has absolute minimum at $k_{\mathrm{CV1}} = 9$, CV_2 has no local minimum, and CV_m has absolute minimum at $k_{\mathrm{CVm}} = 13$.
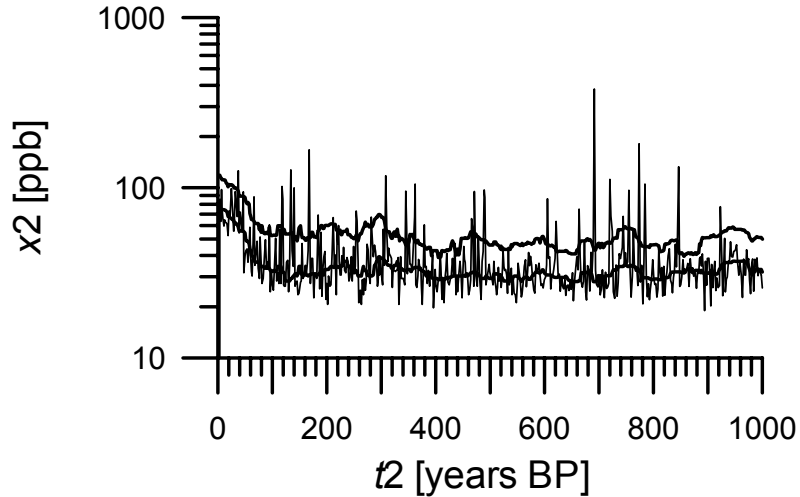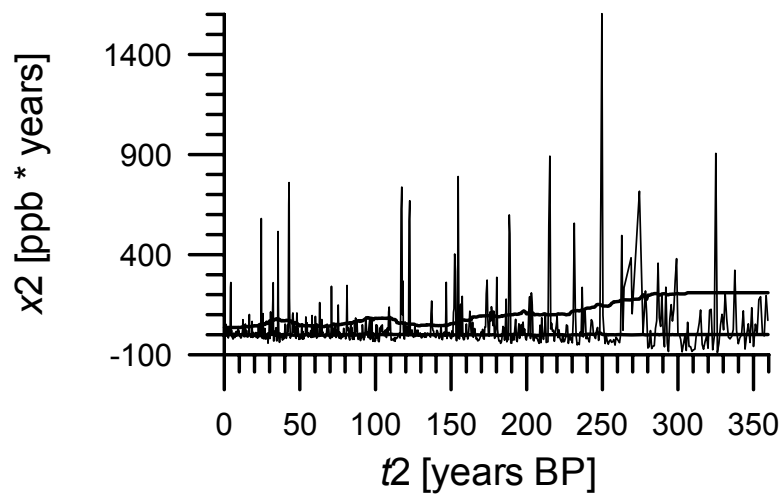

Example 1: GISP2 ice core, extreme events detection (young part plotted here). The lower thick line is time-dependent background, $b2$, estimated using $k = 13$; the upper thick line is time-dependent threshold, $b2 + z \cdot v2$, using Hampel's (1985) rule ($z = 3.5$). The background contains non-volcanic, weaker sulfate signals (e. g., from oceanic sources).

Part 2: Exteme events detection
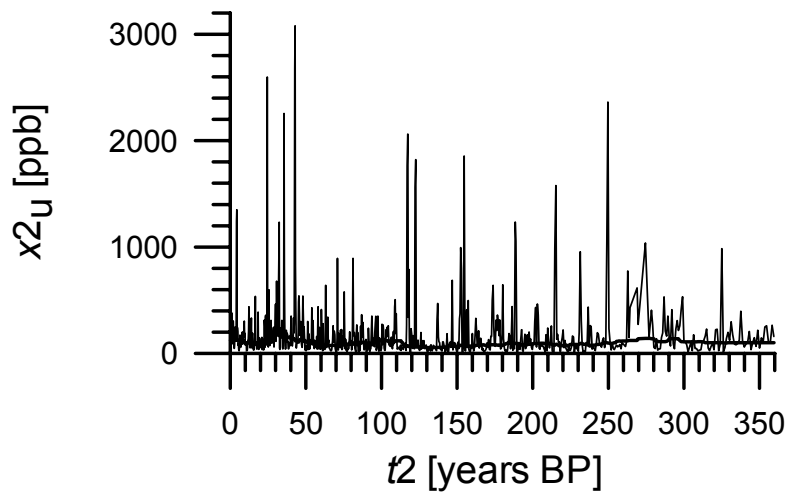
Example 2: CCB ice core, extreme events detection. The thick line is untransformed time-dependent background, $b2u$, estimated using $k_u = 27$. The background contains calcium ions not derived from Saharan dust.

Example 2: CCB ice core, extreme events detection. The lower thick line is transformed time-dependent background, $b2$, estimated using $k = 50$; the upper thick line is time-dependent threshold, $b2 + z \cdot v2$, using Hampel's (1985) rule ($z = 3.5$).

Part 3: Magnitude classification

3.1 Access

Part 3 is accessible for datatypes:

- **ordinary**

- **segmented**

- **extreme**

3.2 Structure

3.2.1 Ordinary, segmented
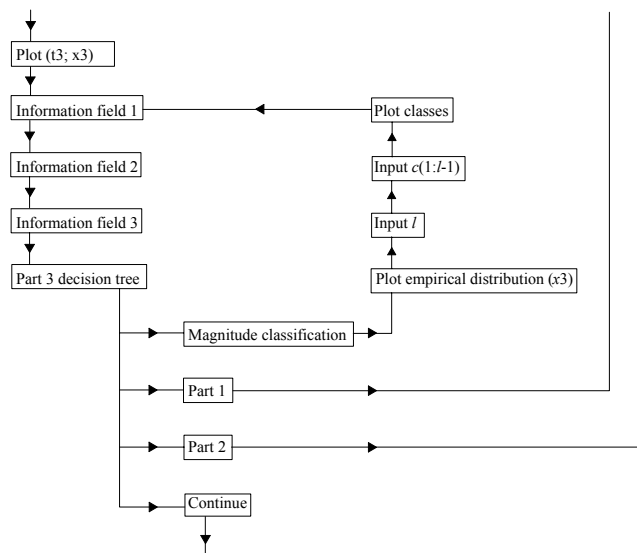
Part 3: Magnitude classification

3.2.2   Extreme
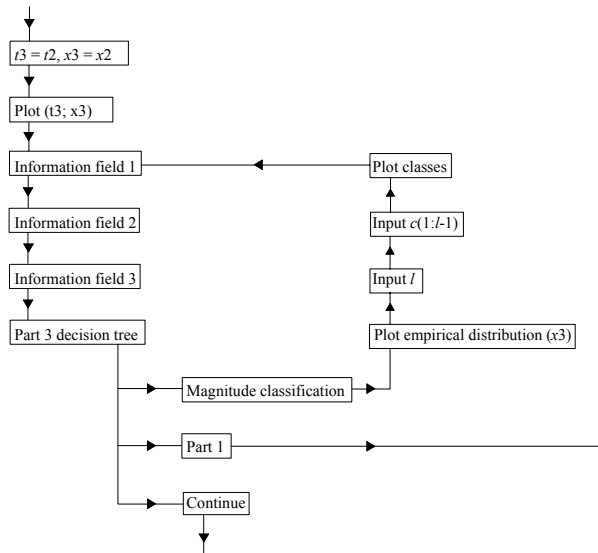
3.3   Information field 3

This information field displays for each class $i = 1, \ldots, l$:

- class intervals $[c(i-1);\ c(i)]$

- number of points, $n4(i)$

- (if set (Part 4)) kernel bandwidth $h(i)$

- (if set (Part 5)) number of bootstrap simulations $n_{\text{sim}}(i)$

3.4   Purpose

To analyse the x coordinate of your data (see Preface). For example, you may be interested to compare the occurrence rate of major flood events with that of minor flood events. Obviously, your choice of the number of classes, l, and the class bounds, c, will depend on the number of points, $n3$. Furthermore, your intimate knowledge about the data will be a helpful guide when you experiment with l and c.

3.5 Notes

3.5.1 Maximum number of classes

The maximum number of classes allowed is 6.

3.5.2 Magnitude class bounds

Class bounds selection may be made automatically (equidistant c) or by hand. Eventually, you may want to use output file `XTREND01.DAT` (from a former XTREND run with the same data) for analysing outside of XTREND the distribution of $x3$ in more detail.
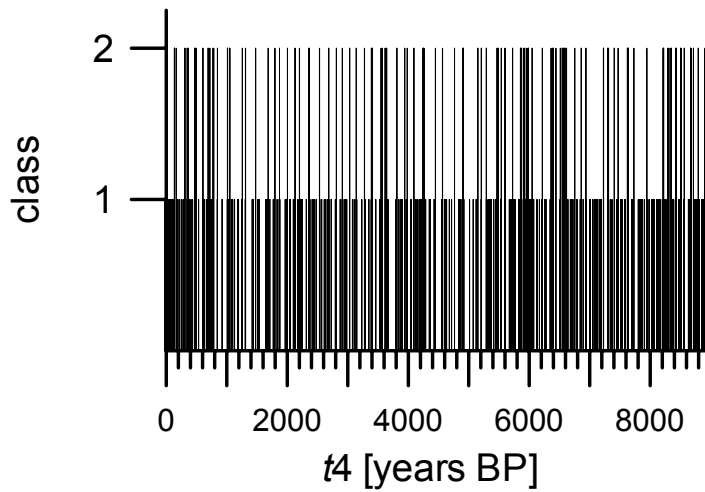
3.6 Examples
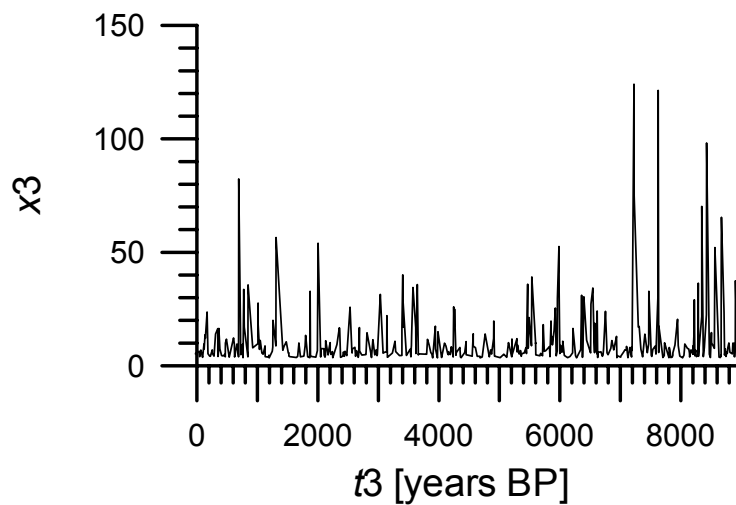
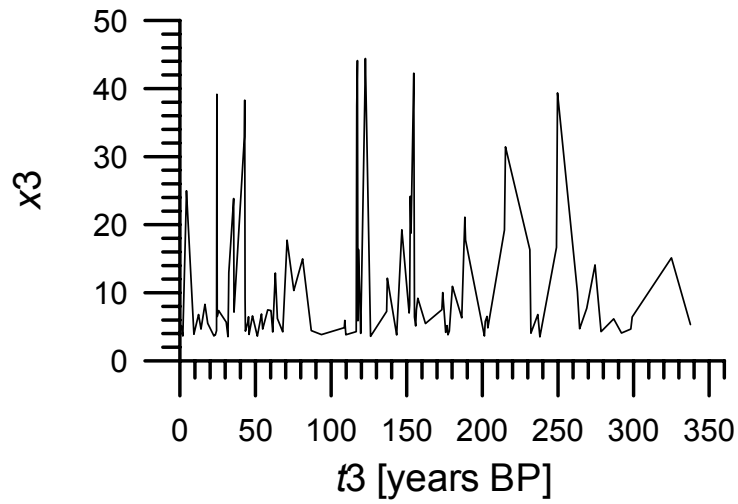Example 1: GISP2 ice core, detected extreme events ($n3 = 415$).
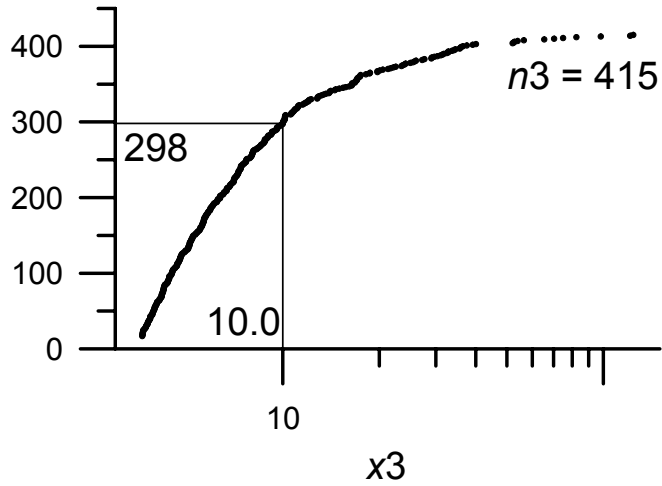
Example 1: GISP2 ice core, extreme events, empirical distribution function. Magnitude class bound is also shown.

Example 1: GISP2 ice core, classified extreme events ($l = 2$).

Example 2: CCB ice core, detected extreme events ($n3 = 103$).
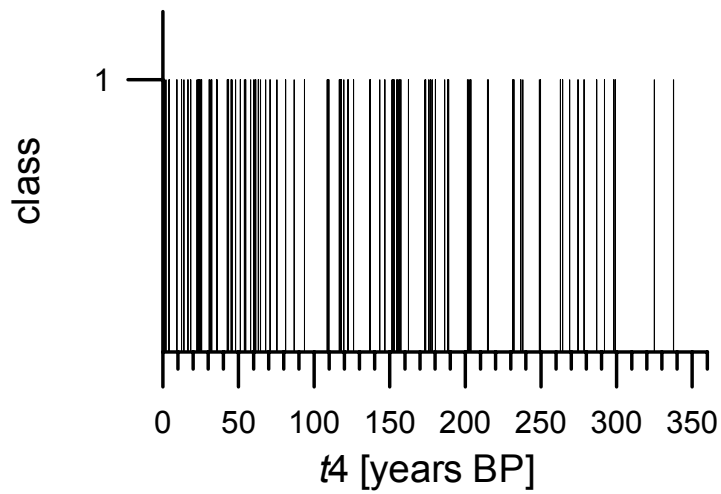
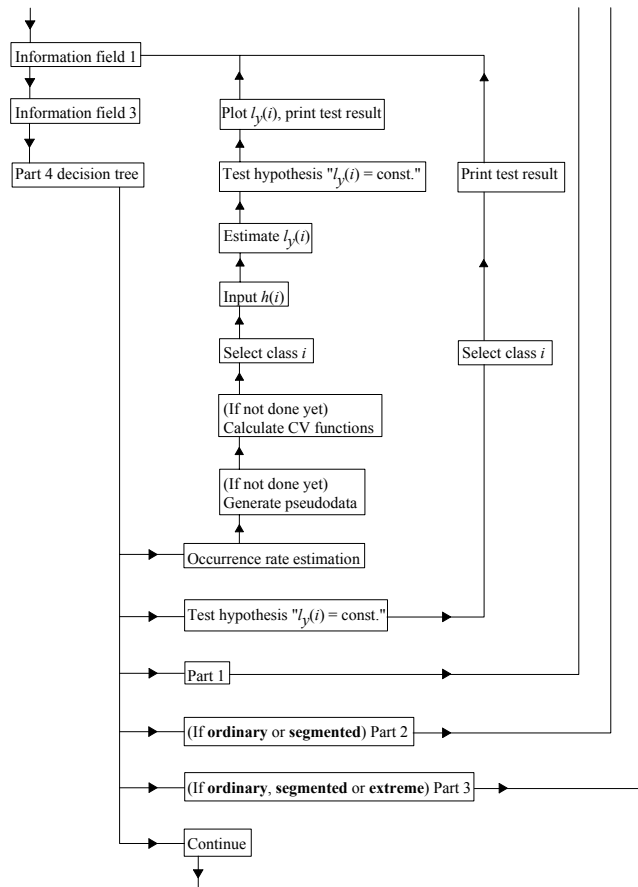Example 2: CCB ice core, classified extreme events ($l = 1$).

Part 4
Occurrence rate estimation

4.1 Access

Part 4 is accessible for datatypes:

- **ordinary**
- **segmented**
- **extreme**
- **times**

4.2 Structure

Information field 1			
Information field 3	Plot $l_y(i)$, print test result		
Part 4 decision tree	Test hypothesis "$l_y(i)$ = const."	Print test result
	Estimate $l_y(i)$	
	Input $h(i)$	
	Select class i	Select class i
	(If not done yet) Calculate CV functions	
	(If not done yet) Generate pseudodata	
Occurrence rate estimation		
Test hypothesis "$l_y(i)$ = const."		
Part 1		
(If **ordinary** or **segmented**) Part 2		
(If **ordinary**, **segmented** or **extreme**) Part 3		
Continue		

4.3   Purpose

The occurrence rate, l_y, is estimated using a kernel function (Diggle 1985):

$$\widehat{l_y}(l_x) = \sum_{i=1}^{n} K_h(l_x - t(i)), \qquad K_h(\cdot) = h^{-1} K(\cdot/h),$$

where t are the pseudodata-augmented event times (number: n) (see Subsection 4.4.1), l_x are the design points (see Subsection 4.4.2), K_h is the kernel function (see Subsection 4.4.3) and h is the kernel bandwidth (see Subsection 4.4.4). (The "hat" indicates the estimate.[3])

The advantages of that method over histogram binning are well known from the context of density estimation (Silverman 1986). For example, why should a decade (in which you count the number of extreme events) start at AD 1880 and not at AD 1878? Mudelsee *et al.* (2001) give an illustrating example in case of flood events.

XTREND additionally tests the hyopthesis "l_y is constant within the observation interval" (see Subsection 4.4.6). This test is not possible for extracted time intervals.

4.4   Notes

4.4.1   Pseudodata

Since the extreme event times for a magnitude class, $t4$, are restricted to the observation interval, $[t_{\mathrm{obs-l}}; \; t_{\mathrm{obs-r}}]$, boundary effects (i. e., reduced $\widehat{l_y}$) will occur near (within, say, $\approx 3 \cdot h$) the boundaries. This bias may be severe in case of occurrence rate estimation (in contrast with density estimation) because the observed process may continue outside the observation interval.

A computationally efficient method to reduce boundary effects is to generate pseudodata outside $[t_{\mathrm{obs-l}}; \; t_{\mathrm{obs-r}}]$. The simplest pseudodata method is reflection of data ($t4$) at the boundaries (parameter rule = 'reflection'), for example:

$$t(i) = t_{\mathrm{obs-l}} - \left[t4(i) - t_{\mathrm{obs-l}} \right],$$

where $t = t4 +$ pseudodata (t is sorted finally). Evidently, the reflection method fails to take into account a nonzero slope of l_y at a boundary. The

---

[3]This Report uses the "hat" notation rather loosely.

pseudodata method of Cowling and Hall (1996) overcomes this problem by extrapolating the empirical distribution function of $t4$. Depending on the kind of extrapolation, a variety of pseudodata generating rules are given. Cowling and Hall (1996) found via a Monte Carlo simulation study that their pseudodata method outperforms other methods devised to reduce boundary effects. XTREND has implemented two pseudodata generating rules that were best in the simulation study:

rule 'threepoint':
$$t(i) = t_{\mathrm{obs-l}} - 5\left[t4(i/3) - t_{\mathrm{obs-l}}\right] - 4\left[t4(2i/3) - t_{\mathrm{obs-l}}\right] + 10/3\left[t4(i) - t_{\mathrm{obs-l}}\right],$$

rule 'twopoint ':
$$t(i) = t_{\mathrm{obs-l}} - 9\left[t4(i/3) - t_{\mathrm{obs-l}}\right] + 2\left[t4(i) - t_{\mathrm{obs-l}}\right],$$

fractional data found by linear interpolation with $t4(0) := t_{\mathrm{obs-l}}$. Pseudodata right of $t_{\mathrm{obs-r}}$ are generated correspondingly. As regards the choice of parameter rule, see Subsection 0.7.

4.4.2   Design points

The design points, l_x, where l_y is calculated, cover the interval

$$\left[t_{\mathrm{obs-l}} - 3 \cdot \left[t1(n1) - t1(1)\right] \cdot h_{\mathrm{relmax}};\ t_{\mathrm{obs-r}} + 3 \cdot \left[t1(n1) - t1(1)\right] \cdot h_{\mathrm{relmax}}\right]$$

with constant spacing (n_{grid} points). For a typical value of $h_{\mathrm{relmax}} = 0.5$, this means that the number of output data (l_x, l_y) in $[t_{\mathrm{obs-l}};\ t_{\mathrm{obs-r}}]$ can be as less as $\approx n_{\mathrm{grid}}/4$. This has to be considered when selecting n_{grid} in the configuration file.

4.4.3   Kernel calculation

The kernel function is a Gaussian which enables fast calculation of the occurrence rate in Fourier space (Silverman 1982). XTREND uses the Fast Fourier Transform algorithm of Press *et al.* (1996). n_{grid} has to be a power of two. In most situations, a value of 1024 or 2048 should yield a reasonable tradeoff: finely spaced design points (Subsection 4.4.2) and acceptable RAM size requirements and speed (Section 0.1).

4.4.4   Kernel bandwidth parameter selection

The selection of kernel bandwidth, h, is a crucial step in occurrence rate estimation, similarly to median smoothing parameter selection (Subsection

2.6.3). Brooks and Marron (1991) devised a cross-validation criterion for bandwidth selection based on the L_2 norm which is asymptotically optimal for kernel occurence rate estimation. Thus, h_{CV} minimizes

$$
CV(h) = \int\limits_{t_{\mathrm{obs-l}}}^{t_{\mathrm{obs-r}}} \widehat{l}_{y_h}(l_x)^2 \, dl_x \; - \; 2 \sum_{i=1}^{n} \widehat{l}_{y_h,i}(t(i)), \qquad (4.1)
$$

where

$$
\widehat{l}_{y_h,i}(l_x) = \sum_{\substack{j=1 \\ j \neq i}}^{n} K_h(l_x - t(j))
$$

is the leave-one-out estimator. $CV(h)$ is determined for $h = [t1(n1) - t1(1)] \cdot 0.001$ to $[t1(n1) - t1(1)] \cdot h_{\mathrm{relmax}}$ ($n_{h\mathrm{srch}}$ search values). We note the following points (which apply also to median smoothing parameter selection):

- computing-time intensive step

- advise: experiment with different h values

- local minima of $CV(h)$ may point to a relevant structure in data

4.4.5   CV calculation

The second term of the right-hand side in Equation 4.1 constitutes a sum of exponentials over a rectangle (i, j) (not a square because of the pseudodata). The terms near the upper left or lower right corner are small ($\propto \exp\left\{-\left[(t(i) - t(j))/h\right]^2/2\right\}$), the terms near the 1:1 line are around unity. This led to the following idea to reduce computational effort: Calculate the terms only in the intermediate range, set the terms near ("near" defined by machine precision) the 1:1 line equal to 1, omit the terms near the two corners. However, for typical data sizes in Geosciences (less than a few thousand) and typical machine precisions, the gain is negligible and therefore not implemented in XTREND. However, XTREND takes into account the variable term size in the summation to reduce roundoff errors.

4.4.6 Hypothesis test

Under the null hypothesis "l_y = constant", the test statistic

$$u = \frac{\sum t4(i)/n4 \; - \; [t_{\text{obs}-\text{r}} + t_{\text{obs}-\text{l}}]\,/2}{[t_{\text{obs}-\text{r}} - t_{\text{obs}-\text{l}}] \; \sqrt{(1/(12 \cdot n4))}}$$

tends rapidly to the standardized normal form (e. g., Cox and Lewis 1966). XTREND reports in addition to u the respective one-sided probability, p.

4.4.7 Segmentation problem

Mudelsee (1999) devised an idea for deciding whether, in the case of **segmented** data and more than one magnitude class (evenly spaced), a segment contains in reality, say, two events of lower class instead of one event of higher class: Compare the probability, $r1$, for one class-two event (calculated using l_y(class $l = 2$) with the probability, $r2$, for two class-one events (using $l_y(l = 1)$) and redistribute $t4$ if $r2 > r1$. This idea is not implemented in XTREND because of following problems.

Limited precision:

1. Positions of new $t4$ are unknown within segment.

2. Magnitude classification uses Heaviside function. For example: event with $x3 = 1.1$ falls into class $2 = \,]1.0;\,2.0]$ and is redistributed as two class-1 events. On the other hand, two events $x3 = 0.9$ and $x3 = 1.0$ (sum $= 1.9$) also result in two class-1 events.

3. Even class spacing may limit the analysis seriously.

Implementation problems and computational burden:

1. After $t4$ redistribution, a class may contain less than n_{min} points.

2. After $t4$ redistribution, in principle, new pseudodata and new CV functions calculations are necessary and, further, new selection of bandwidths, h. Using different h, different l will result, and, eventually, the above condition $r2 > r1$ will not halt anymore.

Therefore, it might in practice not be possible to "look into the segments". The "segmentation problem", however, has to be considered when interpreting results.

4.4.8 Autocorrelation

XTREND treats extreme event times, $t3$, as realizations of a (non-homogeneous) Poisson process, that is, as independent from each other. This assumption should be tested when interpreting results. A CV function with very small h_{CV} might indicate positive autocorrelation. As noted in Subsection 2.6.3, climatic and meteorological time series often show positive autocorrelation. One reason for that is a high sampling rate of a climatological process, for example, in a sediment core. Be therefore advised to check and correct for autocorrelation outside of XTREND, using your knowledge about the data archive. One remedy might be to re-sample by hand the data at a lower resolution.

4.5 Examples
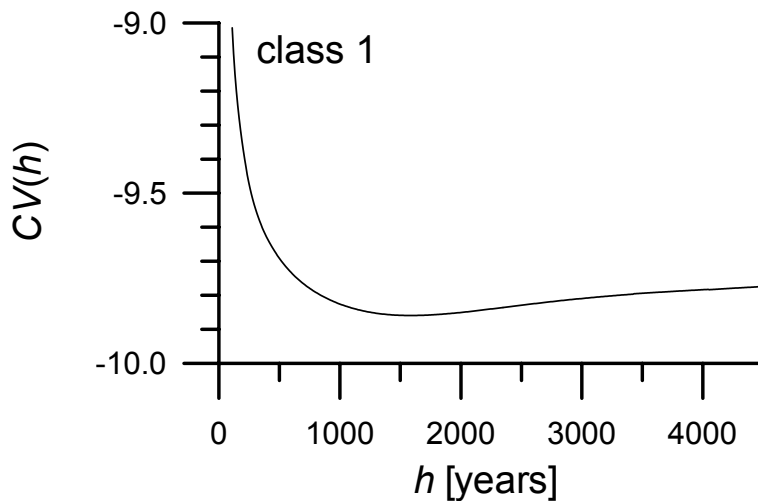
Example 1: GISP2 ice core, magnitude class 1, CV function (kernel bandwidth selection). $CV(h)$ has absolute minimum at $h_{\mathrm{CV}} = 1591.0$ years.
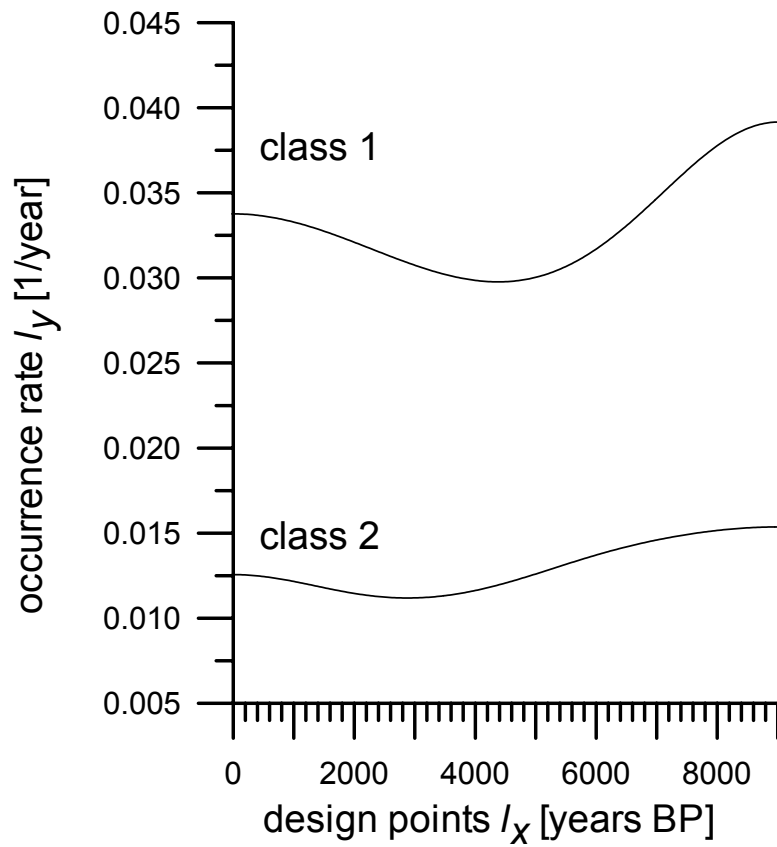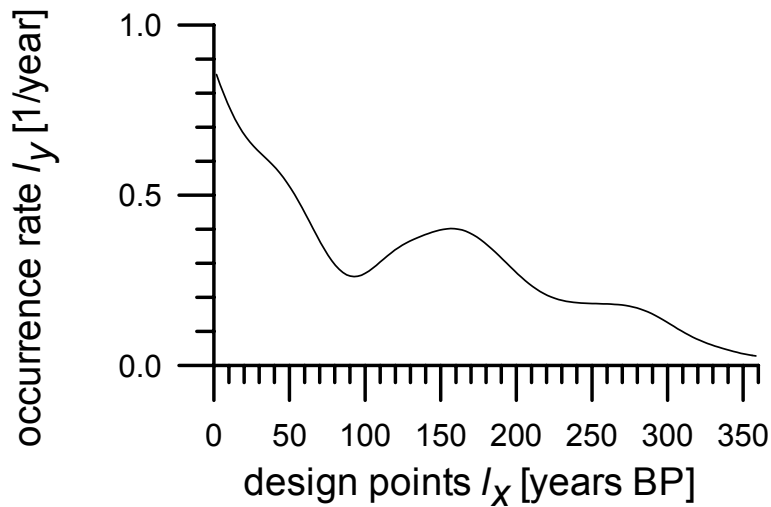
Example 1: GISP2 ice core, estimated occurrence rates ($h = 1591.0$ years and pseudodata rule "reflection" used for both classes).
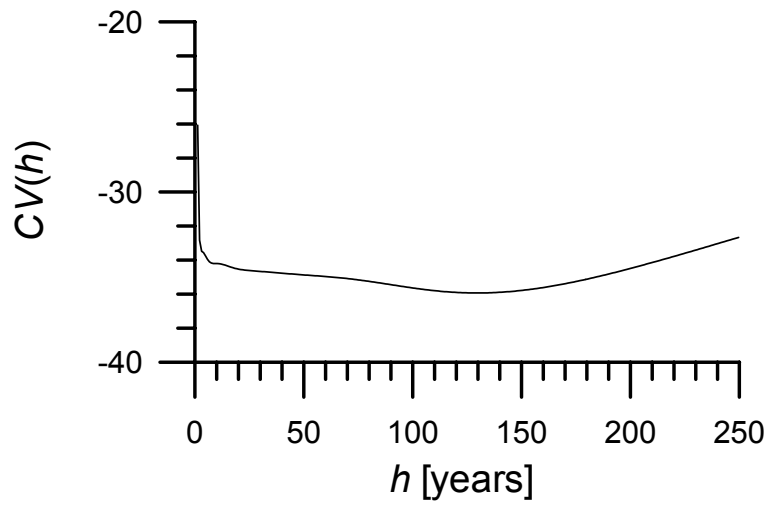
Example 2: CCB ice core, CV function (kernel bandwidth selection). $CV(h)$ has absolute minimum at $h_{\mathrm{CV}} = 129.9$ years; considerably low CV values exist for $h \gtrsim 15.0$ years.

Example 2: CCB ice core, estimated occurrence rate. $h = 20.0$ years was used instead of h_{CV} to inspect short-term changes and reduce boundary effects; pseudodata rule: "reflection".
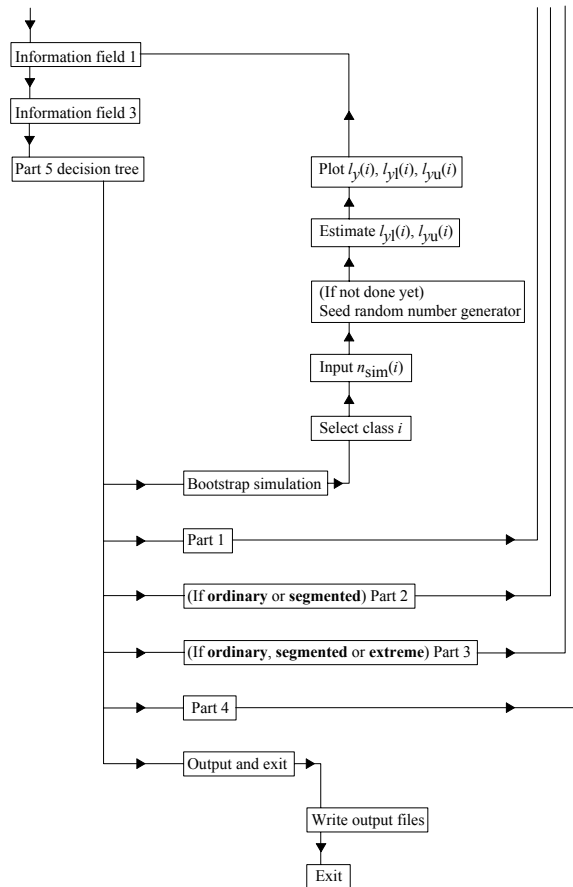
Part 5
Bootstrap simulation

5.1   Access

Part 5 is accessible for datatypes:

- **ordinary**

- **segmented**

- **extreme**

- **times**

5.2   Structure

5.3   Output files

Output file `XTREND01.DAT` is written for datatypes **ordinary** and **segmented**:

- Information field 1

- Information field 2

- Information field 3

- Data from Part 2 (Extreme events detection): $t2$, $x2$, $b2$, $v2$, $b2 + z \cdot v2$, $d2$ (**segmented**), $x2u$ (**segmented**), $b2u$ (**segmented**), $t3$, $x3$

Output file `XTREND02.DAT`:

- Information field 1

- Information field 2

- Information field 3

- Data from Part 3 (Magnitude classification): $t4$, t for all classes

Output file `XTREND03.DAT`:

- Information field 1

- Information field 2

- Information field 3

- Test result (H0: "constant occurrence rate"): u, p

- Data from Parts 4 (Occurrence rate estimation) and 5 (Bootstrap simulation): l_y, l_{yl}, l_{yu} for all classes

5.4   Purpose

A confidence band around the estimated occurrence rate is essential for interpreting results. For example, you might want to know whether a "trough" in the estimated occurrence rate of heavy rainfall events during the Maunder Minimum ($\sim$ AD 1650 to AD 1715) is realistic or, instead, came by chance into your data. XTREND uses bootstrap simulations to produce a set of simulated events, t^*. Occurrence rate estimation (using same kernel bandwidth

as for the original data) is carried out, yielding an estimate, $\widehat{l}_y^{\,*}$. The procedure simulation–estimation is repeated, n_{sim} times in total. The variability of $\widehat{l}_y^{\,*}$ allows to determine confidence bands (level α) of $\widehat{l}_y$. The methodology has been developed by Cowling *et al.* (1996).

A number of simulated events is used which is equal to the number of data, n, for the original estimation. Then, t^* is drawn from the set t with replacement[4]. The confidence band calculated is a percentile-t type: Let $E(\widehat{l}_y^{\,*})$ be the average occurrence rate (at point $l_x \in [t_{\text{obs}-1};\ t_{\text{obs}-\text{r}}]$). Calculate the studentized quantity,

$$T \;=\; \left[\widehat{l}_y^{\,*} - E(\widehat{l}_y^{\,*})\right] \Big/ \sqrt{\widehat{l}_y^{\,*}},$$

which is approximately pivotal, that is, independent of l_y and l_x. That allows you to calculate the t_α-percentiles of $|T|$. Finally, the confidence bounds are calculated as

$$l_{y\text{l}} \;=\; \max\left(0,\ E(\widehat{l}_y^{\,*}) - t_\alpha \sqrt{\widehat{l}_y}\right),$$

$$l_{y\text{u}} \;=\; E(\widehat{l}_y^{\,*}) + t_\alpha \sqrt{\widehat{l}_y}.$$

5.5   Notes

5.5.1   Other methods

In a simulation study, the method used here ("Method 3") gave nearly identical results as two other methods (Cowling *et al.* 1996). It is implemented into XTREND because of its simplicity. That study compared also two other types of confidence band with the type used here ("Type 1"); again, the differences were small.

5.5.2   Bias

Cowling *et al.* (1996) show that $\widehat{l}_y$ has a bias of approximately $h^2 l_y''(l_x)/2$ (Gaussian kernel) where l_y'' is the second derivative of l_y. You may explicitly correct for that bias (outside of XTREND) with an estimate for l_y'' obtained

---

[4]The bootstrap. Efron and Tibshirani (1993) is an excellent monograph on that resampling method.

from using a larger bandwidth in its construction. A simpler method, however, is undersmoothing (small h) which reduces bias effectively (Cowling *et al.* 1996).

5.5.3 Very small kernel bandwidth

In XTREND's subroutine that calculates the kernel occurrence rate (modified routine `denest` in Silverman (1982)), the smallest $\widehat{l_y}$ value is set equal to tiny value ϵ ($= 1.0\mathrm{e}{-}12$). This ensures that T can be calculated without risking math errors. This safeguarding, however, needs additional measures to ensure meaningful results:

1. A rough estimate of the upper bound of the average occurrence rate is given by the ratio $n1/(t_{\mathrm{obs-r}} - t_{\mathrm{obs-l}})$, that is, approximately $1/($average spacing$)$. That value should be considerably larger than tiny. XTREND checks that.

2. Regions l_x where no/few data points exist, in case of small kernel bandwidth, tend to produce $\widehat{l_y} = \epsilon$. Likewise, there: $\widehat{l_y}^* \approx \epsilon$. That means, there: $T \approx 0$. It follows that t_α will be smaller in comparison with the true value (i. e., calculated without numerical errors). The confidence band, therefore, is smaller than the true. This effect is more serious the larger the proportion of the number of points l_x where $\widehat{l_y} = \epsilon$ to the number of points $l_x \in [t_{\mathrm{obs-l}}; t_{\mathrm{obs-r}}]$ becomes in relation to confidence level α. In such cases XTREND issues a warning message. It is possible to select a larger kernel bandwidth to reduce this effect. However, for highly clustered data (say, 60 % of data in the first 10 % of the observed time interval, 40 % in the last 5 %), it might then become impossible to see the structure within a cluster. In such cases, the observations should be cut into pieces (containing one cluster), and the fine structure within each cluster analysed separately with XTREND.
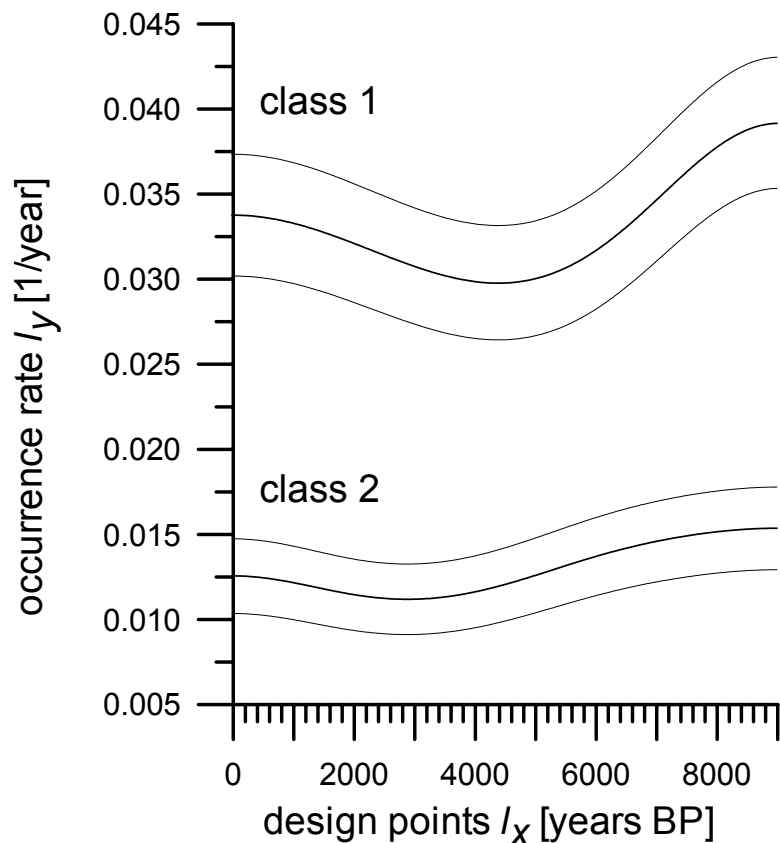
5.5.4 Computing time

The bootstrap simulations do not generally take longer than the CV calculations.

5.5.5 Exit

You may exit XTREND without performing bootstrap simulation.

5.6   Examples

Example 1: GISP2 ice core, estimated occurrence rates and confidence bands ($\alpha = 0.90$, $n_{\text{sim}} = 10000$). Hypothesis "$l_y = $ constant" (Subsection 4.4.6) cannot be rejected with reasonably small p values (class 1, $p = 0.16$; class 2, $p = 0.13$). The bootstrap simulations, however, reveal: At around 9000 years BP the occurrence rate of explosive volcanic eruptions, indicated by sulfate ions, was high (in accordance with Zielinski *et al.* (1994)). Then, as new finding here, a decrease towards mid-Holocene and again an increase towards present occurred. In case of class 1 events, this trend is significant. (Setting $l = 1$ magnitude class revealed a similar trend, at the same significance level.)
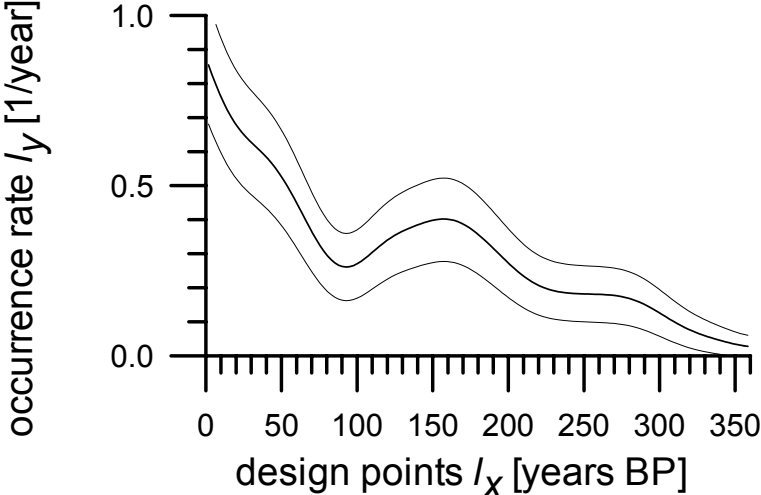
Example 2: CCB ice core, estimated occurrence rate and confidence band ($\alpha = 0.90$, $n_{\text{sim}} = 2000$). Hypothesis "$l_y = \text{constant}$" can be rejected ($p < 1.0\text{e}{-}7$) against H1: "l_y increases towards the present". The bootstrap simulations confirm the overall increasing occurrence rate and, in addition, reveal that this was interrupted by a significant "trough" at around 1900 AD ($l_x = 96.0$ years). Avila and Peñuelas (1999) found an increase in the occurrence rate of Saharan dust peaks in rainfall data from northeastern Spain during the past few decades.[5]

---

[5]A caveat against this study is that it compares data from *different* sites, that is, the data may not be homogeneous.

References

Allen MR, Stott PA, Mitchell JFB, Schnur R, Delworth TL (2000) Quantifying the uncertainty in forecasts of anthropogenic climate change. Nature 407:617–620

Avila A, Peñuelas J (1999) Increasing frequency of Saharan rains over northeastern Spain and its ecological consequences. Sci Tot Environ 228:153–156

Brooks MM, Marron JS (1991) Asymptotic optimality of the least-squares cross-validation bandwidth for kernel estimates of intensity functions. Stoch Pr Ap 38:157–165

Cowling A, Hall P (1996) On pseudodata methods for removing boundary effects in kernel density estimation. J Roy Sta B 58:551–563

Cowling A, Hall P, Phillips MJ (1996) Bootstrap confidence regions for the intensity of a Poisson point process. J Am Stat A 91:1516–1524

Cox DR, Lewis PAW (1966) The Statistical Analysis of Series of Events. Methuen, London, pp 285

Cuomo V, Di Bello G, Lapenna V, Piscitelli S, Telesca L, Macchiato M, Serio C (2000) Robust statistical methods to discriminate extreme events in geoelectrical precursory signals: Implications with earthquake prediction. Natural Hazards 21:247–261

Diggle P (1985) A kernel method for smoothing point process data. Appl Stat 34: 138–147

Easterling DR, Meehl GA, Parmesan C, Changnon SA, Karl TR, Mearns LO (2000) Climate extremes: Observations, modeling, and impacts. Science 289:2068–2074

Efron B, Tibshirani RJ (1993) An Introduction to the Bootstrap. Chapman and Hall, London, pp 436

Embrechts P, Klüppelberg C, Mikosch T (1997) Modelling Extremal Events for Insurance and Finance. Springer, Berlin, pp 648

Härdle W, Steiger W (1995) Optimal median smoothing. Appl Stat 44:258–264

Hampel FR (1985) The breakdown points of the mean combined with some rejection rules. Technomet 27:95–107

Houghton JT, Meira Filho LG, Callander BA, Harris N, Kattenberg A, Maskell K (eds) (1996) Climate Change 1995: The Science of Climate Change. Cambridge University Press, Cambridge, pp 572

Marron JS (1987) What does optimal bandwidth selection mean for nonparametric regression estimation? In: Dodge Y (ed) Statistical Data Analysis Based on the L_1-Norm and Related Methods. North-Holland, Amsterdam, 379–392

Marron JS (1988) Automatic smoothing parameter selection: A survey. Emp Econom 13:187–208

Mudelsee M (1999) On an interesting statistical problem imposed by an ice core. Institute of Mathematics and Statistics, University of Kent at Canterbury, IMS Technical Report UKC/IMS/99/21, Canterbury, pp 12

Mudelsee (in press) TAUEST: A computer program for estimating persistence in unevenly spaced weather and climate time series. Comput Geosci

Mudelsee M, Börngen M, Tetzlaff G (2001) On the estimation of trends in the frequency of extreme weather and climate events. In: Raabe A, Arnold K (eds) Wissenschaftliche Mitteilungen, Band 22. Institute of Meteorology, University of Leipzig, Leipzig, 78–88 [in German]

Olsen JR, Lambert JH, Haimes YY (1998) Risk of extreme events under nonstationary conditions. Risk Anal 18:497–510

Press WH, Teukolsky SA, Vetterling WT, Flannery BP (1996) Numerical Recipes in Fortran 90. Cambridge University Press, Cambridge, pp 935–1446

Reiss R-D, Thomas M (1997) Statistical Analysis of Extreme Values. Birkhäuser, Basel, pp 316

Silverman BW (1982) Kernel density estimation using the Fast Fourier Transform. Appl Stat 31:93–99

Silverman BW (1986) Density Estimation for Statistics and Data Analysis. Chapman and Hall, London, pp 175

Simonoff JS (1996) Smoothing Methods in Statistics. Springer, New York, pp 338

Tukey JW (1977) Exploratory Data Analysis. Addison-Wesley, Reading, Massachusetts, pp 688

von Storch H, Zwiers FW (1999) Statistical Analysis in Climate Research. Cambridge University Press, Cambridge, pp 484

Wagenbach D, Preunkert S, Schäfer J, Jung W, Tomadin L (1996) Northward transport of Saharan dust recorded in a deep Alpine ice core. In: Guerzoni S, Chester R (eds) The Impact of Desert Dust Across the Mediterranean. Kluwer, Dordrecht, 291–300

Zielinski GA, Mayewski PA, Meeker LD, Whitlow S, Twickler MS, Morrison M, Meese DA, Gow AJ, Alley RB (1994) Record of volcanism since 7000 B.C. from the GISP2 Greenland ice core and implications for the volcano–climate system. Science 264:948–952

Zheng Zg, Yang Y (1998) Cross-validation and median criterion. Statist Sinica 8:907–921