# On an interesting statistical problem imposed by an ice core


Manfred Mudelsee

Institute of Mathematics and Statistics
University of Kent,
Canterbury CT2 7NF
United Kingdom

E-mail: M.Mudelsee@ukc.ac.uk


April 1999


**Summary**

An ice core had been sampled in segments of constant length but increasing duration. The record exhibits Saharan dust peaks against a background. This report gives ideas for solving methodical problems that arise when a marked point process is to be estimated as a model for the peaks. The main one is the "segmentation problem": how to decide whether a given segment, which is represented by one single measurement, contains several weak events or, instead, only a few events of higher magnitude. Further investigated problems are the variable dilution of peaks by the background and peak definition/background estimation.



**Institute of Mathematics and Statistics,**
**University of Kent, Canterbury,**
**IMS Technical Report UKC/IMS/99/21**

# 1 The ice core

Ice cores are important archives of past climate. The ice accumulates in stratigraphic layers which supply us with the temporal information: young top, old bottom. Various physical and chemical parameters of the ice can be measured which provide us with the climatic information for the respective time, such as air temperature or atmospheric carbon dioxide concentration. Oeschger and Langway (1989) give an overview about the potential of ice cores as environmental recorders.

The CCB ice core (Wagenbach *et al.* 1996) was drilled into the glacier of the Colle Gnifetti mountain, Alps. It was divided into segments of constant length, 0.05 mwe, which is determined by the minimum material necessary for concentration measurements. (The mwe depth unit means “meter water equivalent”.) The calcium ($Ca^{2+}$) record against depth is shown in Fig. 1. A few segments have no measurement.

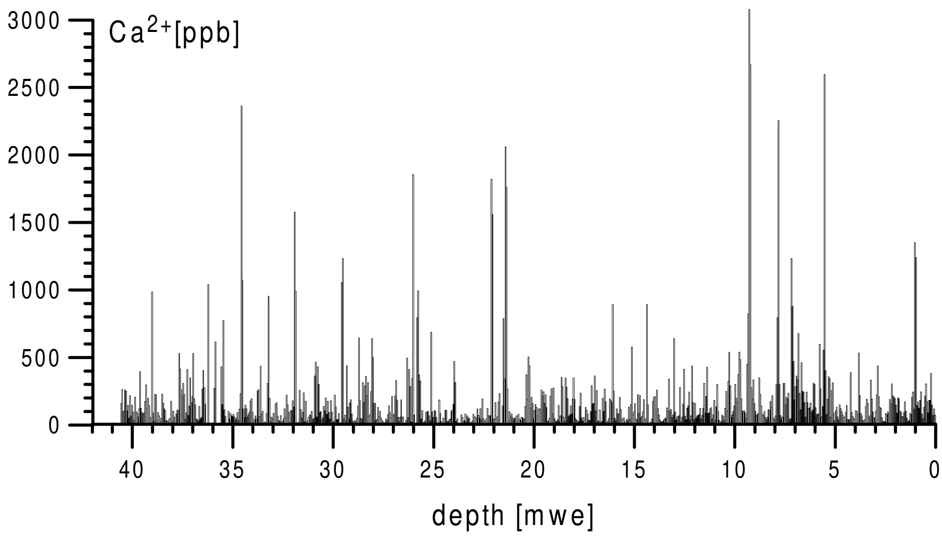


Figure 1: The CCB ice core $Ca^{2+}$ record against depth. The number $N$ of measurements is 793. The number of gaps is 19.

The record exhibits a number of prominent peaks against a background. These peaks reflect (Wagenbach *et al.* 1996) events of Saharan dust input. The dust events are of relevance for the local environment: They increase the avalanche risk, they lower the albedo (reflectivity) thus influencing the glacier's mass balance, and they constitute an extra flux of essential trace elements and nutrients to the alpine ecosystem (Wagenbach *et al.* 1996).

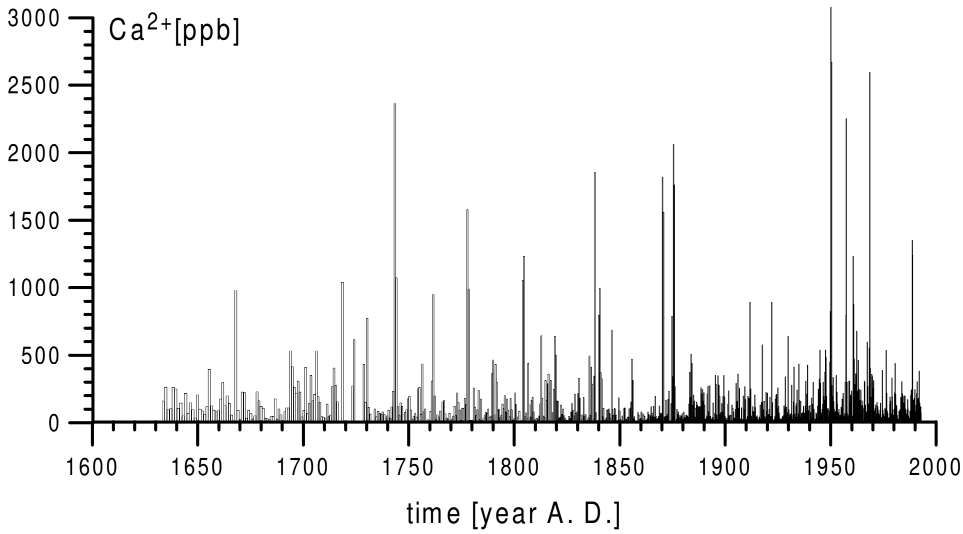


Figure 2: The CCB ice core $Ca^{2+}$ record against time (logarithmic depth–time model).

The physical properties of the ice accumulation and aging process determine the relation between core depth and time. In general, the ice is being compacted by its mass. Thus, the duration which a core segment represents, increases with depth (see Fig. 3). For the CCB core we use a simple logarithmic model (Wagenbach 1998: personal communication),

$$t \;=\; 1993 \text{ A. D. } + \frac{H}{A} \ln\left(1 - \frac{y}{H}\right),$$

where $t$ is time, $y$ is depth, $H$ = maximum depth = 49 mwe and $A$ = ice accumulation rate = 0.24 mwe/year. Fig. 2 shows the CCB record against time. Of course the depth–time model selection is a crucial part. However, we do not discuss it here. Due to the nonlinear relation, the dating uncertainty increases strongly with core depth; Wagenbach *et al.* (1996) give approximate uncertainties of 2 years for the top and 50 years for the bottom.

We focus on how to estimate the frequency (which is allowed to vary with time) of the Saharan dust peaks and also their magnitude which both are of climatological interest. This leads us to the subject of marked point processes (Section 4) where we discuss the main problem caused by the segmentation of the ice core: how to decide whether a given segment (i. e., one measurement) contains several weak events or, instead, only a few events of higher magnitude. But, in view of the nonlinear depth–time relation, we have to consider first the physics of the calcium concentration values to extract (by a transformation) a meaningful climate signal (Section 2). We have also to estimate the signal background (Section 3) which we allow to vary with time since that is more realistic (Wagenbach *et al.* 1996). Section 3, in addition, deals with the definition of a dust peak. This report merely looks on methodical problems and gives ideas for solution. Concrete results for the ice core will be presented elsewhere.

## 2 Transformation

Table 1 lists the used notation. The measured $Ca^{2+}$ concentration is (dust peak and background contribution):

$$C(i) \;=\; \frac{N_p(i) + N_b(i)}{N_p(i) + \bar{N}_p(i) + N_b(i) + \bar{N}_b(i)}.$$

**Proposition 1:** The number of non–$Ca^{2+}$ ions from the background dominates,

$$\bar{N}_b(i) \;\gg\; N_p(i) + \bar{N}_p(i) + N_b(i).$$

This is certainly fulfilled since the bulk of material is water. Then we have

$$C(i) - C_b(i) \;\simeq\; \frac{N_p(i)}{\bar{N}_b(i)}, \qquad (1)$$

Table 1: Notation.

| | |
|---|---|
| $i = 1, \ldots, N$ | segment (measurement) number (top: $i = 1$) |
| $N$ | number of segments |
| $j$ | counts the events in a segment |
| $t$ | time |
| $y$ | core depth |
| $T_u(i)$ | time (upper end of segment) |
| $T_l(i)$ | time (lower end of segment) |
| $T(i)$ | average time (segment) (NB: nonlinear timescale) |
| $D(i)$ | $\equiv T_u(i) - T_l(i)$ = duration (segment) |
| $D_p(i)_j$ | duration of the $j$th dust event in segment |
| $N_p(i)_j$ | number of $Ca^{2+}$ ions from $j$th event in segment |
| $N_p(i)$ | number of $Ca^{2+}$ ions from Saharan dust peaks $= \sum_j N_p(i)_j$ |
| $\bar{N}_p(i)$ | number of non–$Ca^{2+}$ ions from Saharan dust peaks |
| $N_b(i)$ | number of $Ca^{2+}$ ions from background |
| $\bar{N}_b(i)$ | number of non–$Ca^{2+}$ ions from background |
| $C(i)$ | measured $Ca^{2+}$ concentration (segment) |
| $C_b(i)$ | $Ca^{2+}$ background concentration (segment) |
| $A(i)$ | area (right angle) and "susceptibility" to Saharan dust flux |
| $\bar{A}(i)$ | area (right angle) and "susceptibility" to background flux |
| $F(i)_j$ | flux of $Ca^{2+}$ ions from $j$th dust event in segment |
| $\bar{F}_b(i)$ | flux of non–$Ca^{2+}$ ions from background |
| $x'(i)$ | climate variable (dust events) before peak definition |
| $x(i)$ | climate variable (dust events) after peak definition |

where $C_b(i)$ is the time-dependent background concentration:

$$C_b(i) \quad = \quad \frac{N_b(i)}{N_p(i) + \bar{N}_p(i) + N_b(i) + \bar{N}_b(i)}.$$

We consider first the case of a single dust event recorded in segment $i$. For a very short duration, $D_p(i)_{j=1} \ll D(i)$, in addition to the background we have also Saharan input of material ($Ca^{2+}$ and non–$Ca^{2+}$). The flux of Saharan-borne $Ca^{2+}$ ions is

$$F(i)_{j=1} \quad = \quad \frac{N_p(i)_{j=1}}{A(i)\ D_p(i)_{j=1}}$$

where $N_p(i)_{j=1}$ is the number of such ions, here $N_p(i)_{j=1} = N_p(i)$. The area $A(i)$ shall include not only the right-angle projection but also other kinds of "susceptibility" of the then exposed segment to the flux. For example, the core site might had been elevated slowly in time, resulting in a better exposure. The flux of background non–$Ca^{2+}$ ions is

$$\bar{F}(i) \quad = \quad \frac{\bar{N}_b(i)}{\bar{A}(i)\ D(i)}.$$

Then we have from Equation 1

$$C(i) - C_b(i) \quad \simeq \quad \frac{F(i)_{j=1}\ A(i)\ D_p(i)_{j=1}}{\bar{F}_b(i)\ \bar{A}(i)\ D(i)}. \qquad (2)$$

**Proposition 2:** The "susceptibility" ratio is constant,

$$A(i)/\bar{A}(i) \quad = \quad \text{const.}$$

This is certainly fulfilled to a good order of approximation because the dust is scavenged by precipitation (Wagenbach *et al.* 1996), that means, Saharan-derived $Ca^{2+}$ and background-derived non–$Ca^{2+}$ is transported by the same carrier ("dirty water") into the exposed segment. From Equation 2 we have then the approximate relationship

$$[C(i) - C_b]\, D(i) \quad \propto \quad \frac{F(i)_{j=1}\ D_p(i)_{j=1}}{\bar{F}_b(i)}.$$

The generalisation (several events) is straightforward: approximately,

$$[C(i) - C_b]\, D(i) \quad \propto \quad \frac{\sum_j F(i)_j\ D_p(i)_j}{\bar{F}_b(i)}. \qquad (3)$$

**Proposition 3:** The background water flux is constant,

$$\bar{F}_b(i) \quad = \quad \text{const.}$$

This is certainly not fulfilled exactly. However, we can be confident that the variation of the numerator of the right-hand side of Equation 3 is considerably larger (peak vs. no peak) than the variation of $\bar{F}_b(i)$; Wagenbach (1989) estimates $\bar{F}_b(i)$ to be 0.2–0.4 (arbitrary units) for the core site. Of course, in cases where $\bar{F}_b(i)$ can be measured (by an accurate dating), a corresponding normalization can be made. Proposition 3 and Equation 3 give the transformed values,

$$x'(i) \quad \equiv \quad [C(i) - C_b(i)]\, D(i) \tag{4}$$

which, to a fairly well approximation, should reflect the climatic signal of the dust events,

$$x'(i) \quad \propto \quad \sum_j F(i)_j \; D_p(i)_j .$$

In terms of physics, the transformation corrects for the dilution of peak values by background values, the degree of the dilution depending on $D(i)$.

# 3 Background estimation

Equation 4 demands the estimation of the time-dependent background signal, $C_b(i)$. We wish also to count as dust peaks only those values that lie significantly above the background. I propose to use a window which contains a constant number of $k$ points and is being shifted pointwise along the time axis.

The time-dependent background is estimated by the median of the points inside the window:

$$\begin{aligned}\hat{C}_b(i) \quad &= \quad med(i) \\ &= \quad \text{median}\,[C(i-K), \ldots, C(i+K)]\,, \quad i = K+1, \ldots, N-K,\end{aligned}$$

where $k = 2K + 1$ ($k$–nearest-neighbour median smoothing). Generated pseudodata outside the interval (Cowling and Hall 1996) can be used for extending the range to $i = 1, \ldots, N$.

To define a peak, we use the *mad*,

$$mad(i) \;=\; \text{median}\left[|C(i-K) - med(i)|, \ldots, |C(i+K) - med(i)|\right].$$

A peak is defined by the rule

$$\begin{aligned} \text{if } C(i) > med(i) + z\, mad(i) \;&\Rightarrow\; \text{peak at } T(i) \text{ with magnitude} \\ &\quad\;\; X(i) = [C(i) - med(i)]\, D(i), \\ \text{else} \;&\Rightarrow\; \text{no peak at } T(i) \end{aligned}$$

(cf. Equation 4).

The choice of $z$ is partially subjective and knowledge about the data, for example possible background sources, will certainly help. Otherwise we may use Hampel's (1985) X84 rule ($z = 3.5$), and try also other values.

Robust measures such as *med* and *mad* have to be used since the distribution comprises two components (peak, background) and is highly non-normal.

Zheng and Yang (1998) have given a cross-validation criterion for selecting the smoothing parameter $k$ in case of median smoothing. One other argument, for using a small value of $k$, arises from the peak definition inside the window. Segments with large duration $D(i)$ will show a stronger dilution of a peak than those with small $D(i)$. A window represents $k$ segments and therefore a range of values of $D(i)$. If $k$ is too large, the dilution bias may become too strong. A plot of $D(i)$ vs. depth (Fig. 3) helps to asses this effect.

For large $y$, that means large $D(i)$, peak definition is therefore difficult. It might be possible to use adaptive smoothing ($k$ variable) but also this approach has drawbacks. The effective running median calculation algorithm of Härdle and Steiger (1995) would also not be applicable. More helpful could be to experiment with different values of $k$.

The proposed background estimation procedure applies also when there are missing data.

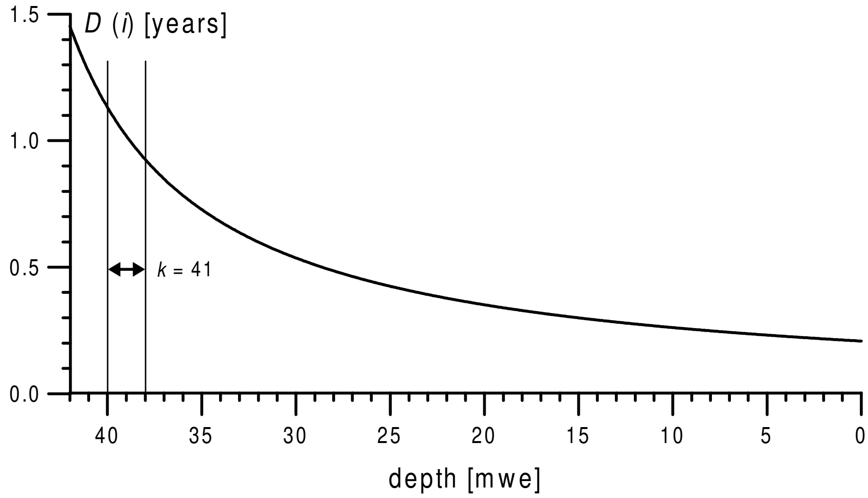


Figure 3: CCB ice core, duration against depth. This curve helps to evaluate the variable dilution by background values within a window.

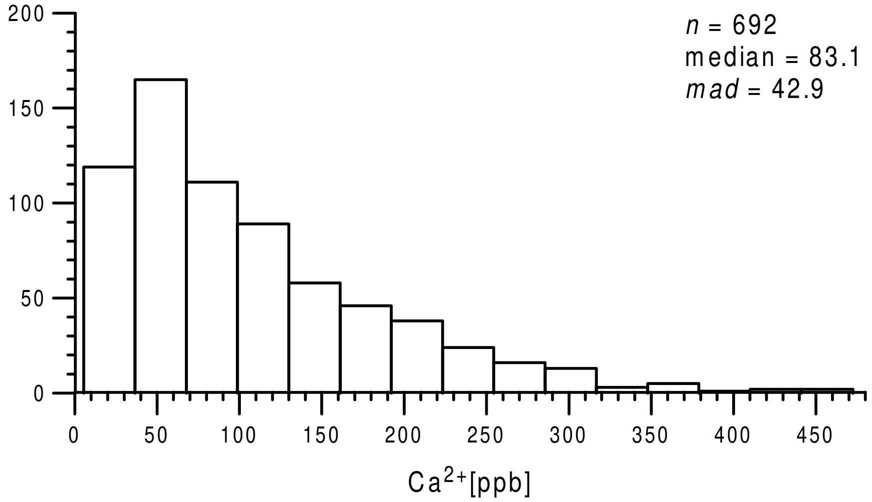


Figure 4: Histogram for the non-peak concentration values $C(i)$. The peaks were defined by median smoothing ($z = 3.5$, $k = 41$). The distribution helps to evaluate an upper limit of the statistical uncertainty of $\hat{C}_b(i)$.

# 4 Marked point process estimation

Given the Saharan dust peaks as $T(i)$ (average time of segment), $D(i)$ (duration) and $X(i)$ (climate variable magnitude), the main climatological interest concerns—in general and not only in case of the CCB record—a trend in the frequency of the events. Measuring the shape of such a trend would help to assess the natural climate variability and a potential human influence. An error band for the estimated trend is important to interpret results.

For using not only the event dates $T(i)$ but also the climatic information of the magnitude, $X(i)$, I propose to divide the events into a number of (equidistant) classes of magnitude. The various sources of uncertainty of the data suggest this. (1) An upper limit of the statistical error of $\hat{C}_b(i)$ may be evaluated with the help of the distribution (Fig. 4) of the non-peak calcium concentration values. (2) The choice of $z$ for peak definition leads to a systematic error. (3) $T(i)$ and $D(i)$ have errors due to the uncertain timescale (Section 1). (4) Proposition 3, $\bar{F}_b(i) =$ const., is not fulfilled exactly (Section 2). (5) The concentration values show measurement errors; however, these are small. From the statistical viewpoint, too many classes would imply too less events of the respective magnitude class and, hence, too large error bands. It should be useful to experiment with the total number of classes, for example with 1 (unmarked), 2 and 3 classes, to assess that dependency.

The event times are analysed as realisations of a time-dependent point process (e. g. Cox and Lewis 1966; Cox and Isham 1980). The fact that the ice core had been sampled as segments of nonzero length causes the "segmentation problem", that is, the problem of estimating whether a given segment (value $X(i)$) contains several events of minor class of magnitude or, instead, only a few events of higher magnitude. I suggest the following algorithm for solving that problem.

Let the classes be denoted by roman numbers, with a high number meaning a high magnitude.

Step 1

Estimate $\lambda_{\mathrm{I}}^{(1)}(t)$, the time-dependent intensity of a point process for class–I events. The superscript (1) indicates that this is the first estimation, before any subdivision of a class–I event has been made (see step 2). Estimation may be carried out by kernel intensity estimation (Diggle 1985). Boundary effects may be reduced by the generation of pseudodata outside the time interval (Cowling and Hall 1996). The cross-validation bandwidth selector by Brooks and Marron (1991) may be used. (The error introduced by setting the time of an event equal to $T(i)$ is small provided that $D(i)$ is small against the bandwidth.) Estimate $\lambda_{\mathrm{II}}^{(1)}(t), \lambda_{\mathrm{III}}^{(1)}(t), \ldots$ similarly.

Step 2

For all segments $i$ with class–II peaks: calculate the probability $p$ of two class–I peaks within the segment. This is carried out using the values of $D(i)$ and $\hat{\lambda}_{\mathrm{I}}^{(1)}(t = T(i))$ and the Poisson distribution. (If the event times are later found to be dependent, the calculations have to be altered accordingly and repeated.) Calculate the probability $q$ of one class–II event within the segment. If $p > q$: subdivide the class–II peak at $T(i)$ into two class–I peaks at $T(i)$. (The equidistance of the magnitude classes allows the subdivision.)

Step 3

New estimation is carried out: $\hat{\lambda}_{\mathrm{I}}^{(2)}(t)$ and $\hat{\lambda}_{\mathrm{II}}^{(2)}(t)$.

Step 4

For all segments $i$ with class–III peaks: calculate the probability $p$ of three class–I peaks within the segment, the joint probability $q$ of one class–I and one class–II peak within the segment and the probability $r$ of one class–III peak within the segment. Make an eventual subdivision of the class–III peak into smaller peaks at $T(i)$ with respect to the highest of those probabilities.

Step 5

New estimation is carried out: $\hat{\lambda}_{\mathrm{I}}^{(3)}(t)$, $\hat{\lambda}_{\mathrm{II}}^{(3)}(t)$ and $\hat{\lambda}_{\mathrm{III}}^{(2)}(t)$.

Step 6

Repeat the double-step subdivision–new estimation for the other magnitude classes.

Step 7

Repeat steps 1–6 until $\hat{\lambda}_{\mathrm{I}}(t)$, $\hat{\lambda}_{\mathrm{II}}(t), \ldots$ remain unchanged (no subdivision). (We then omit the superscript.) I assume quick convergence.

The result are the intensity trends for the different magnitude classes. Confidence bands can be constructed with bootstrap methods (Cowling *et al.* 1996). There is potential for developing bootstrap methods which account also for the uncertainty of the $T(i)$; eventually, the subdivision of peaks by the algorithm (steps 2 and 4) can be simulated also. Additional investigations should include statistical tests of the "no trend" hypothesis (e. g. Cox and Lewis 1966; Loader 1992) and quantification of a serial dependency between events (e. g. Cox and Lewis 1966).

In the case of missing data, the timescale $T(i)$ has to be shortened accordingly before the estimation.

# 5 Acknoweledgements

I thank Dietmar Wagenbach and Marko Scholze (Institute for Environmental Physics, University of Heidelberg, FRG) for discussions and release of the data. This study has been carried out while the author was Marie Curie Research Fellow (EU–TMR grant ERBFMBICT 971919).

## 6 References


Brooks MM, Marron JS (1991) Asymptotic optimality of the least-squares cross-validation bandwidth for kernel estimates of intensity functions. Stochastic Processes and their Applications 38:157–165.

Cowling A, Hall P (1996) On pseudodata methods for removing boundary effects in kernel density estimation. Journal of the Royal Statistical Society Series B 58:551–563.

Cowling A, Hall P, Phillips MJ (1996) Bootstrap confidence regions for the intensity of a Poisson point process. Journal of the American Statistical Association 91:1516–1524.

Cox DR, Isham V (1980) Point Processes. Chapman and Hall, London, 188 pp.

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

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

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

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

Loader CR (1992) A log-linear model for a Poisson process change point. The Annals of Statistics 20:1391–1411.

Oeschger H, Langway CC, Eds. (1989) The Environmental Record in Glaciers and Ice Sheets. Wiley, Chichester, 401 pp.

Wagenbach D (1989) Environmental records in Alpine glaciers. In: Oeschger H, Langway CC, Eds., The Environmental Record in Glaciers and Ice Sheets. Wiley, Chichester, 69–83.

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.

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