Geophys. J. R. astr. Soc. (1978) 52, 205–213

An objective method for smoothing palaeomagnetic data

R. M. Clark *Department of Mathematics, Monash University, Clayton, Victoria, Australia 3168*

R. Thompson *Department of Geophysics, University of Edinburgh, Edinburgh EH9 3JZ*

Received 1977 March 14

Summary. This paper describes a new method of smoothing noisy data, such as palaeomagnetic directions, in which the optimum degree of smoothing is determined objectively from the internal evidence of the data alone. As well as providing a best-fitting smooth curve, the method indicates, by means of confidence limits, which oscillations or fluctuations in the fitted curve are real. The procedure, which is illustrated by an analysis of palaeomagnetic declination directions from Lake Windermere, has potential applications throughout the Earth Sciences. It may be used in any investigation requiring the estimation of a smooth function from noisy data, provided certain basic assumptions are reasonably satisfied.

1 Introduction

Palaeomagnetic directions are dispersed by geomagnetic secular changes and random errors, such as orientation and measurement errors, and secondary magnetizations. Time-averaged directions of the ancient geomagnetic field have been accurately calculated from the mean directions of large collections of several hundred samples. Early palaeomagnetic results produced a time-average geocentric axial dipole field model. However, as palaeomagnetic results have increased in quality and quantity a number of small deviations from the geocentric axial dipole model have been consistently found. Cox (1975) has recently outlined the persistence, symmetry and importance of these geomagnetic nondipole components as deduced from studies of continental lava flows. Lake sediments offer an environment for obtaining ordered records of these geomagnetic features on a timescale of 10^3–10^5 yr. On account of the inherent scatter of palaeomagnetic directions some averaging is needed to estimate reliably the ancient geomagnetic field directions. However, too much smoothing of the time series obscures the secular changes of the geomagnetic field which the investigator is striving to elucidate. Smoothing supposedly equivalent palaeomagnetic direction data sets, with unequal signal-to-noise ratios, is a particularly common problem which becomes highly subjective. For example, the lower limit of detectibility of geomagnetic inclination

variation in Lake Windermere has been placed at the low-amplitude (3°) variations at 340 and 360 cm depth (Thompson 1973, Fig. 2) by Creer & Kopper (1974), whereas Thompson (1975) suggests a variation of over 6° amplitude, such as between 100 and 140 cm, was a more realistic lower limit. A more extreme example stems from Swedish sediments where Noel (1975, Fig. 6) noted 'A striking feature of the Starno postglacial profile is the presence of 3½ regular cycles in magnetic inclination with a 4° amplitude'. In contrast Thompson (1976) considers 'The Starno Post-Glacial (inclination) data are ... simply ... scatter', implying a lower limit of detectability in excess of 20°, due largely to the paucity of Noel's data. An objective statistical treatment based on the continuity of palaeomagnetic directions is clearly needed to establish the best-fitting function.

In this paper, we illustrate such a curve-fitting method, given by Clark (1977), by applying it to a set of palaeomagnetic declination directions (Fig. 1) from Lake Windermere, measured by Mackereth (1971).

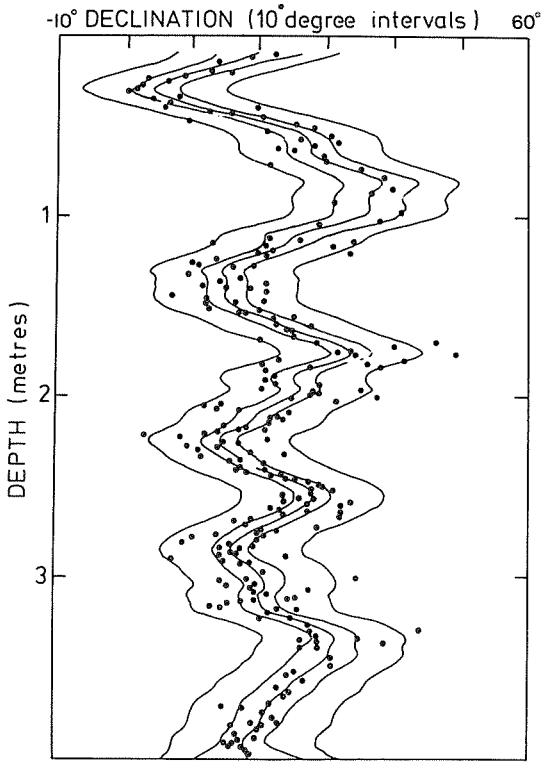

Figure 1. Declination as a function of depth of sediment at Lake Windermere. Central curve is the best-fitting smoothed estimate of the true function F relating declination to depth, as derived by the method described in this paper. The two extreme curves are the 95 per cent prediction limits for future independent observations of declination; the remaining two curves are the 95 per cent confidence limits for F.

2 Mathematical model and assumptions

The given sequence of paired measurements $(x_i, y_i), i = 1, \ldots, n$ of depth and declination are assumed to satisfy the model

$$y_i = F(x_i) + e_i \quad i = 1, \ldots, n \tag{1}$$

where x_i denotes the measured depth of the ith sample from the core and y_i the corresponding measured declination. The variable e_i represents the overall error of measurement in the observation y_i, that is, the net effect of errors due to instrumental noise, sample orientation and sedimentological variability. The $\{e_i\}$ are assumed to be independent normally distributed random variables with zero mean and constant variance σ^2 (initially unknown), while the $\{x_i\}$ are assumed to be known exactly, with *no* error of measurement. The function F represents the underlying idealized functional relationship between declination and depth for Lake Windermere, describing implicitly the secular variation of the geomagnetic field. In other words, at each depth x, $F(x)$ denotes the true geomagnetic declination, relative to the chosen reference direction, at Lake Windermere at the time corresponding to that depth. Our only assumption concerning the function F is that it is 'smooth'; our aim is to estimate, and place confidence limits on, this unknown function, under this minimal assumption.

These assumptions, although an approximation to the real world, are readily justified, and seem reasonable in practical terms. For example, although there would be some error of measurement, however small, in the observed depths $\{x_i\}$, any such errors would be negligible compared with the overall errors in the declinations, and so the former may be safely ignored. Similarly, while it is most unlikely that the $\{e_i\}$ are *exactly* normally distributed, it is likely, essentially because of the Central Limit Theorem (Cramér 1946, p. 231), that their probability distribution would be a very close approximation to the normal distribution. The assumption of normality, or equivalently, of a Fisherian distribution (Fisher 1953) with high concentration parameter, has been considered acceptable for typical palaeomagnetic data (e.g. Collinson, Creer & Runcorn 1967; McElhinny 1973).

In general, the precision of the declination measurements, or equivalently, the variance of the residual errors $\{e_i\}$, need not necessarily be the same for all samples from a given core. For instance, sedimentological processes such as micro-slumping, differential compaction or diagenesis need not be uniform along the core, implying that errors due to these factors alone may be greater in some parts of the core than others. In such cases, the smoothing procedure described below may be readily modified to account for these or similar effects, provided the *ratios* of the error variances to one another are known, or can be reasonably estimated. In the present case of the Windermere declination record, there was no *a priori* reason to suggest that any particular measurements of the declination would be less precise than others. Subsequently statistical analyses confirmed that the assumption of constant error variance was indeed justifiable for this particular set of data.

We make no assumptions concerning the form of F, the underlying empirical function or curve relating declination to depth. For instance, we do *not* assume that F is a polynomial, or a sum of exponential terms. Our assumption that F is 'smooth', i.e. differentiable, reflects a basic tenet of geophysics, namely that past secular variations in geophysical parameters (such as geomagnetic directions) have been smooth and gradual.

3 Smoothing procedure

When smoothing a given set of data, one must decide on (i) the *method* of smoothing and (ii) the *degree* or *extent* of smoothing. We consider these issues separately.

3.1 METHOD OF SMOOTHING

We consider four possible methods of smoothing, namely, by fitting (i) a polynomial of appropriate degree by least-squares, (ii) an interpolating cubic spline to smoothed points

obtained by Fourier-filtering (Thompson & Kelts 1974), (iii) an approximating cubic spline directly (Reinsch 1971; Greville 1969; Schoenberg 1964), and (iv) the estimator developed by Clark (1977). The first method was rejected principally because polynomials are generally unsatisfactory as approximations to empirical functions (Rice 1969, p. 123). The second method is intuitively appealing, but has the disadvantage that the choice of the parameters in the Fourier-filtering was arbitrary. Further, it would appear that this method would not have the optimal properties of method (iii).

Although there are strong theoretical arguments in favour of the third method, we in fact used the fourth method, based on convolution-smoothing of a first-order interpolating spline. The properties of this method are similar to those of the optimal approximating cubic spline of Reinsch & Schoenberg, particularly for moderate degrees of smoothing, but the new method is simpler and quicker computationally. This reduction in computing effort can be considerable, especially when a large number of smoothed curves must be derived in a trial-and-error process, as in the cross-validation method described below.

The smoothing formula given by Clark (1977) may also be regarded as a special type of moving average, as it can be written as

$$f(x) = \sum u_i(x) y_i \tag{2}$$

where $f(x)$ is the smoothed estimate of $F(x)$, the 'theoretical' declination at depth x, and the $\{u_i\}$ are known (piecewise polynomial) functions of x. The extent of the smoothing is determined by a single parameter called the *bandwidth*. In general, the only non-zero terms in the summation (2) are those corresponding to observations within roughly one bandwidth of the current depth x; in short, the larger the bandwidth, the heavier the smoothing.

3.2 DEGREE OF SMOOTHING

The *degree* of smoothing of the data is more crucial than the method of smoothing to be employed. If the smoothing is too heavy, some of the fine detail of F may be lost, while too little smoothing may produce spurious kinks or undulations as the fitted curve 'chases' the random noise in the data. The *cross-validation* method is a remarkably adaptable and successful technique for determining objectively the appropriate degree of smoothing, using only the internal evidence of the data.

To determine the appropriate bandwidth for smoothing the Windermere declination record, we used a modification, described in Clark (1977), to the cross-validation technique used for example by Wahba & Wold (1975) and reviewed by Stone (1974). Briefly, the method is as follows. Ten per cent of the observations, chosen at random by computer are temporarily deleted from the record, and a smooth curve is fitted to the remaining points, using (2) and a given bandwidth. One then examines collectively the differences between the declination predicted by this curve and the actual observed declination at each of the points originally deleted. This procedure is then repeated for different random selections and different bandwidths. The optimum bandwidth is then the one which minimizes the average sum of squares of the differences between 'observed' and 'predicted' declinations. Further details are given in Clark (1977, Section 4).

Experiments with artificially-generated data reported by Clark (1977) indicate that this method is remarkably effective in finding the 'right' degree of smoothing in a variety of situations. If the data are 'good', i.e. the signal-to-noise ratio is relatively high and/or the spacing between abscissae is relatively small, our smoothing procedure (using (2) and the optimum bandwidth) will recover most of the fine structure of the underlying function F.

If the data are 'poor', only the main trend of F can be recovered, as expected. On the other hand, the method does not produce spurious kinks, in either situation.

In principle, the method of cross-validation can be used to determine the right degree of smoothing for *any* method of smoothing. For example, Wahba & Wold (1975) use cross-validation to determine the smoothing parameter (their λ) of Reinsch's approximating cubic spline. Their very extensive simulation experiment gives compelling evidence of the effectiveness of the cross-validation procedure.

In addition, the method of cross-validation enables one to estimate σ^2, the variance of the error terms $\{e_i\}$, measuring the scatter of the observations about the true curve F (see Appendix A). As well as being of intrinsic interest, this estimate is essential for the derivation of confidence limits and prediction limits (see Appendix B).

4 Summary of results

The results of cross-validation smoothing of the declination measurements from Lake Windermere are summarized in Figs 1 and 2. The confidence limits and prediction limits should be interpreted as follows. At any given depth x_0, one can be 95 per cent sure that a future *independent* observation of declination at that depth will lie within the given prediction limits. Similarly, the true geomagnetic declination $F(x_0)$ at depth x_0 lies, with 95 per cent probability, within the given confidence limits for that depth. For example, at a depth of one metre, we can be 95 per cent sure that the true declination lies between 31 and 41° (relative to our arbitrary reference direction), while a future independent measurement of declination at that depth should lie between 25 and 47°, with 95 per cent probability.

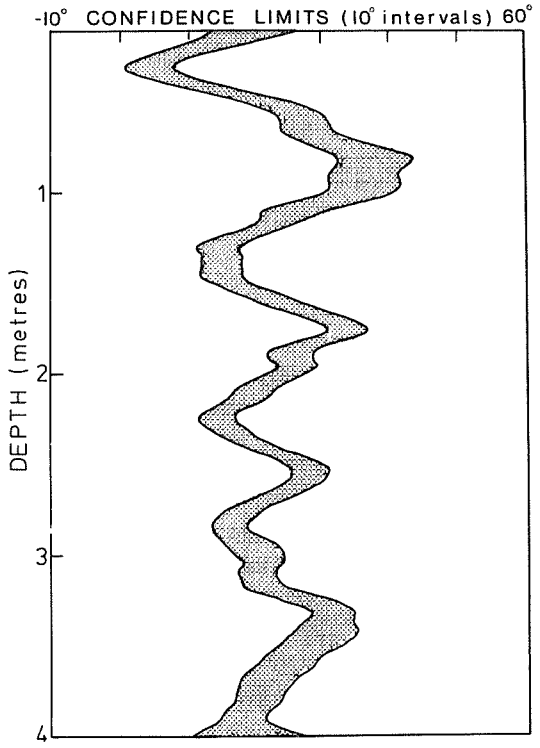

Figure 2. 95 per cent confidence band for the true functional relationship F between declination and depth at Lake Windermere.

This does not necessarily mean that 95 per cent of our original observations should be within the prediction limits, since these limits only apply to measurements whose random components are statistically independent of the random error in the original measurements. Experience with artificially-generated data indicates that on average about 97 or 98 per cent of the original measurements should lie within these prediction limits. In fact, 241 or 96.4 per cent of our 250 measurements lie within the prediction limits of Fig. 1.

Although the central curve of Fig. 1 is our best point-estimate of the function F in (1), the unknown function F could lie anywhere within the shaded confidence band of Fig. 2, with 95 per cent probability. It follows that minor oscillations or kinks, such as those at 190 or 310 cm, are not necessarily real, and may simply be the result of the particular random errors in our particular observations. However, as judged by the confidence band of Fig. 2, the major fluctuations at 30, 90, 135, 180, 230, 260 and 340 cm are clearly reproducible, and may be realistically considered as of geomagnetic origin.

The overall random scatter of the declination measurements about the true curve F corresponds to a standard deviation estimated as in Appendix A to be 5° (approximately). As noted previously, these overall errors of measurement, represented by $\{e_i\}$ in (1), are due to instrumental noise, variations in sample orientation, and variations in the sediment. Instrumental error is small compared with other sources of variation and in this example can be taken as having a standard deviation of 1° (Gough 1967; Collinson 1975). Orientation errors are more difficult to estimate, but will be less than in conventional palaeomagnetic investigations and the standard deviation should not exceed 3° (Collinson 1975; Gough 1967). It follows that sedimentological variability corresponds to a standard deviation of approximately 4° (assuming that these three factors are statistically independent of one another).

5 Prospect

The Lake Windermere palaeomagnetic directions provide the longest, most continuous and well-dated record in Europe and are consequently used as a 'master dating curve'. The cross-validation curve-fitting, as well as providing the best-fitting master curve, indicates which oscillations are 'real'. Significant oscillations, when found in separate cores, can be realistically interpreted as of geomagnetic origin. The cross-validation best-fitting curve can be used as a master curve to which other undated data groups can be correlated and hence dated. The corresponding confidence limits clearly show that at the 95 per cent probability level certain local minor fluctuations cannot be justified by the data set, and so cannot be used for modelling past changes of the geomagnetic field or as a dating tool.

To specify completely the direction in space of the geomagnetic field, both the declination and inclination must be given. Measurements of the inclination at Lake Windermere are available (Thompson 1973), covering the same period of time as the declination data of Fig. 1. At first glance, it seems not unreasonable to smooth these inclination measurements independently of, but by the same method as, the declinations, and to use the smoothed curves for declination and inclination jointly as master curves describing past variation in magnetic direction. However, recent work (to be reported elsewhere) has shown that such separate smoothing can be very misleading, since the final results may depend critically on the choice of reference directions. Since the basic measurements are of a direction in space, that is, a vector, the data should be treated as such, and the declinations and inclinations should be smoothed simultaneously. Work is currently in hand on the necessary modifications to the smoothing procedures of this paper to do this. Unfortunately, the measurements of declination and inclination presently available from Lake Windermere

cannot be paired with one another, as these were obtained independently. Simultaneous smoothing is thus not possible with the available data from Lake Windermere, but investigations are in progress to obtain the relevant paired measurements of declination and inclination.

As more data become available, we intend to develop the smoothing method of this paper (i) to test whether two or more independent data sets are recording the same signal, and (ii) to match an undated sequence of lacustrine declinations and inclinations against a master curve, allowing for possible differences in sedimentation rates, thereby dating the unknown sequence. With these objective methods of smoothing and comparing groups of noisy data, we will thus be able further to refine our descriptive model of past geomagnetic field changes and to check the underlying geophysical assumptions.

Many investigations in the Earth Sciences, such as palaeomagnetism, involve the estimation of some unknown smooth function from noisy data satisfying the model (1). We believe that these investigations could benefit significantly from the smoothing procedures outlined in this paper.

References

Clark, R. M., 1977. Non-parametric estimation of a smooth regression function, *J. R. stat. Soc. B*, **39**, 1, 107–113.

Collinson, D. W., 1975. Instruments and techniques in paleomagnetism and rock magnetism. *Rev. Geophys. Space Phys.*, **13**, 5, 569–686.

Collinson, D. W., Creer, K. M. & Runcorn, S. K., 1967. *Methods in palaeomagnetism*, Elsevier, Amsterdam.

Cox, A., 1975. The frequency of geomagnetic reversals and the symmetry of the nondipole field, *Rev. Geophys. Space Phys.*, **13**, 35–51.

Cramér, H., 1946. *Mathematical methods of statistics*, Princeton University Press.

Creer, K. M. & Kopper, J. S., 1974. Palaeomagnetic dating of cave paintings in Tito Bustillo Cave, Asturias, Spain, *Science*, **186**, 348–350.

Fisher, R. A., 1953. Dispersion on a Sphere, *Proc. R. Soc. Lond. A.*, **217**, 295–305.

Gough, D. I., 1967. Notes on rock sampling for palaeomagnetic research, *Methods in palaeomagnetism*, pp. 3–7, eds Collinson, D. W., Creer, K. M. & Runcorn, S. K., Elsevier, Amsterdam.

Greville, T. N. E., 1969. Introduction to spline functions, *Theory and applications of spline functions*, pp. 1–36, ed. Greville, T. N. E., Academic Press, New York.

Mackereth, F. J. H., 1971. On the variation in direction of the horizontal components of remanent magnetization in lake sediments, *Earth planet. Sci. Lett.*, **12**, 332–338.

McElhinny, M. W., 1973. *Palaeomagnetism and plate tectonics*, Cambridge University Press.

Noel, M., 1975. The Palaeomagnetism of varved clays from Blekinge, Southern Sweden, *GFF*, **97**, No. 4, 357–367.

Reinsch, C. H., 1971. Smoothing by spline functions. II, *Numer. Math.*, **16**, 451–454.

Rice, J. R., 1969. *The approximation of functions*, Vol. II, Addison, Wesley, Reading, Massachusetts.

Schoenberg, I. J., 1964. Spline functions and the problem of graduation, *Proc. Nat. Acad. Sci.*, **52**, 947–950.

Stone, M., 1974. Cross-validatory choice and assessment of statistical predictions, *J. R. stat. Soc. B*, **36**, 111–147.

Thompson, R., 1973. Palaeolimnology and palaeomagnetism, *Nature*, **242**, 182–184.

Thompson, R., 1975. Long period European geomagnetic secular variation confirmed, *Geophys. J. R. astr. Soc.*, **43**, 847–859.

Thompson, R., 1976. The palaeomagnetism of varved clays from Blekinge, Southern Sweden. A Comment. *GFF*, **98**, No. 3, 283–284.

Thompson, R. & Kelts, K., 1974. Holocene sediments and magnetic stratigraphy from Lakes Zug and Zurich, Switzerland, *Sedimentology*, **21**, 577–596.

Wahba, G. & Wold, S., 1975. A completely automatic French curve: fitting spline functions by cross-validation, *Comm. Stat.*, **4**, 1–17.

Appendix A: estimation of the error variance σ^2

We use the same notation as Clark (1977). For a given bandwidth b, consider a particular random partition P of the data into an *estimation sample* of n_1 observations (re-numbered (x_{1i}, y_{1i}), $i = 1, 2, \ldots, n_1$) and a *validation sample* of n_2 observations (re-numbered (x_{2i}, y_{2i}), $i = 1, 2, \ldots, n_2$), with $n_1 + n_2 = n$, and in our case, $n_2 = 0.1n$. The corresponding estimate of F is

$$f^*(x) = \sum_{j=1}^{n_1} u_{1j}(x) y_{1j} \tag{A1}$$

where the subscript 1 on the u functions denotes that these functions (and consequently f^*) are computed from the estimation sample only. The aim of the cross-validation procedure is to minimize (as a function of b) $\bar{C}(b)$, the average over m partitions P of

$$C(b, P) = \frac{1}{n_2} \sum_{i=1}^{n_2} \{y_{2i} - f^*(x_{2i})\}^2. \tag{A2}$$

For this partition, we define parameters $b_i, g_i, i = 1, 2, \ldots, n_2$, such that

$$\text{Var}\{f^*(x_{2i})\} = g_i \sigma^2$$

$$b_i = E\{f^*(x_{2i})\} - F(x_{2i}) = \sum_{j=1}^{n_1} u_{1j}(x_{2i}) F(x_{1j}) - F(x_{2i})$$

where E denotes expectation and Var denotes variance. The $\{g_i\}$ are easily computed, since by (A1) and the assumptionsof our model,

$$g_i = \sum_{j=1}^{n_1} \{u_{1j}(x_{2i})\}^2.$$

The parameters $\{b_i\}$, representing the *bias* of $\{f^*(x_{2i})\}$, are unknown, as they depend on the unknown function F.

Since, by the assumptions of model (1), y_{2i} and $f^*(x_{2i})$ are independent,

$$E[\{y_{2i} - f^*(x_{2i})\}^2] = \sigma^2(1 + g_i) + b_i^2, \quad i = 1, 2, \ldots, n_2$$

giving

$$E\{C(b, P)\} = \sigma^2\{1 + \bar{g}(P)\} + \bar{B}(P), \tag{A3}$$

where $\bar{g}(P)$ and $\bar{B}(P)$ denote the average for this partition of the $\{g_i\}$ and $\{b_i^2\}$ respectively.

Finally, averaging (A3) over the m partitions P, we obtain

$$E\{\bar{C}(b)\} = \sigma^2\{1 + \bar{\bar{g}}(b)\} + \bar{\bar{B}}(b) \tag{A4}$$

where $\bar{\bar{g}}(b)$ and $\bar{\bar{B}}(b)$, the averages over P of $\bar{g}(P)$ and $\bar{B}(P)$, are implicit functions of the bandwidth b.

Equation (A4) expresses the mean or expected value of $\bar{C}(b)$ as the sum of two initially unknown factors. The first factor arises from the random scatter in the observations $\{y_i\}$, while the second allows for possible systematic discrepancies between the fitted curves $f^*(x)$ and the true curve $F(x)$. Even if the $\{y_i\}$ could be measured *exactly*, that is, the $\{e_i\}$ in (1) were identically zero, the fitted curve need not necessarily coincide exactly with the true curve.

In general, as b decreases, $\bar{\bar{g}}$ increases while $\bar{\bar{B}}$ decreases. By Lemma 2 of Clark (1977), the bias terms $\{b_i\}$ are zero if, roughly speaking, F is linear throughout suitable intervals of

length approximately $2b$. If F is differentiable, as assumed, it may be approximated by straight lines over sufficiently narrow intervals, and so $\bar{\bar{B}}$ will be arbitrarily close to zero if b is sufficiently small.

Accordingly, we estimate σ^2 by solving the equation

$$\bar{C}(b) = \hat{\sigma}^2\{1 + \bar{\bar{g}}(b)\} \tag{A5}$$

for some suitably 'small' value of b. We then estimate the bias factor $\bar{\bar{B}}(\hat{b})$ corresponding to the optimum bandwidth $\hat{b}$ by solving

$$\bar{C}(\hat{b}) = \hat{\sigma}^2\{1 + \bar{\bar{g}}(\hat{b})\} + \bar{\bar{B}}(\hat{b}). \tag{A6}$$

Example. For the Lake Windermere declinations, we find, for $b = 2.0$ cm, $\bar{C}(b) = 37.34$ with $\bar{\bar{g}}(b) = 0.552$, giving $\hat{\sigma}^2 = 24.06$ from (A5). For the optimum bandwidth $\hat{b} = 9.85$ cm, $\bar{C}(\hat{b}) = 29.4$, $\bar{\bar{g}}(\hat{b}) = 0.151$, giving $\bar{\bar{B}}(\hat{b}) = 1.71$.

Although the mean of the random variable $\bar{C}(b)$ is known and given by (A4), its sampling distribution is likely to be rather complex and difficult to evaluate. Simulation experiments indicate that equation (A5) leads to reasonable estimates of σ^2, being usually within 20 per cent of the correct value, provided n, the total number of observations, is at least 100.

Appendix B: construction of confidence intervals

The construction of a confidence interval for the ordinate $F(x_0)$ corresponding to a given abscissa x_0, is based on the difference

$$d_0 = f(x_0) - F(x_0) = \sum u_i(x_0)y_i - F(x_0) \tag{B1}$$

where the functions $u_i(.)$ from (2) are computed using the optimum bandwidth $\hat{b}$. Under our assumptions, d_0 is normally distributed with known variance $\sigma_0^2 = \sigma^2 \Sigma u_i^2(x_0)$ but with unknown mean

$$b_0 = E\{f(x_0)\} - F(x_0). \tag{B2}$$

Although b_0 is unknown, its magnitude should be similar on average (averaging over various choices of x_0) to the bias terms $\{b_i\}$ of Appendix A. Using the estimated mean-square $\bar{\bar{B}}(\hat{b})$ of these bias terms, we may estimate the mean-square-error of d_0, that is $\sigma_0^2 + b_0^2$, by

$$\sigma_F^2 = \sigma^2 \sum u_i^2(x_0) + \bar{\bar{B}}(\hat{b}). \tag{B3}$$

Hence we obtain approximate confidence limits for $F(x_0)$ by treating d_0 as normally distributed with zero mean and effective variance σ_F^2, with σ^2 and $\bar{\bar{B}}(\hat{b})$ being estimated from (A5) and (A6) respectively.

Any error in the resulting 95 per cent confidence interval from treating d_0 as normally distributed but with zero mean and variance equal to its mean-square-error is negligible, provided the ratio of the bias to the standard deviation, that is $|b_0|/\sigma_0$, is less than about 0.9. Since for the Windermere declination data the corresponding ratio $\{\bar{\bar{B}}/\hat{\sigma}_0^2\}^{1/2}$ lies mostly within this range, the confidence limits of Figs 1 and 2 should be sufficiently accurate for all practical purposes.