THE AUSTRALIAN NATIONAL UNIVERSITY

CENTRE FOR RESOURCE AND ENVIRONMENTAL STUDIES

CANBERRA

ANUSPLIN VERSION 4.36

USER GUIDE

M.F. Hutchinson

The ANUSPLIN package contains FORTRAN programs for fitting surfaces to noisy data as functions of one or more independent variables. The package includes programs for interrogating the fitted surfaces in both point and grid form. Procedures for calculating standard error surfaces have also been developed.

The programs are normally distributed as binary executables for:
Sun Microsystems Solaris 2.x on SPARC hardware
Silicon Graphics Irix 6.x on MIPS-3 hardware (or later)
Compaq Digital Unix on Alpha hardware
AIX5.x on IBM PowerPC
Linux on Intel or AMD hardware.
Microsoft Windows 95, 98, 2000, NT, ME, XP

Last revision to this document: 17 July 2006

The publishing program of the Centre for Resource and Environmental Studies at the Australian National University is designed to present the results of the Centre's research and the proceedings of conferences and workshops. Views expressed in CRES publications are the views of the authors and are not necessarily those of the Centre or any associated institution.

Director: Prof Will Steffen

Executive Officer: Sharon McInnes

ISBN 086740 512 0

All CRES publications are available from:

Publications Section
Centre for Resource and Environmental Studies
Australian National University
Canberra ACT 0200

Tel. +61 2 6125 4277

Fax +61 2 6125 0757

Email: publications@cres.anu.edu.au

URL: cres.anu.edu.au

TABLE OF CONTENTS

INTRODUCTION

The aim of the ANUSPLIN package is to provide a facility for transparent analysis and interpolation of noisy multi-variate data using thin plate smoothing splines. The package supports this process by providing comprehensive statistical analyses, data diagnostics and spatially distributed standard errors. It also supports flexible data input and surface interrogation procedures.

The original thin plate (formerly Laplacian) smoothing spline surface fitting technique was described by Wahba (1979), with modifications for larger data sets due to Bates and Wahba (1982), Elden (1984), Hutchinson (1984) and Hutchinson and de Hoog (1985). The package also supports the extension to partial thin plate splines based on Bates *et al*. (1987). This allows for the incorporation of parametric linear sub-models (or covariates), in addition to the independent spline variables. This is a robust way of allowing for additional dependencies, provided a parametric form for these dependencies can be determined. In the limiting case of no independent spline variables (not currently permitted), the procedure would become simple multi-variate linear regression.

Thin plate smoothing splines can in fact be viewed as a generalisation of standard multi-variate linear regression, in which the parametric model is replaced by a suitably smooth non-parametric function. The degree of smoothness, or inversely the degree of complexity, of the fitted function is usually determined automatically from the data by minimising a measure of predictive error of the fitted surface given by the generalised cross validation (GCV). Theoretical justification of the GCV and demonstration of its performance on simulated data have been given by Craven and Wahba (1979).

An alternative criterion is to minimise the generalised maximum likelihood (GML) developed by Wahba (1985,1990). It is based on a Bayesian formulation for the thin plate smoothing spline model and has been found to be superior to GCV in some cases (Kohn *et al*. 1991). Both criteria are offered in this version of ANUSPLIN.

A comprehensive introduction to the technique of thin plate smoothing splines, with various extensions, is given in Wahba (1990). A brief overview of the basic theory and applications to spatial interpolation of monthly mean climate is given in Hutchinson (1991a). More comprehensive discussion of the algorithms and associated statistical analyses, and comparisons with kriging, are given in Hutchinson (1993) and Hutchinson Gessler (1994). Recent applications to annual, monthly and daily precipitation data have been described by Hutchinson (1995, 1998ab) and Price *et al.* (2000). The book by Schimek (2000) provides a good overview of the subject of smoothing and non-parametric regression with extensive references.

It is often convenient, particularly when processing climate data, to process several surfaces simultaneously. If the independent variables and the relative weightings of the data are the same for each surface then many surfaces can be calculated for little more computation than one surface. ANUSPLIN allows for arbitrarily many such surfaces with significant savings in computation. ANUSPLIN also introduces the concept of "surface independent variables", to accommodate independent variables that change systematically from surface to surface. ANUSPLIN permits systematic interrogation of these surfaces, and their standard errors, in both point and grid form.

ANUSPLIN also permits transformations of both independent and dependent variables and permits processing of data sets with missing data values. When a transformation is

applied to the dependent variable ANUSPLIN permits back-transformation of the fitted surfaces, calculates the corresponding standard errors, and corrects for the small bias that these transformations induce. This has been found to be particularly convenient when fitting surfaces to precipitation data and other data that are naturally positive or non-negative.

A summary of the eight programs that make up the ANUSPLIN package is tabulated in the following section, accompanied by a flow chart showing the main connections between the programs. This is followed by detailed documentation for each program in the package. The User Guide concludes with a comprehensive discussion of example smoothing spline analyses of uni-variate data and multi-variate climate data. The data supporting these analyses are supplied with the package. These analyses can be used as a tutorial on the basic concepts of data smoothing, with particular applications to the spatial interpolation of climate.

PROGRAM SUMMARY

Table 1. The eight programs making up the ANUSPLIN package.

PROGRAM	DESCRIPTION
SPLINA	A program that fits an arbitrary number of (partial) thin plate smoothing spline functions of one or more independent variables. Suitable for data sets with up to about 2000 points although data sets can have arbitrarily many points. The degree of data smoothing is normally determined by minimising the generalised cross validation (GCV) or the generalised maximum likelihood (GML) of the fitted surface.
SPLINB	An approximate version of SPLINA designed for larger data sets and for data sets with missing data values. It uses knots initially selected by SELNOT that can optionally be updated by ADDNOT or DELNOT. Suitable for data sets with up to about 10,000 data points, with up to about 2000 knots, although data sets can have arbitrarily many points.
SELNOT	Selects an initial set of knots for use by SPLINB.
ADDNOT	Updates knot index file when additional knots are selected from the ranked residual list produced by SPLINB.
DELNOT	Adjusts knot index file when points are removed from the data file to be used by SPLINB.
GCVGML	Calculates the GCV or GML for each surface and the average GCV or GML over all surfaces, for a range of values of the smoothing parameter. It can be applied to surfaces fitted by either SPLINA or SPLINB. The GCV or GML values are written to a file for inspection and plotting.
LAPPNT	Calculates values and Bayesian standard error estimates of partial thin plate smoothing spline surfaces at points supplied in a file.
LAPGRD	Calculates values and Bayesian standard error estimates of partial thin plate smoothing spline surfaces on a regular rectangular grid.

The flow chart in Figure 1 shows the main data flows through the programs described in Table 1. The overall analysis proceeds from point data to output point and grid files suitable for storage and plotting by a geographic information system (GIS) and other plotting packages. The analyses by SPLINA and SPLINB produce output files that provide statistical analyses, support detection of data errors, an important phase of the analysis, and facilitate determination of additional knots by ADDNOT for SPLINB. The output surface coefficients and error covariance matrices enable systematic interrogation of the fitted surfaces by LAPPNT and LAPGRD. The GCV or GML files output by GCVGML can also assist detection of data errors and revision of the specifications of the spline model.

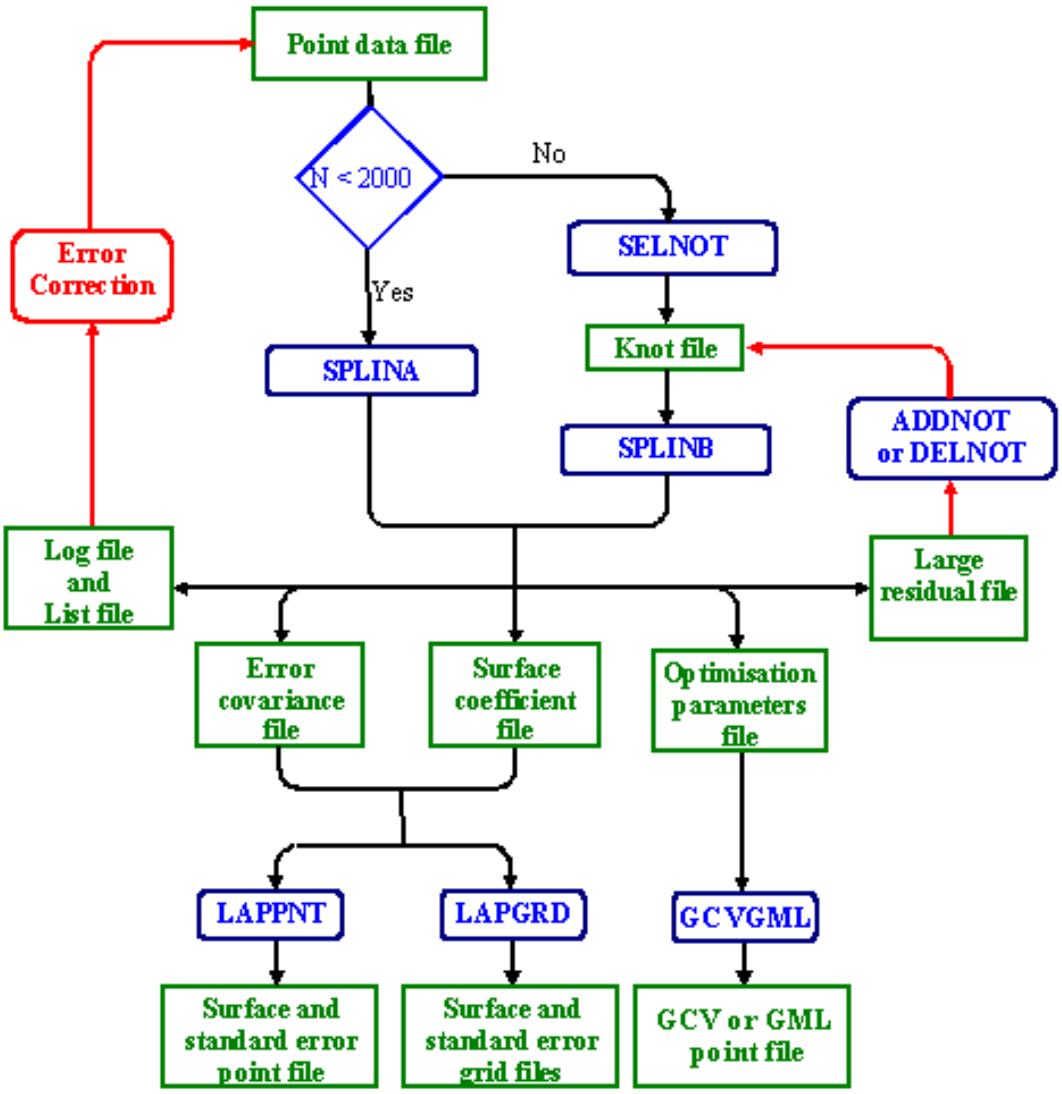

Figure 1. Main data flows through the ANUSPLIN package. N denotes the number of data points. Error covariance files are currently only produced by SPLINA. SPLINB may be applied to data sets with fewer than 2000 points to save computer time and storage space or to perform more robust analyses of poor data.

SPLINA and SPLINB

SPLINA is a FORTRAN 90 program that fits partial thin plate smoothing spline surfaces to multi-variate noisy data.

SPLINB is an approximate version of SPLINA designed for larger data sets. It fits partial thin plate smoothing spline surfaces, constructed from a set of knots, to multi-variate noisy data. It can also be applied to smaller data sets to save computer time and storage space. It can also be used to provide more robust analyses of poor data.

User directives are read from standard input. Users are strongly advised to use a command file for the user directives so that program output can be saved in an output log file. The log file provides a record of the directives supplied to the program and includes essential statistical analysis of the fitted surfaces, including large residuals and standard error estimates.

The program is best executed by starting it from a command-line shell. Under Unix, use any terminal emulator window showing a shell prompt. Under Microsoft Windows, start an MS-DOS shell or a command prompt window. To run splina, for example, type,

```
splina < job.cmd > job.log
```

where job.cmd is the input command file and job.log is the output log file.

Program Inputs

These include the numbers of independent spline variables and covariates, the lower and upper limits for each independent variable, optional transformations of each independent variable, and of the dependent variable, the order of derivative to be minimised, the number of surfaces, and the method to be used to determine the amount of data smoothing for each surface. Input and output file specifications, including the knot index file name in the case of SPLINB, are also required. Data points at positions that lie outside the user supplied independent variable limits are rejected. These limits can be used to fit a surface to a subset of the data without having to create a separate data file. These limits may include margins to allow for the development of overlapping surface patches that can be required for very large data sets. The user-supplied limits also give a simple check on the specified data format and the order of the independent variables in the data file. An error in these specifications would be indicated if fewer than the expected number data points were selected.

With the incorporation of standard FORTRAN 90 ANUSPLIN has dynamically allocated memory for most data and working arrays. Accordingly, both SPLINA and SPLINB can accommodate arbitrarily many surfaces fitted to arbitrary numbers of data points. However, it is advisable to limit the number of data points to no more than about 2000 data points for SPLINA, and to no more than about 10,000 data points. SPLINB may be applied to data sets with fewer than 2000 points provided the number of knots is sufficient to adequately approximate the fitted spline function. The main storage requirements for SPLINA are proportional to the square of the number of data points and the processing time is proportional to cube of the number of data points. Storage requirements for SPLINB are proportional to the square of the number of knots and processing time is approximately proportional to the cube of the number of knots, since a tri-diagonal decomposition of a matrix of this order is required. The required linear

algebra routines are contained within the double precision LINPACK library (Dongarra *et al*., 1979).

SPLINA does not permit data sets containing coincident data points or missing data values. SPLINB does permit data sets with coincident data points and/or missing data values. Coincident knots are not permitted.

Program Outputs

Summary statistics and a list of the largest residuals, ranked in descending order, are always written to standard output that should always be saved in an output log file. A list of the data and fitted values with Bayesian standard error estimates may also be written to an output list file. Files containing the coefficients of the fitted surfaces and the Bayesian error covariance matrices of the fitted surface coefficients may also be written. Optimisation parameters, that are used to determine the optimum smoothing parameter, may also be written. The surface coefficients and their error covariance matrices are used to calculate values and standard errors of the fitted surfaces by LAPGRD and LAPPNT. The ranked list of largest residuals is particularly useful in detecting data errors. The list of data and fitted values is also useful for this task, especially when fitting a surface to a new data set.

Knot Selection for SPLINB

The following, somewhat heuristic, procedure works well in practice:

1. Use SELNOT to choose an initial set of knots from the data points. This program selects a specified number of data positions as knots by attempting to equi-sample the independent spline variable space. While it is theoretically possible to choose knots that are not at data positions, this is not currently supported by SPLINB. In the limiting case where every data position is a knot, then the actions of SPLINA and SPLINB are identical. Choose an initial set of knots about 1/4 to 1/3 the size of the data set. The number of knots required depends on the spatial complexity of the data being fitted - more knots for a more complex surface. If the signal of the final result is within 10% of the number of knots (the maximum possible signal), then the process should be re-started with a larger initial knot set.

2. Run SPLINB, with the output list of data and fitted values, and examine the largest residuals for data errors. Re-fit the surface if necessary after data errors have been corrected. Use ADDNOT to add to the knot index file the indices of the largest 20-50 residuals that are not already knots and re-fit the surface with these additional knots. These indices may be read by ADDNOT from the large residual list output by SPLINB. Note that each knot index refers to the sequence number in the original data file and not the sequence number of the points actually selected. Knot indices may also be added to the knot index file using a text editor, but this is not generally recommended. If an editor is used, it is recommended that the increasing order of the sequence numbers in the knot index file be maintained. Residuals that are already associated with a knot are identified by a minus sign, both in the output ranked residual list, and in the list of data and fitted values. ADDNOT ignores these residuals when adding new knots to the set of knots.

3. Repeat the procedure of adding to the knot list the indices of the largest 20-50 residuals and re-fitting the surface until the solution stabilises or the variance estimates output by the program are in approximate agreement with *a priori* estimates. This should normally be done only once or twice, since there is a risk of overfitting to erroneous data if it is done too many times, especially if there is short range correlation in the data.

Interpretation of Output Statistics

The output statistics are best interpreted in relation to the partial spline model for N observed data values z_i given by

$$z_i = f(x_i) + b^T y_i + e_i \qquad (i = 1,...,N) \tag{1}$$

where each x_i is a d-dimensional vector of spline independent variables, f is an unknown smooth function of the x_i, each y_i is a p-dimensional vector of independent covariates, b is an unknown p-dimensional vector of coefficients of the y_i and each e_i is an independent, zero mean error term with variance $w_i\sigma^2$, where w_i is termed the relative error variance (known) and σ^2 is the error variance which is constant across all data points, but normally unknown (Hutchinson, 1991a). The model reduces, on the one hand, to an ordinary thin plate spline model when there are no covariates (p=0) and to a simple multivariate linear regression model, on the other hand, when $f(x_i)$ is absent. The latter possibility is not currently permitted by ANUSPLIN.

The function f and the coefficient vector b are determined by minimising

$$\sum_{i=1}^{N} \left(\frac{z_i - f(x_i) - b^T y_i}{w_i} \right)^2 + \rho\, J_m(f) \tag{2}$$

where $J_m(f)$ is a measure of the complexity of f, the "roughness penalty" defined in terms of an integral of mth order partial derivatives of f and ρ is a positive number called the smoothing parameter. As ρ approaches zero, the fitted function approaches an exact interpolant. As ρ approaches infinity, the function f approaches a least squares polynomial, with order depending on the order m of the roughness penalty. The value of the smoothing parameter is normally determined by minimising a measure of predictive error of the fitted surface given by the generalised cross validation (GCV).

The vector $\hat{z}$ of fitted function values can be written

$$\hat{z} = Az \tag{3}$$

where A is an N x N matrix called the *influence matrix*. By analogy with linear regression (Wahba 1990), the number of degrees of freedom of the fitted spline, or the effective number of parameters, is given by

$$SIGNAL = trace\ (A). \tag{4}$$

The number of degrees of freedom of the weighted residual sum of squares, the first term of equation (2), is given by

$$ERROR = trace\ (I - A) = N - trace\ (A). \tag{5}$$

The weighted mean residual sum of squares is given by

$$MSR = \left\| W^{-1} \left(I - A \right) z \right\|^2 / N \tag{6}$$

where W is the diagonal matrix given by

$$W = diag\left(w_1 , ..., w_N \right) \tag{7}$$

The *SIGNAL* degrees of freedom and the *ERROR* degrees of freedom for each surface add up to N (the number of data points).

The GCV is calculated for each value of the smoothing parameter ρ by implicitly removing each data point and calculating the residual from the omitted data point of a surface fitted to all other data points using the same value of ρ. The *GCV* is then a suitably weighted sum of the squares of these residuals (Craven and Wahba 1979, Wahba 1990). The *GCV* is actually calculated by the formula

$$GCV = \frac{\left\| W^{-1} \left(I - A \right) z \right\|^2 / N}{\left[tr\left(I - A \right) / N \right]^2} . \tag{8}$$

The surface fitting procedure is normally considered to have failed to find a genuine optimum value of the smoothing parameter if either the smoothing parameter is very small and the signal is the maximum possible (equal to the number of data points in the case of SPLINA or the number of knots in the case of SPLINB) or the smoothing parameter is very large and the signal is the minimum possible (a number which depends on the number of independent variables and the order of the roughness penalty). Both of these conditions are flagged by an asterisk in the output log file. Hutchinson (1993) and Hutchinson and Gessler (1994) recommend that the signal should not exceed about half the number of data points. Signals larger than this can indicate insufficient data or positive correlation in data errors.

The variance σ^2 of the data error e_i in equation (1) is estimated by

$$VAR = \frac{\left\| W^{-1} \left(I - A \right) z \right\|^2}{tr\left(I - A \right)} \tag{9}$$

If σ^2 is known, or estimated, an unbiased estimate of the "true" mean square error of the fitted function across the data points is given by

$$MSE = \left\| W^{-1} \left(I - A \right) z \right\|^2 / N - 2\sigma^2 tr\left(I - A \right) / N + \sigma^2 . \tag{10}$$

Craven and Wahba (1979) have shown that under suitable conditions the formula

$$GCV = MSE + \sigma^2 \tag{11}$$

holds approximately. Thus minimising *GCV*, which does not depend on knowing σ^2, is approximately equivalent to minimising *MSE*, the true mean square error.

The generalised cross validation *(GCV)*, mean square residual *(MSR)* and the data error variance estimate *(VAR)* are written to the output log file together with their square roots *(RTGCV, RTMSR, RTVAR)* which are in the units of the data values. *VAR* is the estimate of σ^2 given by equation (9). The mean square residual given by equation (6) is weighted according to the relative variance estimates w_i as provided in the data file. For the GCV calculation these relative variances are rescaled to have average value 1 in order to facilitate comparisons of *GCV* values across different models. If the relative variance estimates are actual estimates of the absolute value of the error variance (so that σ^2 =1), then *VAR* and *RTVAR* should be approximately 1.

The goodness of fit of the fitted model may be checked by comparing the scaled residual sum of squares ($N.MSR/\sigma^2$ where N is the number of data points) with the critical points of a chi-square variable with *df* degrees of freedom, where *df* is the error degrees of freedom, given by equation (5), as output by the program, and σ^2 is an *a priori* estimate of the error variance.

This variance corresponds to the "nugget" in standard kriging analyses. It is rarely known *a priori*, since it includes two distinct components. The first of these is error inherent in the data, such as measurement error. This may be known or reasonably estimated beforehand. However, the second component is the error in the underlying spline function. This error is essentially unknown, and decreases as the number of data points increases. In different situations one of these components can be dominant, or they can be equally important, as is often the case when interpolating climate statistics (Hutchinson 1995).

When an estimate of σ^2 is available an alternative strategy is to provide the corresponding standard deviation estimate σ to the program. The program then minimises an unbiased estimate of the true mean square error, *MSE* given by equation (10) instead of the *GCV*. This is not normally recommended since it depends on having a reasonably accurate estimate of σ^2. It is generally preferable to minimise *GCV*, since this appears to be more robust and does not depend on knowing σ^2. An *a priori* estimate of σ^2 can be better used to check the goodness of fit of the model as described above. On the other hand, specifying the error standard derivation may be preferable when there is no local minimum of the *GCV*, as can happen when fitting surfaces to very small data sets (less than about 20-30 data points). Note that SPLINB should not be used for small data sets unless there are coincident data points.

SPLINA and SPLINB provide in the output log file the coefficients of any covariates as well as the estimate of the mean square error of the smoothed data values *(MSE)*. This estimate depends on the value of error variance *(VAR)* as estimated by equation (9) or the input error standard deviation estimate when this has been provided by the user.

Calculation of Standard Errors

Using a Bayesian argument, Wahba (1983) and Silverman (1985) have adopted appropriate multi-variate Gaussian prior distributions for the vector z of data values, so that the error covariance matrix of the vector $\hat{z}$ of fitted values is given by

$$AW\sigma^2 \qquad (12)$$

where A is the influence matrix described in equation (3) and $W\sigma^2$ is the assumed error covariance matrix of the data errors. Here W is described by equation (7) and σ^2 is estimated by equation (9).

Spatially distributed standard errors for surfaces fitted by SPLINA are calculated using the method described by Hutchinson (1993). SPLINA calculates the error covariance matrix of the coefficients of the fitted spline surface by expressing the surface coefficients as a linear transformation of the vector $\hat{z}$ of fitted values. This includes the error covariance matrix of the coefficients of any covariates, from which standard error estimates of the coefficients of the covariates may be directly calculated. The error covariance matrices of the fitted surfaces are written by SPLINA to a separate binary file, as shown in Figure 1.

The value z_x of a spline surface at an arbitrary position x can be written

$$z_x = a_x^T c \qquad (13)$$

where a_x is a vector depending on x and c is the vector of fitted surface coefficients. The standard error estimate of the surface value z_x is then calculated by LAPPNT and LAPGRD using the formula

$$(a_x^T V a_x)^{1/2} \qquad (14)$$

where V is the error covariance matrix of the surface coefficients calculated by SPLINA. This standard error is called the model standard error, since it relates to the error in estimating the model given by equation (1). The prediction standard error is calculated by LAPPNT and LAPGRD using the formula

$$(a_x^T V a_x + \sigma^2)^{1/2} \qquad (15)$$

where σ^2 is the variance of the data error. This estimate is only applicable if the values being predicted have a uniform variance of σ^2 about the fitted spline function. This normally occurs when W is the identity matrix. Non-uniform error variances, such as those for the model discussed by Hutchinson (1995), must be accommodated using a separate calculation. Alternatively, non-uniform error variances may be directly accommodated in LAPPNT and LAPGRD using one of the transformations of the dependent variable described in the following section.

Confidence intervals of the calculated spline values are estimated by multiplying either the model standard error or the prediction standard error by 1.96 corresponding to the 95 percent two-sided confidence interval of the standard normal distribution.

The mean of an arbitrary number of fitted surface values is a linear function of the fitted surface coefficients. It can be expressed in the form

$$a^T c \qquad (16)$$

where a is the mean of the vectors a_x in equation (13). The standard error of the mean is therefore given by

$$(a^T V a)^{1/2}. \qquad (17)$$

This formula is used by LAPPNT and LAPGRD to calculate the standard error of the mean of the surface values when there is no dependent variable transformation. It is *not* the mean of the standard errors of the individual surface values.

Dependent Variable Transformations

Three dependent variable transformations, the square root, the natural logarithm and an occurrence transformation, are currently permitted by ANUSPLIN. Any of these transformations may be applied by SPLINA or SPLINB to the data before a spline surface is fitted. The square root and the natural logarithm transformations can reduce positive skew in measured values, as can arise when fitting data that are naturally non-negative or positive. The occurrence transformation is defined by setting all positive data value to 1.0 and ignoring all negative data values.

These transformations are automatically coded into the fitted surface coefficients file so that LAPPNT and LAPGRD can calculate either transformed surface values or back-transformed values. For the square root and natural logarithm transformations, these are obtained by applying the inverse dependent variable transformation (square or exponential) to the calculated surface values. When either of these inverse transformations is applied a correction for bias is made. Hutchinson (1998a) has found that applying the square root transformation to daily rainfall data, before fitting a thin plate smoothing spline, could reduce interpolation error by about 10 percent.

For the occurrence transformation, the back-transformation consists of setting output spline values to 0.0 or 1.0 depending on whether or not the fitted spline values are respectively less than or greater than the threshold value of 0.5. Standard errors are not available for the back-transformed occurrence values.

If the surface values are chosen to back-transformed using the inverse of the square root or natural logarithm transformations then standard errors are calculated by LAPPNT and LAPGRD accordingly. Formulae appropriate for the square root transformation have been demonstrated by Hutchinson (1998a). If the interpolated square root value is given by X, with standard error s, then an estimate of the standard error of X^2 is given by

$$SE(X^2) = 2s(X^2 + s^2/2)^{1/2}. \qquad (18)$$

This can be applied with s as either model standard error, or predictive standard error, as defined in the preceding section. The second term in this expression is negligible except when X is close to zero or s^2 is relatively large. Relative errors are thus given approximately by the formula

$$RE(X^2) = 2s/X \qquad (19)$$

that is twice the relative error in the square root surface values.

For the square root transformation, absolute standard error estimates are calculated. It can be seen from these two formulae that smaller surface values will be estimated with smaller *absolute* standard error, while larger surface values will be estimated with smaller *relative* error. Approximate confidence intervals are calculated in this case by assuming that the errors of the interpolated square root values are distributed according

to a normal distribution. It follows that the 95 percent confidence interval for the squared values is given by

$$[X^2 - CI, X^2 + CI] \quad (20)$$

where

$$CI = 4X\ 1.96\ s(X^2 + s^2/2)^{1/2}. \quad (21)$$

Analogous standard error estimates are calculated by LAPPNT and LAPGRD when the natural logarithm has been applied to the data values and the exponential transformation is applied to the interpolated values. If the interpolated logarithmic value is given by X, with standard error s, then LAPPNT and LAPGRD calculate the standard error in the value $exp(X)$ using the formula

$$SE(exp(X)) = exp(X + s^2/2)\ (exp(s^2) - 1)^{1/2}. \quad (22)$$

Relative confidence intervals, that must be applied multiplicatively, are calculated in this case by assuming that the errors of the interpolated logarithmic values are distributed according to a normal distribution. The two-sided 95 percent confidence interval is then given by

$$[exp(X)/CI, exp(X).CI] \quad (23)$$

where

$$CI = exp(1.96s). \quad (24)$$

LAPPNT and LAPGRD provide the absolute standard error estimate given by equation (22) and the relative confidence interval given by equation (24).

Fitting Climate Surfaces

The ANUSPLIN package was primarily developed for this task. There are normally at least two independent spline variables, longitude and latitude, in this order and in units of decimal degrees. A third independent variable, elevation above sea-level, is normally appropriate when fitting surfaces to temperature or precipitation. This is normally included as a third independent spline variable, in which case it should be scaled to be in units of kilometres. Minor improvements can sometimes be had by slightly altering this scaling of elevation. This scaling was originally determined by Hutchinson and Bischof (1983) and has been verified by Hutchinson (1995, 1998b).

Over restricted areas, superior performance can sometimes be achieved by including elevation not as an independent spline variable but as an independent covariate. Thus, in the case of fitting a temperature surface, the coefficient of an elevation covariate would be an empirically determined temperature lapse rate (Hutchinson, 1991a). Other factors that influence the climate variable may be included as additional covariates if appropriate parameterizations can be determined and the relevant data are available. These might include, for example, topographic effects other than elevation above sea-level. Other applications to climate interpolation have been described by Hutchinson *et al*. (1984ab, 1996a) and Hutchinson (1989a, 1991ab). Applications of fitted spline climate surfaces to global agroclimatic classifications and to the assessment of

biodiversity are described by Hutchinson *et al.* (1992, 1996b). They have also been used to develop spatially detailed climate change scenarios (Houser *et al*. 2004).

To fit multi-variate climate surfaces, the values of the independent variables are needed only at the data points. Thus meteorological stations should be accurately located in position and elevation. Errors in these locations are often indicated by large values in the output ranked residual list. Recent applications have examined the utility of using elevation and variables related to slope and aspect obtained from digital elevation models at various horizontal resolutions (Hutchinson 1995, 1998b). Thin plate spline interpolation of monthly mean precipitation and temperature has been favourably compared with other methods by Price *et al.* (2000).

The LAPGRD program can be used to calculate regular grids of fitted climate values and their standard errors, for mapping and other purposes, provided a regular grid of values of each independent variable, additional to longitude and latitude, is supplied. This usually means that a regular grid digital elevation model (DEM) is required. A technique for calculating such DEMs from elevation and stream line data has been described by Hutchinson (1988, 1989b, 1996, 2001).

SPLINA and SPLINB User Directives

User Directive	Type	Description
Title of fitted surfaces	60 characters	Title recorded in surface coefficient file to document surface.
Surface value units code and optional missing data value	0 – 8, real number	0 - undefined 1 - metres 2 - feet 3 - kilometres 4 - miles 5 - degrees 6 - radians 7 - millimetres 8 – megajoules Data values less than or equal to the missing data value are removed from the analysis. If a dependent data transformation is specified then data values outside the natural domain of the transformation are automatically removed. Thus negative data values are automatically removed if the square root dependent variable transformation is specified.
Number of independent spline variables	Non-negative integer	May not exceed specified limit (currently 10).
Number of independent covariates	Non-negative integer	Limit depends on the number of spline variables.
Number of surface independent spline variables	Non-negative integer	Surface independent variables take different values for each surface.
Number of surface independent covariates	Non-negative integer	Surface independent variables take different values for each surface.
Independent variable lower and upper limits, transformation code, units code, optional margins.	Two real numbers, two non-negative integers (0-8), up to two real numbers for each independent variable	Lower limit precedes upper limit. Data points outside these limits, augmented by margins, are ignored. One or both margins may be omitted. If one margin is supplied it is used as the common lower and upper margin. If both margins are omitted the transformation code and units code may also be omitted. Units code as for surface value units code.

Independent variable transformation parameters	One or two real numbers	Required for each independent variable for which the transformation code is positive. The possible transformations for each independent variable *x* are: 0 - no transformation *1 - x/a* *2 - ax* *3 -a log (x +b)* *4 - (x/b)* ^{a} *5 - a exp (x/b)* *6 - a tanh (x/b)* 7 - anisotropy angle in degrees 8 - anisotropy factor - in the direction specified by the anisotropy angle.
Dependent variable transformation	0, 1, 2 or 5	0 - no transformation. 1 - fit surface to natural logarithm of the data values. 2 - fit surface to the square root of the data values. 5 – occurrence – transform data values by setting all positive value to 1.0 and ignoring all negative values.
Order of spline	Positive integer	Usually 2. Lower limit specified by the program.
Number of surfaces	Positive integer	Any positive number of surfaces permitted.
Number of relative error variances	Non-negative integer	0 - data points uniformly weighted for each surface. 1 - the same weighting is applied to each surface. Number of surfaces - a different weighting is applied to each surface.
Optimization directive	0 – 2	0 - common smoothing parameter for all surfaces. 1 - common smoothing directive for all surfaces (default). 2 - different smoothing directive for each surface.
Smoothing directive for each surface	0 – 4	0 - fixed smoothing parameter -supply value. 1 - minimise GCV (default). 2 - minimise true mean square error using supplied error standard deviation estimate. 3 - fixed signal - supply value. 4 - minimise GML.

Data file name	255 characters	Must be supplied.
Maximum number of data points	Positive integer	Used to allocate memory for data and working arrays.
Number of characters in site label	0 - 20	If positive, an alphanumeric site label is expected for each data point in the data file. These labels are printed in the output data list and large residual files.
Data file format	255 characters	If non-blank the provided FORTRAN format statement is used to read in order: the site label (if number of characters in site name is positive), the independent variables (spline variables before covariates), the surface independent variables (spline variables before covariates), the data values and the relative variances as specified above. A uniform weighting of 1 for each data point may be specified by having zero relative variances. If the format is blank, the data file is read in list directed free format in the same order as for formatted reads.
Knot index file	255 characters (required only for SPLINB)	File of indices of data points selected as knots. The indices refer to the sequence numbers in the data file. Initially obtained by SELNOT.
Maximum number of knots	Positive integer (required only for SPLINB)	Used to allocate memory for knots and working arrays.
Input bad data flag file	255 characters (required only for SPLINB)	File used to remove particular data values from the analysis. Each record has a site label followed by binary number (0 or 1) for each surface, with each 1 indicating a corresponding data value to be removed. This permits removal of suspicious data values without altering the data file.

Output bad data flag file	255 characters (required only for SPLINB)	File contains all bad data flags from the input bad data flag file augmented by a flag for each data value that differs from the corresponding fitted surface value by more than 3.6 standard deviations. This file can be used as an input bad data flag file for a subsequent run of SPLINB after inspection and possible changes by the user.
Output large residual file name	255 characters	Blank if not required. Used to check for data errors. May be read directly by ADDNOT to add knots to an existing knot file for use with SPLINB.
Output optimisation parameters file	255 characters	Blank if not required. File containing parameters used to calculate the optimum smoothing parameter(s). This file can be used with GCVGML to calculate GCV or GML values as a function of the smoothing parameter.
Output surface coefficients file	255 characters	Normally required but may be blank if surface coefficients are not required. Contains the coefficients defining the fitted surfaces. These are used to calculate values of the surfaces by LAPPNT and LAPGRD.
Output data list file name	255 characters	Blank if not required. List of data and fitted values with Bayesian standard error estimates. Useful for checking for data errors.
Output error covariance file name	255 characters	Error covariance matrices of fitted surface coefficients. Used by LAPPNT and LAPGRD to calculate spatially distributed standard error estimates of the fitted surfaces.
Input validation data file name	255 characters	Blank if not supplied. If non-blank, residuals of the validation data points from the fitted surfaces are calculated, and summary statistics are written to the log file. This file normally holds data points that are not in the data file used to fit the surface. The validation data can provide independent validation of the output surface statistics.
Maximum number of validation data points	Positive integer (not required if the validation file name is blank)	Used to allocate memory for validation data and working arrays.

Number of characters in validation site label	0 - 20 (not required if the validation file name is blank)	If positive, an alphanumeric site label is expected for each validation data point.
Validation data format	255 characters (not required if the validation file name is blank)	As for the data file format above but no relative variances.
Output validation data list file name	255 characters (not required if the validation file name is blank)	If non-blank then a list of validation data and surface values is written to this file in standard format.

GCVGML

GCVGML calculates values of the GCV or GML statistic for surfaces produced by SPLINA or SPLINB. Values are tabulated as a function of the common logarithm of the smoothing parameter and written in columns to an output text file, with one column for each surface, in a format suitable for easy plotting by commonly available plotting packages. If there is more than one surface, the averages of the GCV or GML values over all surfaces are written to a final column. The GCV is the usually recommended statistic as it is more stable over different model structures and knot sets in the case of SPLINB.

GCVGML User Directives

User Directive	Type	Description
Optimisation parameters file name	255 characters	Name of optimisation parameters file produced by SPLINA or SPLINB.
Statistic	1 or 4	1 – GCV 4 – GML
Output GCV or GML file name	255 characters	Name of output text file with columns of GCV or GML values.

SELNOT

SELNOT is a program that selects an initial set of knots for use by the SPLINB program. As for SPLINB, multiple surfaces and multiple relative error variances are permitted. Independent and dependent variables are specified exactly as for SPLINB.

SELNOT selects knots by successively rejecting one point from the closest remaining pair of points in the independent spline variable space until the specified number of knots remain. Distances in the independent spline variable space are calculated after any specified transformations of the independent variables have been performed. Overall computational cost of the procedure is proportional to the square of the number of data points. The procedure was first described in Hutchinson (1984) and applied to rainfall interpolation by Hutchinson and Bischof (1983). It can also be used to select withheld data for validation of fitted surfaces (Hutchinson 1995, 1998ab).

SELNOT User Directives

User Directive	Type	Description
Number of independent spline variables	Non-negative integer	May not exceed specified limit (currently 50).
Number of independent covariates	Non-negative integer	Limit depends on the number of spline variables.
Number of surface independent spline variables	Non-negative integer	Surface independent variables take different values for each surface.
Number of surface independent covariates	Non-negative integer	Surface independent variables take different values for each surface.
Independent variable lower and upper limits, transformation code, units code, optional margins.	Two real numbers, two non-negative integers (0-8), two real numbers for each independent variable	Lower limit precedes upper limit. Data points outside these limits, augmented by margins, are ignored. One or both margins may be omitted. If one margin supplied it is used as the common lower and upper margin. If margins are omitted transformation code and units code may be omitted. Units code as for surface value units code.

Transformation parameters	One or two real numbers	Required for each independent variable for which the transformation code is positive. The possible transformations for each independent variable *x* are: 0 - no transformation *1 - x/a* *2 - ax* *3 - alog (x +b)* *4 - (x/b)* [a] *5 - a exp (x/b)* *6 - a tanh (x/b)* 7 - anisotropy angle in degrees 8 - anisotropy factor
Dependent variable transformation	0, 1, 2 or 5	0 - no transformation 1 - natural logarithm 2 - square root 5 - occurrence
Number of surfaces	Positive integer	Any positive number of surfaces permitted.
Number of relative error variances	Non-negative integer	0 - data points uniformly weighted for each surface 1 - the same weighting is applied to each surface. Number of surfaces - a different weighting is applied to each surface.
Data file name	255 characters	Must be supplied.
Maximum number of data points	Positive integer	Used to allocate memory for data and working arrays.
Number of characters in site name	0 - 20	If positive, a site name is expected for each data point in the data file.
Date file format	255 characters	Specify format for data and relative error variances, exactly as for SPLINB. Use blank to specify list directed free format.
Output knot file	255 characters	Name of output knot index file.

Rejected points file	255 characters	Not normally required. If non-blank, lists the index of each data point rejected as a knot together with the index of the closest knot. Points are listed in reverse order so the file begins with the last rejected point.
Number of knots	Positive integer	Normally within the range, calculated by the program, of 1/4 to 1/3 of the number of data points contained within the specified coordinate limits.

ADDNOT

The ADDNOT program adds data point indices to an existing knot file that has been initially calculated either by SELNOT for use with SPLINB. Knot indices may be read from standard input, or preferably read from the large residual list produced by a previous run of SPLINB. In this case the user must specify the number of knots to be added.

ADDNOT User Directives

User Directive	Type	Description
Old knot index file name	255 characters	Name of old knot file
Maximum number of knots in old knot file	Positive integer	Used to allocate memory for old knot index arrays.
Number of characters in site name	0-20	If positive, a site name is expected for each data point. Specify exactly as for SPLINB.
Large residual file	255 characters	Name of large residual file, as produced by a previous run of SPLINB. If blank, additional knot indices are read from standard input.
Number of additional knots	Positive integer	Required if the large residual file name is not blank. Number of knots to be added from the specified large residual file.
New knot file name	255 characters	Name of new augmented knot file.
Data point indices	Positive integers, with site names	Site names are required if the number of characters in the site name is positive. The lists of indices and site names can be supplied in an input command file. If ADDNOT is run interactively, terminate the list with a data point index of 0.

DELNOT

The DELNOT program can be used to adjust a knot index file when data points are removed from a data file to be used by SPLINB. This can avoid repetition of a run of SELNOT when bad data points are found late in the process of knot selection. However, if more than about 10-20 data points are removed it is recommended that the entire knot selection process be repeated.

DELNOT User Directives

User Directive	Type	Description
Old knot index file name	255 characters	Name of old knot index file
Maximum number of knots in old knot file	Positive integer	Used to allocate memory for knot index arrays.
Number of characters in site label	0-20	If positive, a site label is expected for each data point.
Name of new knot index file	255 characters	Name of new knot index file
Data point indices	Positive integers	Index of each data point that is to be removed from the date file for SPLINB. List indices in increasing order, one per record. These can be supplied in an input command file. If DELNOT is run interactively, terminate the list with 0.

LAPPNT

LAPPNT calculates values and spatially distributed errors of (partial) thin plate smoothing spline surfaces at points whose position coordinates are provided in a file. The spline surface coefficients are read from an ASCII file calculated by SPLINA or SPLINB. The error covariance matrices of the surface coefficients are read from a binary file calculated by the same run of SPLINA or SPLINB. All surfaces can be calculated by specifying 0 for the surface number.

Calculation time for each surface value is proportional to the number of knots. Calculation time for each error value is proportional to the square of the number of knots.

An alphanumeric label may be read from the user supplied point file and written to the output point file. No alphanumeric labels are read or written if the number of characters in the label is specified to be 0. Points outside the position limits in the surface coefficients file are ignored. The position coordinates are optionally written to the output point file. The program writes the number of points and summary statistics to standard output.

LAPPNT User Directives

User Directive	Type	Description
Surface file name	255 characters	Name of the surface coefficients file.
Surface numbers	Non-negative integers	Surface numbers to be calculated, in increasing order. Specify 0 if values of all surfaces are to be selected.
Type of surface calculation	0 or 1	0 - summary statistics only. 1 - calculate surface values.
Back-transform surface and error values	0 or 1 (not required if no surface transformation)	0 - do not apply surface back-transformation 1 - apply surface back-transformation.
Error covariance file name	255 characters	Blank if there is no covariance file or if no errors are to be calculated.
Type of error calculation	0 - 4 (not required if covariance file name is blank).	0 - calculate standard error of the average surface value only. 1 - calculate model standard errors. 2 - calculate prediction standard errors. 3 - calculate 95% model confidence intervals. 4 - calculate 95% prediction confidence intervals.

Maximum standard errors	Blank or maximum standard errors for all selected surfaces (not required if covariance file name is blank).	Surface values and error values are not calculated if the standard error exceeds the provided maximum error. When there is a surface transformation then maximum errors are applied to the error surface fitted to the transformed values.
Position coordinates file	255 characters	User supplied file with position coordinates.
Label size	Non-negative integer	Specifies the number of characters in the label attached to each set of coordinates in the position file. If label size is set to 0, then no label is read from the file.
Position file format	255 characters	Format of coordinates in position file. If label size is positive, then the format must include an initial alphanumeric format descriptor with number of characters set to the label size. If format is blank then the site label, if required, and the position coordinates are read in free format.
Output point file name	255 characters	Name of output point file.
Include position coordinates	0 or 1	0 - position coordinates are not included in the output point file. 1 - position coordinates are included in the output point file.
Output point file format	255 characters	Output format for writing both the input positions, with label when specified, and the output calculated surface values. Blank for free format.

LAPGRD

LAPGRD calculates values and spatially distributed errors of a regular two-dimensional grid of a (partial) thin plate smoothing spline surface. Coefficients defining the partial spline surface are read from an ASCII file calculated by SPLINA or SPLINB. The error covariance matrices of the surface coefficients are read from a binary format calculated by the same run of SPLINA or SPLINB. Calculation time for surface values is proportional to the number of knots times the number of grid points. Calculation time for error values is proportional to the square of the number of knots times the number of grid points.

Values of additional independent variables required to define the spline may be set to user supplied constants or read from user supplied grid files with the same number of rows and columns as the grid being calculated by LAPGRD. User supplied grids must be in row format, since they are read one row at a time to save storage space. All grids are read and written by rows from maximum Y to minimum Y.

Grid points may be specified to lie either at the corners or at the centres of grid cells. Normal usage with modern packages, including Arc/Info, Grass and Idrisi, is that grid points are specified to lie at the centres of grid cells. Points at corners of grid cells are a common option in older systems that generate vector output to display grids.

LAPGRD User Directives

User Directive	Type	Description
Surface file	255 characters	Name of the surface coefficients file.
Surface numbers	Non-negative integers	Surface numbers to be calculated in increasing order. Specify 0 if values of all surfaces in the surface coefficients file are to be selected.
Type of surface calculation	0 or 1	0 - summary statistics only. 1 - calculate surface values.
Back-transform surface and error values	0 or 1	Not required if there is no surface transformation. 0 - do not apply surface back-transformation 1 - apply surface back-transformation.
Error covariance file name	255 characters	Blank if there is no covariance file or if no errors are to be calculated.
Type of error calculation	0 - 4	0 - calculate standard error of the average surface value only. 1 - calculate model standard errors. 2 - calculate prediction standard errors. 3 - calculate 95% model confidence intervals. 4 - calculate 95% prediction confidence intervals.

Maximum standard errors	Blank or maximum standard errors for all selected surfaces	Surface values and error values are not calculated if the standard error exceeds the provident maximum error. When there is a surface transformation then maximum errors are applied to the errors of the surface fitted to the transformed values.
Grid position option	0 or 1	0 - grid points at cell corners. 1 - grid points at cell centres. Normally 1 for Arc/Info, Grass and Idrisi.
Index of first grid variable	Non-negative integer	If positive, identifies the independent variable of the spline which increments across each row of the output grid - normally 1. If zero then values of this independent variable are read from a grid.
Limits and spacing of first variable	3 real numbers	Lower limit, upper limit and spacing respectively of first grid independent variable.
Index of second grid variable	Non-negative integer	If positive, identifies the independent variable of the spline which increments along each column of the output grid - normally 2. If zero then values of this independent variable are read from a grid.
Limits and spacing of second grid variable	3 real numbers	Lower limit, upper limit and spacing respectively of second grid independent variable. N.B. The spacing of the first and second variable must be equal when reading or writing Arc/Info or Idrisi grids.
Mode of mask grid	0 - 3	0 - mask grid not supplied. 1 - generic mask grid. 2 - Arc/Info mask grid. 3 - Idrisi mask grid.
Name of mask grid	255 characters (Not required if mode of mask grid is zero)	Grid used to mask out special values. The mask corresponds to the no-data values of the mask grid. Mask grids in standard Arc/Info or Idrisi mode are recommended. If the mask grid is in generic mode, the row format (blank for binary format, non-blank for free ASCII format), no value indicator (0 or 1) and the no data value (real number) are also required.

Specify for each remaining independent variable (if spline has more than two independent variables or if the first and second grid variable indices not both positive.):-

Mode of the independent variable	0 - 3	0 - user supplied constant. 1 - user supplied grid in generic row format with the same size as the grid being calculated. 2 - user supplied Arc/Info grid with same size as the grid being calculated. 3 - user supplied Idrisi image with the same size as the grid being calculated.
Constant	Real number (Only required if mode is 0)	Independent variable grid is set to this constant.
Input grid file name	255 characters (Required if mode is not 0.)	File name of user supplied grid. If the independent variable is a surface independent variable then a separate file name is required for each surface being calculated. Input grids in standard Arc/Info or Idrisi mode are recommended. If the input grid is in generic mode, the row format (blank for binary format, non-blank for free ASCII format), no value indicator (0 or 1) and the no data value (real number) are also required.

If the surface calculation type is 1 then specify:-

Mode of output surface value grids	0 - 3	0 - grid written in X,Y,Z format. 1 - generic grid written by rows. 2 - Arc/Info grid. 3 - Idrisi image. Output grids in standard Arc/Info or Idrisi mode are recommended.
Special value of output grid	Real number (Must be supplied if input grids have special values or maximum errors are specified as above.)	Indicates no data value in output grid.
Output grid file names	255 characters	File names of all output surface value grids.

Output grid format	255 characters	Must be consistent with the format mode of the output grids specified above. If blank then output grid is written as an unformatted binary file. This is normally recommended as it saves time and storage space. Use an ASCII formatted grid when the grid is to be moved between DOS and UNIX platforms.

If the error calculation type is positive then specify:-

Mode of output error grids	0 - 3	0 - grid written in X,Y,Z format. 1 - generic grid written by rows. 2 - Arc/Info grid. 3 - Idrisi image. Output grids in standard Arc/Info or Idrisi mode are recommended.
Special value of output grid	Real number (Must be supplied if input grids have special values or maximum errors are specified as above.)	Indicates no data value in output grid.
Output grid file names	255 characters	File names of all output error surface grids.
Output grid format	255 characters	Must be consistent with the format mode of the output grids specified above. If blank then output grid is written as an unformatted binary file. This is normally recommended as it saves time and storage space. Use an ASCII formatted grid when the grid is to be moved between DOS and UNIX platforms.

ANNOTATED EXAMPLES

In order to test and demonstrate the 8 programs in ANUSPLIN, test data and example command files have been provided in four groups in four separate sub-directories. The example data sets, command files and reference outputs can be found under the test directory in the ANUSPLIN installation root directory. The ANUSPLIN installation root directory is the directory specified when the package was installed and will vary from system to system.

The first group illustrates the basic principles of data smoothing by applying SPLINA to simulated noisy uni-variate data, obtained by randomly perturbing points from a sine curve.

The second group illustrates smoothing of monthly mean temperature data using a tri-variate partial spline function of longitude, latitude and elevation. Analyses with both SPLINA and SPLINB are illustrated.

The third group illustrates smoothing of monthly mean precipitation data using a full tri-variate spline function of longitude, latitude and elevation. Analyses with SPLINB, using knots selected by SELNOT, are illustrated because precipitation data sets are often large. Use of independent variable margins and of the square root transformation of the dependent precipitation values is also illustrated.

The fourth group illustrates smoothing of monthly mean solar radiation data using a bivariate spline function of longitude and latitude only and using a tri-variate spline function with precipitation as a third "surface independent variable". Each analysis is performed using SPLINA.

The examples are intended to test the installation of ANUSPLIN and to provide canonical examples of applications to uni-variate data and multi-variate climate data. Each group of examples contains a table showing all commands and input and output files. Each table is followed by explanatory notes for each command in the proceeding table

Spline smoothing of uni-variate data

To illustrate the basic concepts and procedures for data smoothing, two data files are supplied in the test/math subdirectory:

sine.dat - 101 noisy data values obtained by perturbing points from a single sine curve by random values from a zero mean normal variable with standard deviation 0.2

sine.val - 101 values of the true sine curve

These data are displayed in Figure 2 below.

Six ANUSPLIN command files for processing these data files are provided in the test/math subdirectory and listed in the table below. Each command and its outputs are discussed in the notes following the table. All output files are provided in the test/math/out subdirectory.

Command	Input Files	Output Files
1. splina < sine.cmd > sine.log	sine.dat sine.val	sine.log sine.opt sine.sur sine.lis sine.cov sine.out
2. gcvgml < sinegcv.cmd > sinegcv.log	sine.opt	sine.gcv
3. lappnt < sinepnt1.cmd > sinepnt1.log	sine.sur sine.cov sine.val	sinepnt1.out
4. lappnt < sinepnt3.cmd > sinepnt3.log	sine.sur sine.cov sine.val	sinepnt3.out
5. lappnt < sinepnt2.cmd > sinepnt2.log	sine.sur sine.cov sine.val	sinepnt2.out
6. lappnt < sinepnt4.cmd > sinepnt4.log	sine.sur sine.cov sine.val	sinepnt4.out

Notes

1. This command uses SPLINA to fit a second order smoothing spline to the noisy data points shown in Figure 2. It produces optimisation parameters in the output file sine.opt for use by command 2. It also produces surface coefficients in the output file sine.sur and error covariances in the output file sine.cov for use by commands 3-6. The command also produces an output large residual file in sine.res and an output list file in sine.lis, which lists the data and fitted values, together with Bayesian standard error estimates. These files are normally used, in conjunction with summary statistics in the output log file to aid detection and correction of data errors, as indicated in Figure 1. The largest data residual from the fitted spline is the 61st data point, as listed under the ranked root mean square residuals in the output log file, and in the output file sine.res. This data point has an x value of 216 degrees and can be seen clearly in Figure 3.

The fitted spline curve is plotted in Figure 3, showing good agreement with the original sine curve in Figure 2. This command also calculates values of the fitted spline function compared with the true sine values provided in the input file sine.val. The true and fitted values are written to the output file sine.out, and summary validation statistics are written to the output log file. The largest residual of the true sine values from the fitted spline is the 38th point, as listed under the validation statistics in the output log file. This point has an x value of 133.2 degrees and can also be seen clearly in Figure 3.

Unlike common applications of SPLINA to higher dimensional data, there are no site labels in the data file. In this case each data point is labelled by the program to be its sequence number in the data file. No units are specified for the data values, and no transformations are applied to either the independent variable or the dependent variable. No margins are specified for the independent variable. No weighting is applied to the data values.

The order of the spline is specified to be 2, giving rise to a minimum curvature smoothing spline. This spline can be represented, in the uni-variate case only, by a piece-wise cubic polynomial. This representation is not provided by the ANUSPLIN package, which is primarily designed for general applications to multi-variate data. Efficient "order (n)" cubic spline smoothing of uni-variate data, using a piece-wise cubic representation, can be obtained using the procedure CUBGCV (Hutchinson and de Hoog 1985, Hutchinson 1986).

The amount of data smoothing is determined in this example by minimising the generalised cross validation (GCV). The log file shows that the fitted spline has 8.4 degrees of freedom, or 8.4 effective parameters, as given by the trace of the influence matrix associated with the fitted spline (Wahba 1990). The number of degrees of freedom of the residual is 92.6. These two numbers sum to 101, the number of data points. The signal to noise ratio of this smoothing analysis is 8.4/92.6=0.09. The size of the signal is much less than the half the number of data points, in line with the heuristic recommendation in Hutchinson (1993) and Hutchinson and Gessler (1994). Equivalently, the signal to noise ratio is less than 1.0. The square root of the GCV, or "root mean square predictive error", is listed under RTGCV as 0.183. The root mean square residual of the spline from the data is listed under RTMSR as 0.168, and the estimate of the standard deviation of the noise in the spline model is listed under RTVAR as 0.175. This estimate is reasonably close to the known standard deviation of the noise in the data of 0.2. Further examples of smoothing spline analyses of uni-variate noisy data have been given by Craven and Wahba (1979).

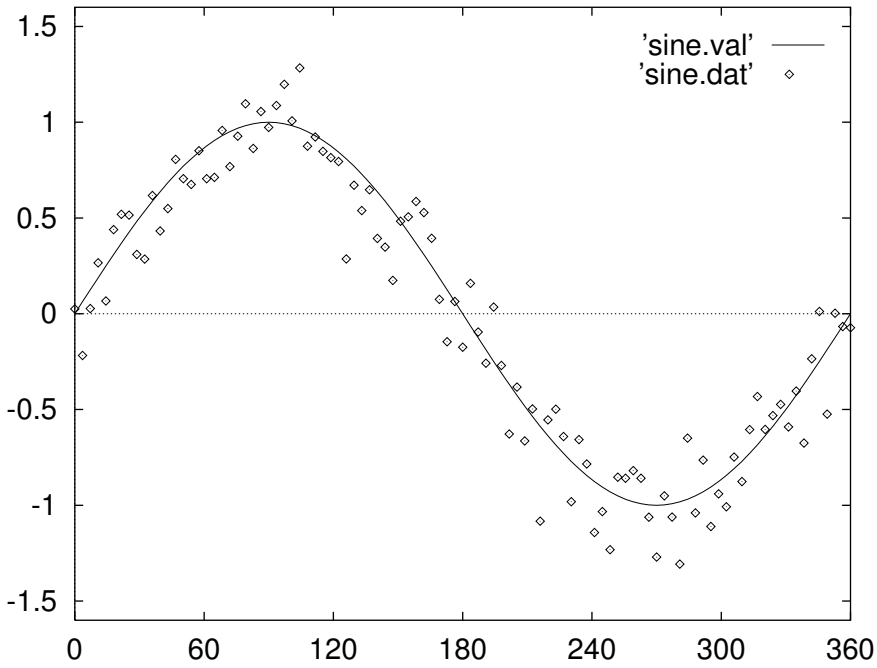

Figure 2. Sine curve and 101 noisy data points perturbed from the sine curve by values from a zero mean normal variable with standard deviation 0.2.

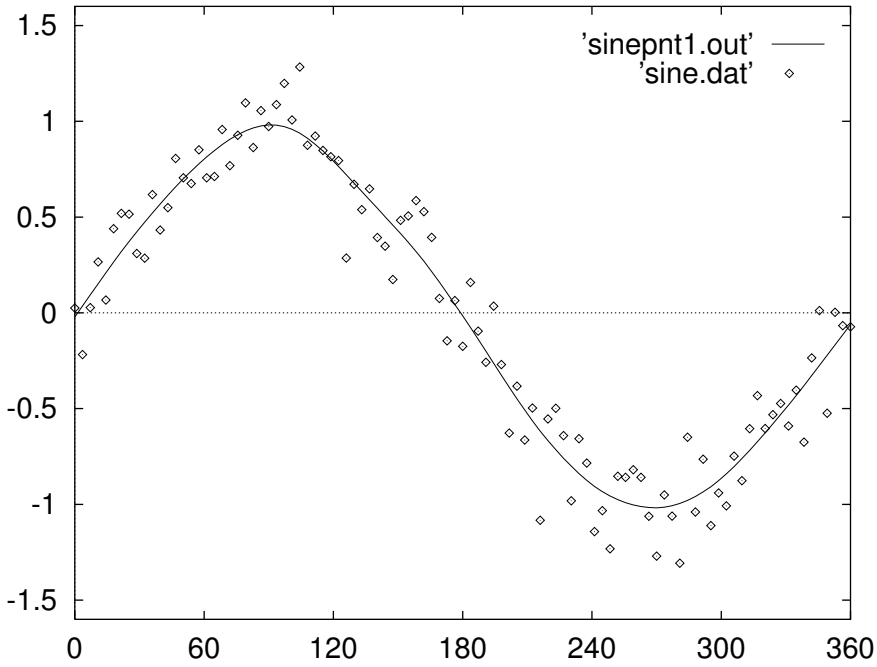

Figure 3. Fitted spline curve with the 101 noisy data points.

The root mean square error estimate of 0.0504, listed under RTMSE, is an estimate of the error in the fitted function after the effects of the noise in the data have been removed from the RTGCV. This is reasonably close to the root mean square residual from the true sine curve, which has been obtained from the values in the file sine.val and is listed under RMS as 0.0432. In this example the variance of the error of the fitted spline is dominated by the variance of the noise in the data values. In many applications, such as the interpolation of rainfall (Hutchinson 1995), error in the spline itself contributes significantly to the estimated error variance. In such cases the error estimate listed under RTMSE would be optimistic. In general, the standard deviation of the true error of the fitted spline will lie somewhere between RTMSE and RTGCV, depending on the relative magnitudes of the error in the noise and the error in the fitted spline.

2. This command uses GCVGML to calculate values of the GCV as a function of the logarithm to base 10 of the smoothing parameter. GCVGML uses the optimisation parameters, as calculated by SPLINA in the file sine.opt, and writes the table of GCV values to the output file sine.gcv. These values are plotted in Figure 4. The GCV normally has a unique local minimum value, which in this case occurs when the logarithm of the value of the smoothing parameter is 4.4. The corresponding value listed under RHO in the SPLINA log file is 0.255E+5. Multiple local minima in GCV curves can indicate significant errors in the data or significant mis-specification of the spline model. SPLINA attempts to determine the smoothest local minimum when there are multiple local minima, in order to choose the model with the least number of effective parameters.

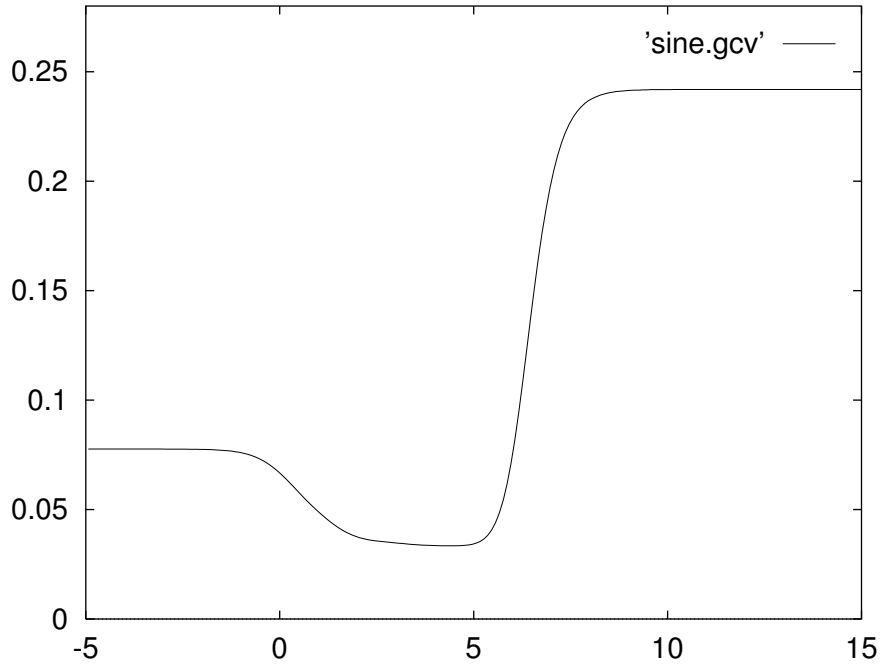

Figure 4. Plot of the GCV as a function of the logarithm of the smoothing parameter.

3. This command uses LAPPNT to calculate values of the fitted spline, and corresponding Bayesian standard error estimates, using surface coefficients provided in the input data file sine.val and error covariances in the input data file sine.cov. The error covariance matrix of the spline coefficients are calculated according to the method described by Hutchinson (1993). Spline values and standard errors are calculated at the *x* values provided in the first column of the input data file sine.val and are written to the output file sinepnt1.out. The output spline values are plotted as the curve in Figure 3.

The Bayesian standard error estimates are plotted in Figure 5. In this case, model standard errors are calculated (error calculation type = 1). These standard errors correspond to standard errors of the fitted parameters of a linear regression model. They are essentially functions of local data density, being approximately 0.05 for most interior points, but rapidly increasing towards 0.1 as points approach the boundary of the data points. The Bayesian standard errors increase without bound at positions beyond the limits of the original data points.

4. This command uses LAPPNT to calculate values of the fitted spline, and corresponding Bayesian 95% confidence intervals, using the same input files as for command 3. Model confidence intervals are specified (error calculation type = 3). Output spline values and confidence intervals are written to the output file sinepnt3.out. The confidence intervals are plotted in Figure 6, together with 101 values of the true sine curve. The 95% confidence intervals are calculated as 1.96 times the model standard errors calculated by command 3. This assumes that the errors are distributed according to a normal distribution. Just 3 of the 101 true sine values lie beyond the 95% confidence intervals, acceptably close to the expected number of about 5.

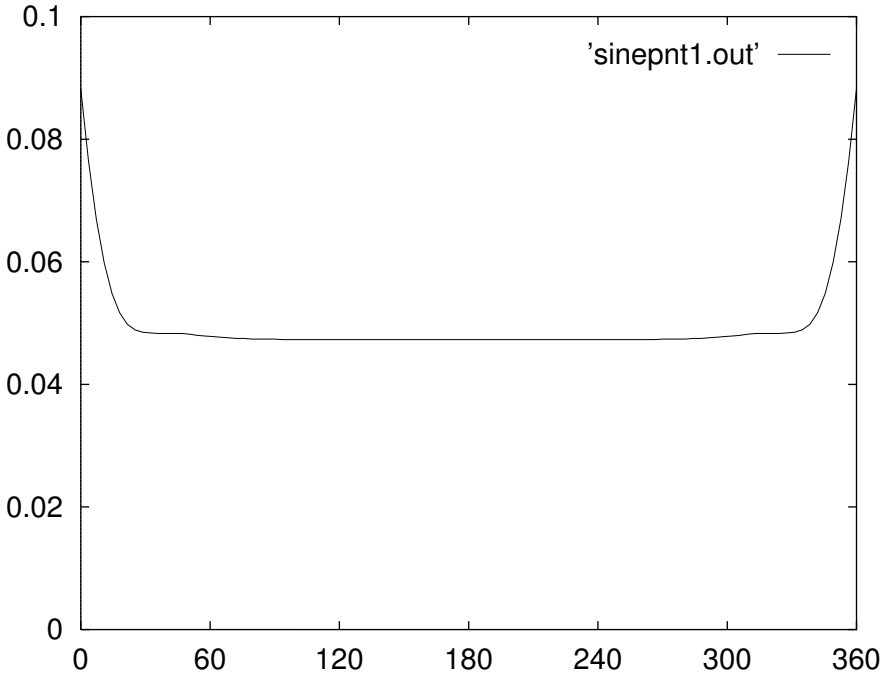

Figure 5. Plot of Bayesian model standard errors of the fitted spline.

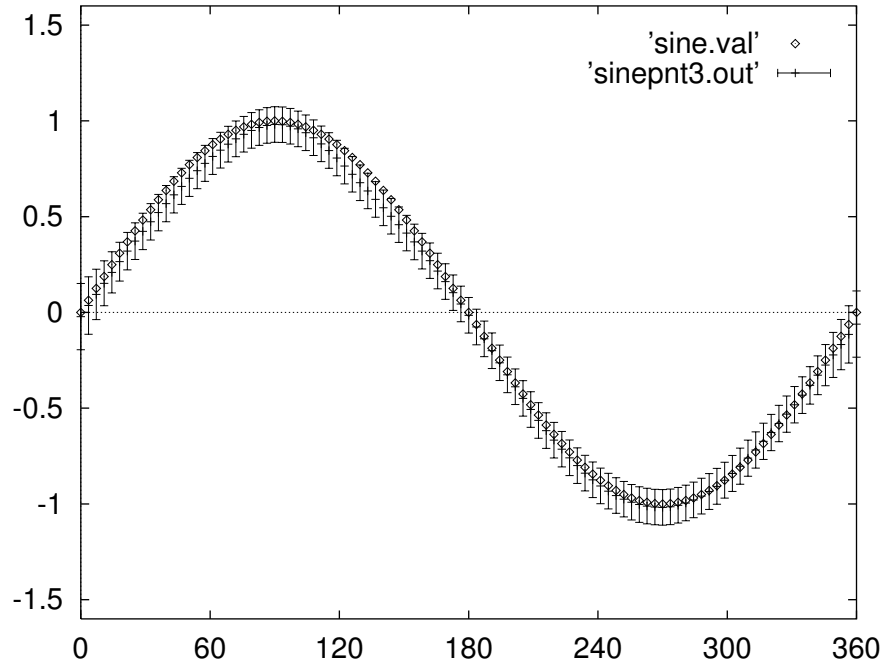

Figure 6. Plot of 95% model confidence intervals together with 101 true sine values.

5. This command uses LAPPNT to calculate values of the fitted spline, and corresponding Bayesian standard error estimates, using the same input files as for command 3. Prediction standard errors are calculated (error calculation type = 2). Output spline values and prediction standard errors are written to the output file sinepnt2.out. These standard errors correspond to standard errors in estimating data from the spline model. The prediction standard errors are obtained from the model standard errors calculated by command 3 using the formula

$$\sigma_p = (\sigma_m^2 + \sigma^2)^{1/2}$$

where σ_p is the prediction standard error, σ_m is the model standard error, and σ=*0.175* is the estimated standard deviation of the data errors. The prediction standard error estimates are plotted in Figure 7. Since the data errors in this case dominate the model standard errors, the prediction standard errors increase only slightly at positions close to the boundary of the data points. However, as for the standard model errors, the prediction standard errors would increase without bound at positions beyond the limits of the original data points.

6. This command uses LAPPNT to calculate values of the fitted spline, and corresponding two-sided Bayesian 95% confidence intervals, using the same input files as for command 3. Prediction confidence intervals are specified (error calculation type = 4). Output spline values and confidence intervals are written to the output file sinepnt4.out. The confidence intervals are plotted in Figure 8, together with 101 data values obtained in a separate simulation from the original data values. The 95% prediction confidence intervals are calculated as 1.96 times the prediction standard errors calculated by command 5. This assumes that the prediction errors are distributed according to a normal distribution. Seven of the 101 simulated noisy data values lie beyond the 95% confidence intervals, acceptably close to the expected number of about 5.

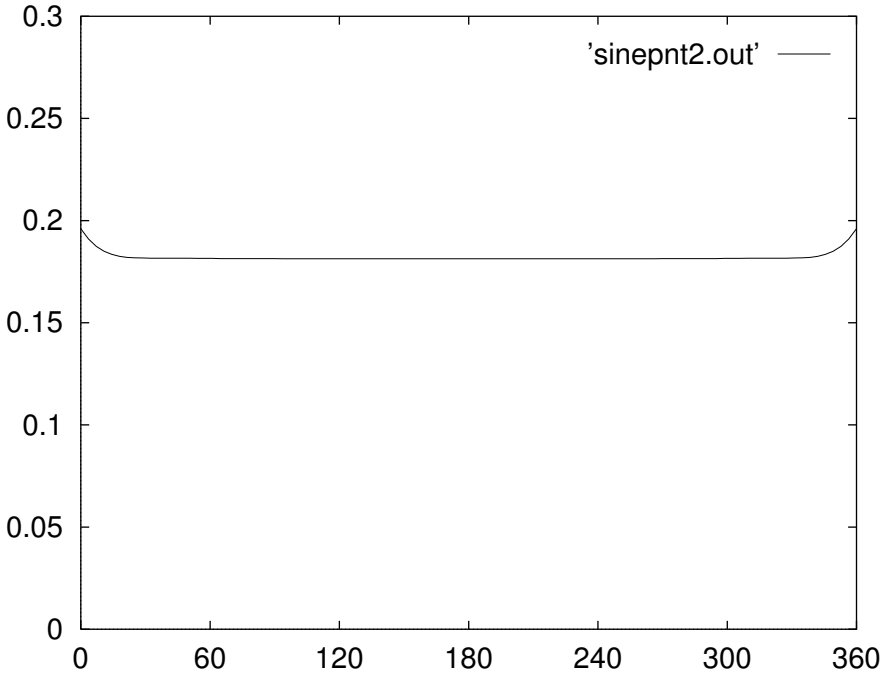

Figure 7. Plot of prediction standard errors of the fitted spline.

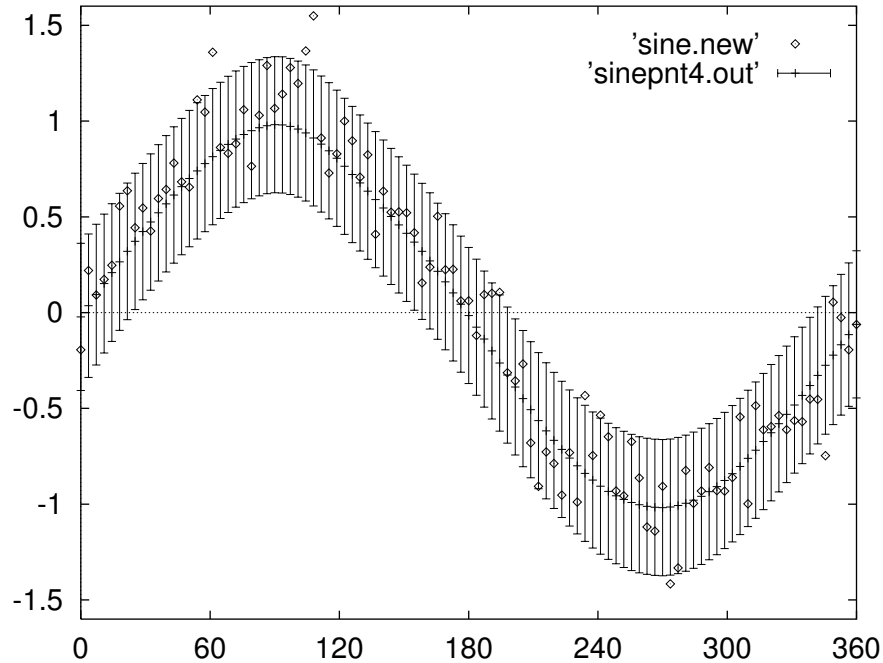

Figure 8. Plot of 95% prediction confidence intervals together with 101 simulated data values distinct from the original data values.

Partial spline smoothing of monthly mean temperature data

To illustrate the tri-variate partial spline smoothing of monthly mean temperature data, five data files are supplied in the test/temp subdirectory:

tmaxa.dat - monthly mean temperature data

tmaxb.dat - monthly mean temperature data with coincident data points

tas4.dem - small DEM in Arc/Info ASCII grid format

tas4x.grd - X coordinates of the small DEM in Arc/Info ASCII grid format

tas4y.grd - Y coordinates of the small DEM in Arc/Info ASCII grid format

tas4.img - same DEM as in tas4.dem, in Idrisi ASCII image format

tas4.doc - standard Idrisi documentation file for tas4.img

Fourteen ANUSPLIN command files for processing these data files are provided in the test/temp subdirectory and listed in the table below. Each command and its outputs are discussed in the notes following the table. All output files are provided in the test/temp/out subdirectory.

Command	Input Files	Output Files
1. splina < tmaxa.cmd > tmaxa.log	tmaxa.dat	tmaxa.log tmaxa.res tmaxa.opt tmaxa.sur tmaxa.cov tmaxa.lis
2. gcvgml < tmaxagcv.cmd > tmaxagcv.log	tmaxa.opt	tmaxagcv.log tmaxa.gcv
3. lappnt < tmaxapnt.cmd > tmaxapnt.log	tmaxa.sur tmaxa.cov tmaxa.dat	tmaxapnt.log tmaxapnt.out
4. lapgrd < tmaxasum1.cmd > tmaxasum1.log	tmaxa.sur tmaxa.cov tas4.img tas4.doc	tmaxasum1.log
5. lapgrd < tmaxasum2.cmd > tmaxasum2.log	tmaxa.sur tmaxa.cov tas4x.grd tas4y.grd tas4.dem	tmaxasum2.log
6. lapgrd < tmaxagrd.cmd > tmaxagrd.log	tmaxa.sur tmaxa.cov tas4.dem	tmaxagrd.log tmaxa.1grd tmaxa7.grd tcova1.grd tcova7.grd

7. lapgrd < tmaxaimg.cmd > tmaxaimg.log	tmaxa.sur tmaxa.cov tas4.doc tas4.img	tmaxaimg.log tmaxa1.doc tmaxa1.img tmaxa7.doc tmaxa7.img tcova1.doc tcova1.img tcova7.doc tcova7.img
8. selnot < tmaxsel.cmd > tmaxsel.log	tmaxb.dat	tmaxsel.log tmaxb.not tmaxb.rej
9. splinb < tmaxb.cmd > tmaxb.log	tmaxb.dat tmaxb.not	tmaxb.log tmaxb.res tmaxb.opt tmaxb.sur tmaxb.cov tmaxb.lis
10. gcvgml < tmaxbgcv.cmd > tmaxabgcv.log	tmaxb.opt	tmaxagcv.log tmaxb.gcv
11. lapgrd < tmaxbgrd.com > tmaxbgrd.log	tmaxb.sur tmaxb.cov tas4.dem	tmaxbgrd.log tmaxb1.grd tmaxb7.grd tcovb1.grd tcovb7.grd
12. addnot < tmaxad1.cmd > tmaxad1.log	tmaxb.not	tmaxad1.log
13. addnot < tmaxad2.cmd > tmaxad2.log	tmaxb.not	tmaxad2.log tmaxad2.not
14. delnot < tmaxdel.cmd > tmaxdel.log	tmaxb.not	tmaxdel.log tmaxdel.not

Notes

1. This command uses SPLINA to fit a partial thin plate smoothing spline, with linear dependence on elevation, to monthly mean values of daily maximum temperature data in the file tmaxa.dat. The data have the same uniform weight for each surface, allowing SPLINA to fit 12 monthly surfaces simultaneously. The data are read using a FORTRAN format.

The log file tmaxa.log contains summary statistics for the analysis, including the number of points read by the program, the generalised cross validation for each monthly surface, the standard error of each fitted monthly mean maximum temperature elevation lapse rate and a ranked list of the largest residuals from the fitted surfaces. The log file should *always* be carefully inspected. Large residuals from the fitted surface often

indicate errors in data positions or values. The ranked list of large residuals is also written to tmaxa.res.

The fitted surface coefficients are stored in tmaxa.sur. The error covariance matrices of the surface coefficients for each surface are stored in tmaxa.cov in binary form only. This file cannot be moved between DOS and Unix platforms. The surface coefficients and the error covariance matrices are used to calculate values of the fitted surfaces and spatially distributed standard errors by LAPPNT and LAPGRD in commands 3,4,5,6 and 7.

The list of data and fitted values is stored in tmaxa.lis. This file also contains a Bayesian standard error estimate for each fitted value. This file can assist detection of data errors when used in conjunction with the large residual list. The optimisation parameters in tmaxa.opt can be used by GCVGML to calculate the GCV as a function of different values of the smoothing parameter, as in command 2.

The log file shows that the signals of the fitted surfaces vary between 6 and 39. Almost all of these values are less than half the number of data points, in agreement with the general recommendation. A signal much larger than half the number of data points indicates either significant data errors or that there are insufficient data to fit the surface model. There is a generally systematic progression in the signals from month to month, although the higher signal in June (surface number 6) indicates some instability in the determination of the smoothing parameter which, in this case, is probably due to data errors.

The square root of the GCV (RTGCV) varies between 0.74 degrees in June and 1.2 degrees in February. These are conservative estimates of overall standard prediction error because they include the data error, as estimated by the procedure. The root mean square model error (RTMSE) is an estimate of standard error after the estimated data error has been removed. This may be likened to a standard error estimate of a fitted coefficient of a parametric model. It is a somewhat optimistic estimate of surface error because the procedure includes deficiencies in the model in the estimated data error. The RTMSE varies between 0.23 degrees in June and 0.57 degrees in February. Standard error estimates less than 0.5 degrees are typical when fitting splines to monthly mean maximum temperature data.

The coefficients of the parametric sub-model, which can be interpreted as temperature lapse rates, are approximately 8 degrees per 1000 metres. This agrees with known process controls on this value. The free air dry adiabatic lapse rate is 10 degrees per 1000 metres. The elevation lapse rate for minimum temperature is generally less than 8 degrees per 1000 metres (Hutchinson 1991a). Note that elevation, the third independent variable, has been scaled to be in kilometres. The standard error estimates of the lapse rates for the 12 surfaces ranges between 0.35 and 0.58, consistent with *a priori* expectations, and with the month to month variation in the fitted covariate values.

The stations with the four largest residuals all have significant elevation errors. The four elevation values for the corresponding points in the data file are 700, 305, 145 and 200 metres. The correct values are 1250, 40, 5 and 80 metres respectively. The departures of the fitted temperature values, as can be seen in the file tmaxa.lis, are consistent with the fitted temperature elevation lapse rates. The fifth largest residual in the large residual list is associated with a point on the coast, where close proximity to the ocean can significantly reduce maximum temperatures.

Fitting temperature with a partial spline dependence on elevation provides a robust analysis of elevation dependence that is very useful for flagging elevation errors in the data. These errors have been corrected in the data file tmaxb.dat used by commands 8 and 9.

2. Uses GCVGML to calculate values and model standard errors for the GCV for each month, as a function of the smoothing parameter, in the file tmaxa.gcv. The optimisation parameters required for this calculation have been obtained from the optimisation parameters file tmaxa.opt, as produced by command 1.

3. Uses LAPPNT to calculate values of the 12 fitted surfaces, fitted by command 1, at positions specified in the file tmaxa.dat. Since this is the same data file used in command 1, the calculated surface values should be identical to the fitted values in the file tmaxa.lis. In this case the data file tmaxa.dat is read using a FORTRAN format statement. The surface coefficients are read from the file tmaxa.sur and the error covariance matrices of the surface coefficients are read from the file tmaxa.cov. The log file includes summary statistics for the output surface and standard error values.

4. Uses LAPGRD to calculate summary statistics of grids of mean daily maximum temperature and standard errors for the four mid-season months. LAPGRD uses the surface coefficients and error covariance matrices calculated by command 1 and the DEM in Idrisi image format in tas4.img. The Idrisi image file is specified in the command file by its stem name tas4. LAPGRD then requires the corresponding Idrisi documentation file tas4.doc. The actual surface and standard error grids are not calculated. The summary statistics consist, for each month, of the number of valid grid points, the mean of the grid of valid surface values and the standard error of the surface mean. Note that this is NOT the mean of the grid of standard errors.

5. Uses LAPGRD to calculate the same summary statistics of grids as calculated by command 4 but uses X and Y coordinates supplied separately as grids in tas4x.grd and tas4y.grd and elevations supplied as a grid in tas4.dem. This can be useful in modelling situations where the X or Y coordinates used to fit the spline surface are not the standard X or Y coordinates but are instead functions of position.

6. Uses LAPGRD to calculate grids of values of mean daily maximum temperature and prediction standard errors for the months of January and July. The grids of surface values depend on the surface coefficients tmaxa.sur calculated by SPLINA in command 1 and the small DEM provided as tas4.dem. The standard error grids also depend on the error covariance matrices in tmaxa.cov calculated by SPLINA.

The file tas4.dem is in standard Arc/Info ASCII GRID format. LAPGRD reads the elevation data from this file, in units of metres, without further specification of format. Special or NODATA values, as specified in the header of this file, are recognised by LAPGRD. Surface values are not calculated by LAPGRD for such values. Binary Arc/Info GRIDS, with accompanying standard ASCII header file, are also recognised by LAPGRD, provided the ASCII header file is provided with the standard file extension ".hdr". The position limits and grid spacing in the Arc/Info header file are checked for compatibility with the position limits and grid spacing specified in the command file tmaxagrd.cmd.

7. Uses LAPGRD to calculate grids of values of mean daily maximum temperature and prediction standard errors for January and July. The grids depend on the surface

coefficients tmaxa.sur, the error covariance matrices in tmaxa.cov calculated by SPLINA in command 1 and the small DEM in standard Idrisi ASCII IMAGE format, provided as tas4.img. Surface values and errors are not calculated if the prediction standard error exceeds the value 0.8. This reduces the number of grid points calculated from 148 to 80 in January and from 148 to 43 in July. This facility is useful in preventing calculation of grid values with very large estimated errors.

The accompanying standard Idrisi Documentation file, called tas4.doc, is expected by LAPGRD. The position limits and grid spacing in this file are checked for compatibility with the position limits and grid spacing specified in the command file tmaxaimg.cmd.

8. Uses SELNOT to select 40 knots from the 72 data points in tmaxb.dat for input to the approximate thin plate spline fitting procedure SPLINB in command 9. SELNOT provides a recommended range of number of knots, between approximately 1/4 and 1/3 of the number of data points within the specified X,Y limits. Indices and names of the selected knots are stored in tmaxsel.not. The independent and dependent variables are specified exactly as in the command files used for SPLINA and SPLINB in commands 1 and 9 respectively.

9. Uses SPLINB to fit an approximate partial thin plate spline, with linear dependence on elevation, to the corrected monthly mean daily maximum temperature data in the data file tmaxb.dat. The approximate spline is constructed from the knots specified in tmaxsel.not, as produced by command 8. Other specifications are exactly as for SPLINA in command 1.

The log file tmaxb.log contains summary statistics for the analysis, including the number of points read by the program, the cross validation for each monthly surface and a ranked list of the largest outliers from the fitted surfaces. The predictive errors obtained from using the corrected data are considerably reduced from those obtained by command 1, with the RTGCV now varying between 0.40 in September to 0.90 in February. The log file should *always* be carefully inspected. Large outliers from the fitted surface often indicate errors in data positions or values. In this case the use of the corrected data has removed all large outliers.

The ranked list of large residuals is written to tmaxb.res. This is used by ADDNOT in command 12 to add knots to the knot file from the points with the largest residuals from the fitted surface.

The fitted surface coefficients are stored in tmaxb.sur. The error covariance matrices of the surface coefficients are stored in tmaxb.cov. The list of data and fitted values is stored in tmaxb.lis. The optimisation parameters in tmaxb.opt can be used by GCVGML to calculate the GCV as a function of different values of the smoothing parameter, as in command 10.

The fitted elevation lapse rates for this analysis are very similar to the analysis discussed for command 1 but the standard errors have been halved because of the corrected data. The use of knots saves computer time, both in fitting the surfaces and in subsequent interrogation of the fitted surfaces. Moreover, it has helped to stabilise the values of the signal, which now show very systematic variation throughout the year.

10. Uses GCVGML to calculate values of the GCV for each month, as a function of the smoothing parameter, in the file tmaxb.gcv. The optimisation parameters required for

this calculation have been obtained from the optimisation parameters file tmaxb.opt, as produced by command 9.

11. Uses LAPGRD to calculate grids of mean daily maximum temperature and standard errors for the months of January and July. LAPGRD uses the surface coefficients in tmaxb.sur and the error covariance matrices in tmaxb.sur calculated by SPLINB in command 9. The mean values for the two grids are quite similar to the means for the corresponding surface grids calculated by LAPGRD in command 6. The standard errors of the means have been halved.

12. Uses ADDNOT to add knots to the knot file tmaxsel.not calculated by SELNOT in command 8. The new knot file is tmaxad1.not. In this case the knots are provided interactively in the command file. They have been obtained, using a text editor, from the largest 10 residuals from the ranked residual list in tmaxb.res, as produced by SPLINB in command 9. These points could also have been obtained from the same ranked residual list at the bottom of tmaxb.log. Points with negative index are already knots and are ignored. Thus only 5 points are selected as additional knots from this list.

13. Uses ADDNOT to add knots to the knot file tmaxsel.not calculated by SELNOT in command 8. The new knot file is tmaxad2.not. In this case the knots are selected directly from the ranked residual list in tmaxb.res, as produced by SPLINB in command 9. The user simply specifies the number of new knots to be added. In this case 5 additional knots are requested and the two output knots files tmaxad1.not and tmaxad2.not are identical.

14. Uses DELNOT to revise the knot file tmaxsel.not after two data points have been removed from the data file tmax.dat. The indices of these points are provided interactively in the command file, in order of increasing index.

Tri-variate spline smoothing of monthly mean precipitation data using knots and the square root transformation

To illustrate the tri-variate spline smoothing of monthly mean precipitation data, four data files are supplied in the test/rain subdirectory:

rain.dat - monthly mean precipitation data

tas4.dem - small DEM in Arc/Info ASCII grid format

tas4.img - same DEM as in tas4.dem, in Idrisi ASCII image format

tas4.doc - standard Idrisi documentation file for tas4.img

Five ANUSPLIN command files for processing these data files are provided in the test/rain subdirectory and are listed in the table below. Each command and its outputs are discussed in the notes following the table. All output files are provided in the test/rain/out subdirectory.

Command	Input Files	Output Files
1. selnot < rainsel.cmd > rainsel.log	rain.dat	rainsel.log rainsel.not rainsel.rej
2. splinb < rain.cmd > rain.log	rain.dat rainsel.not rain.val	rain.log rain.res rain.opt rain.sur rain.cov rain.flg rain.lis rain.out
3. gcvgml < raingcv.cmd > raingcv.log	rain.opt	raingcv.log
4. lapgrd < rainimg.cmd > rainimg.log	rain.sur rain.cov tas4.doc tas4.img	rainimg.log rain1.doc rain1.img rain7.doc rain7.img rcov1.doc cov1.img rcov7.doc rcov7.img
5. lapgrd < raingrd.cmd > raingrd.log	rain.sur rain.cov tas4.dem	rain1.grd rain7.grd rcov1.grd rcov7.grd

Notes

1. Uses SELNOT to select 150 knots from the 243 data points in rain.dat that lie within the specified coordinate limits. SELNOT provides a recommended range of number of knots, from 90 to 130, between approximately 1/4 and 1/3 of the number of valid data points. Indices and site names of the selected knots are stored in the file rainsel.not, for input to SPLINB in command 2. The independent and dependent variables are specified exactly as in the command file used for SPLINB in command 2. Positive margins of 3.0 for longitude and 2.0 for latitude are specified for both SELNOT and SPLINB.

2. Uses SPLINB to fit 12 approximate thin plate smoothing spline functions to 12 sets of monthly mean precipitation in the file rain.dat. The rainfall means are first transformed by the square root transformation. This should only be applied to data with naturally non-negative values. The square root rainfall means are weighted uniformly. The square root transformation reduces the skew in the data and has been found by Hutchinson (1998b) to reduce overall error when interpolating daily precipitation data. The effect of using the square root transformation is to apply more smoothing to large rainfall data values, and less smoothing to small rainfall data values.

The approximate splines are constructed from the knots in the file rainsel.not produced by command 1. Three points in data file rain.dat lie outside the specified X,Y limits. The data are read using a FORTRAN format statement. Summary statistics in the log file are calculated in terms of the square root analysis, and approximate statistics for the untransformed rainfall values are also provided. Validation statistics are calculated in terms of square roots and in terms of untransformed values.

The log file shows that the signals of the fitted surfaces vary between 87 and 104, generally less than the number of knots. There is a generally systematic progression in the signals from month to month. The square root of the GCV (RTGCV) varies between 5.4 mm in February (11% of the network mean) and 13.0 mm in July (13% of the network mean). As for the temperature analyses, these are conservative estimates of overall standard prediction error because they include the data error, as estimated by the procedure. The root mean square model error (RTMSE), which is an estimate of standard error after the estimated data error has been removed, is a more optimistic estimate of error. It ranges between 2.6 mm in February (6% of the network mean) and 6.3 mm in July (6% of the network mean). It would be reasonable to say that the true errors of the fitted surfaces are no more than 10%. Standard errors of around 10% are typical for surfaces fitted to monthly mean precipitation data with adequate network density. It is recommended to quote standard errors of precipitation in terms of percentages, because of the nature of the distribution of precipitation.

3. Uses GCVGML to calculate values of the GCV for each month, as a function of the logarithm to base 10 of the smoothing parameter. The optimisation parameters required for this calculation are read from the file rain.opt produced by command 2. The output GCV values are tabulated in the output text file rain.gcv. The GCV values in this table show unimodal minimums for each surface. This in part reflects the robustness of SPLINB analyses with knots.

4. Uses LAPGRD to calculate the grid files rain1.img, rain7.img , rcov1.img and rcov7.img, in Idrisi image format, of mean precipitation and standard errors for the months January and July. The grids depend on the surface coefficients in the file rain.sur and the covariance matrices in rain.cov produced by command 2 and the small DEM, in Idrisi image format, in the file tas4.img. The log file concludes with summary

statistics of the output grid files. The Idrisi grid documentation files rain1.doc, rain7.doc, rcov1.doc and rcov7.doc are produced automatically by LAPGRD.

5. Uses LAPGRD to calculate the grid files rain1.grd, rain7.grd, rcov1.grd and rcov7.grd, in Arc/Info ASCII grid format, of mean precipitation and standard errors for the months January and July. The grids depend on the surface coefficients in the file rain.sur and the covariance matrices in rain.cov produced by command 2 and the small DEM, in Arc/Info ASCII grid format, in the file tas4.dem. LAPGRD produces grids of untransformed rainfall by squaring the interpolated square root values. Grids of the square root values can also be produced by LAPGRD. The log file concludes with summary statistics of the output grid files.

Bi-variate and tri-variate spline smoothing of monthly mean solar radiation data using surface independent variables

To illustrate bi-variate and tri-variate spline smoothing of monthly mean solar radiation data, five data files are supplied in the test/rad subdirectory:

rainrad.dat - monthly mean solar radiation and precipitation data

rainrad.val - validation monthly mean solar radiation and precipitation data

tas4.dem - small DEM in Arc/Info ASCII grid format

rain1.grd - precipitation grid in Arc/Info ASCII grid format

rain7.grd - precipitation grid in Arc/Info ASCII grid format

The two precipitation grids are calculated in the test/rain/out subdirectory by command 5 of the preceding set of examples. Five ANUSPLIN command files for processing these data files are provided in the test/rad subdirectory and are listed in the table below. Each command and its outputs are discussed in the notes following the table. All output files are provided in the test/rad/out subdirectory.

Command	Input files	Output files
1. splina < rad.cmd > rad.log	rainrad.dat rainrad.val	rad.log rad.res rad.opt rad.sur rad.lis rad.cov rad.out
2. lapgrd < radgrd.cmd > radgrd.log	rad.sur rad.cov tas4.dem	radgrd.log rad1.grd rad7.grd rad1.err rad7.err
3. splina < rainrad.cmd > rainrad.log	rainrad.sur rainrad.val	rainrad.log rainrad.res rainrad.opt rainrad.sur rainrad.lis rainrad.cov rainrad.out
4. gcvgml < rradgcv.cmd > rradgcv.log	rainrad.opt	rradgcv.log rainrad.gcv
5. lapgrd < rradgrd.cmd > rradgrd.log	rainrad.sur rainrad.cov rain1.grd rain7.grd	rradgrd.log rrad1.grd rrad7.grd rrad1.err rrad7.err

Notes

1. Uses SPLINA to fit 12 bi-variate thin plate smoothing spline functions to 12 sets of monthly mean solar radiation data in the file rainrad.dat. Surface coefficients are written to the file rad.sur. Error covariance matrices of the surface coefficients are written to the file rad.cov. Separate validation data have been provided in the input file rainrad.val. The root mean square residuals from the surfaces of the validation data are shown under RMS in the output log file. These residuals are in good agreement with the square roots of the GCV, especially in the winter months. Fitted values at the validation points are written to the file rad.out in a standard format.

2. Uses LAPGRD to calculate the surface grid files rad1.grd and rad7.grd, and prediction standard error grids in rad1.err and rad7.err for the mid-summer and mid-winter months January and July. All grids are written in Arc/Info ASCII grid format. The grids depend on the surface coefficients in the file rad.sur and the error covariance matrices in the file rad.cov, as produced by command 1. Since the bi-variate solar radiation surfaces do not depend on elevation, the DEM in Arc/Info ASCII grid format in the file tas4.dem is used to mask the output grids so that points are only calculated over land. The log file concludes with summary statistics of all output grid files.

3. Uses SPLINA to fit 12 tri-variate partial thin plate smoothing spline functions to 12 sets of monthly mean solar radiation data in the file rainrad.dat. Monthly mean rainfall, transformed by the *tanh* function, is used as a surface independent covariate. This variable varies systematically from month to month. Surface coefficients are written to the file rainrad.sur. Error covariance matrices of the surface coefficients are written to the file rainrad.cov. Separate validation data have been provided in the input file rainrad.val. The root mean square residuals from the surfaces of the validation data are shown under RMS in the output log file. The RTGCV and RMS values show agreement similar to that for the bi-variate analysis in command 1. Fitted values at the validation points are written to the file rainrad.out in a standard format.

The dependence on transformed rainfall allows for the known dependence of solar radiation on cloud associated with rainfall, giving rise to more complex solar radiation patterns in areas with complex terrain (Hutchinson *et al.* 1984a). Signals of the fitted surfaces show a more consistent progression over the months than for the bi-variate analysis. Validation residuals in the summer months are slightly less than corresponding validation residuals for the bi-variate analysis, while validation residuals in the winter months are slightly larger. Some remaining data errors may affect these comparisons. There is some inconsistency in the coefficients of the covariates for different months.

4. Uses GCVGML to calculate values of the GCV for each month, as a function of the logarithm, to base 10, of the smoothing parameter. The optimisation parameters required for this calculation are read from the file rainrad.opt produced by command 3. The output GCV values are tabulated in the output text file rainrad.gcv. The GCV values in this table show multi-modal behaviour for some months, indicative of remaining data errors, and perhaps deficiencies in the modelled dependence on rainfall.

5. Uses LAPGRD to calculate the surface grid files rrad1.grd and rrad7.grd, and prediction standard error grids in rrad1.err and rrad7.err for the mid-summer and mid-winter months January and July. All grids are written in Arc/Info ASCII grid format. The grids depend on the surface coefficients in the file rainrad.sur and the error covariance matrices in the file rainrad.cov, as produced by command 3, as well as the rainfall grids rain1.grd and rain7.grd for the months of January and July. The rainfall grids are the grids produced by command 5 of the preceding group of examples of

rainfall analyses. Since the rainfall grids automatically depend on elevation, there is no need to use a mask grid in this case to ensure that grid points are only calculated over land. The log file concludes with summary statistics of all output grid files.

REFERENCES

Bates D and Wahba G. 1982. Computational methods for generalised cross validation with large data sets. In: Baker CTH and Miller GF (eds). *Treatment of Integral Equations by Numerical Methods*. New York: Academic Press: 283-296.

Bates D, Lindstrom M, Wahba G and Yandell B. 1987. GCVPACK − routines for generalised cross validation. *Commun. Statist. B − Simulation and Computation* 16: 263-297.

Craven P and Wahba G. 1979. Smoothing noisy data with spline functions. *Numerische Mathematik* 31: 377-403.

Dongarra, JJ., Moler, CB., Bunch, JR. and Stewart GW. 1979. *LINPACK Users' Guide*. SIAM, Philadelphia.

Elden L. 1984. A note on the computation of the generalised cross-validation function for ill-conditioned least squares problems. *BIT* 24: 467-472.

Houser P, Hutchinson MF, Viterbo P, Hervé Douville J, and Running SW 2004. Terrestrial data assimilation. In: *Vegetation, Water, Humans and the Climate. Global Change - The IGB Series.* Kabat, P. *et al.* (eds). Springer, Berlin.

Hutchinson MF. 1984. A summary of some surface fitting and contouring programs for noisy data. *CSIRO Division of Mathematics and Statistics, Consulting Report ACT* 84/6. Canberra, Australia.

Hutchinson MF. 1988. Calculation of hydrologically sound digital elevation models. *Third International Symposium on Spatial Data Handling.* Columbus, Ohio: International Geographical Union: 117-133.

Hutchinson MF. 1989a. A new objective method for spatial interpolation of meteorological variables from irregular networks applied to the estimation of monthly mean solar radiation, temperature, precipitation and windrun. *CSIRO Division of Water Resources Tech. Memo.* 89/5*:* 95-104.

Hutchinson MF. 1989b. A new procedure for gridding elevation and stream line data with automatic removal of spurious pits. *Journal of Hydrology* 106: 211-232.

Hutchinson MF. 1991a. The application of thin plate smoothing splines to continent-wide data assimilation. In:. Jasper JD (ed.) *BMRC Research Report No.27, Data Assimilation Systems*. Melbourne: Bureau of Meteorology: 104-113.

Hutchinson MF. 1991b. Climatic analyses in data sparse regions. In:. Muchow RC and. Bellamy JA (eds). *Climatic Risk in Crop Production*, CAB International, 55-71.

Hutchinson MF. 1993. On thin plate splines and kriging. In: Tarter ME and Lock MD.(eds). *Computing and Science in Statistics* 25. University of California, Berkeley: Interface Foundation of North America: 55-62.

Hutchinson MF. 1995. Interpolating mean rainfall using thin plate smoothing splines. *International Journal of Geographic Information Systems* 9: 305-403.

Hutchinson MF. 1996. A locally adaptive approach to the interpolation of digital elevation models. *Third Conference/Workshop on Integrating GIS and Environmental Modeling*. Santa Barbara: NCGIA, University of California. http://www.ncgia.ucsb.edu/conf/SANTA_FE_CD-ROM/santa_fe.html.

Hutchinson MF. 2001. ANUDEM Version 5.0. Centre for Resource and Environmental Studies, Australian National University, Canberra. http://cres.anu.edu.au/outputs/anudem.php

Hutchinson MF. 1998a. Interpolation of rainfall data with thin plate smoothing splines: I two dimensional smoothing of data with short range correlation. *Journal of Geographic Information and Decision Analysis* 2(2): 152-167. http://www.geodec.org/gida_4.htm

Hutchinson MF. 1998b. Interpolation of rainfall data with thin plate smoothing splines: II analysis of topographic dependence. *Journal of Geographic Information and Decision Analysis* 2(2): 168-185. http://www.geodec.org/gida_4.htm

Hutchinson MF. and Bishof RJ. 1983. A new method for estimating the spatial distribution of mean seasonal and annual rainfall applied to the Hunter Valley, New South Wales. *Australian Meteorological Magazine* 31: 179-184.

Hutchinson MF, Booth TH, Nix HA and McMahon JP. 1984a. Estimating monthly mean values of daily total solar radiation for Australia. *Solar Energy* 32: 277-290.

Hutchinson MF., Kalma JD and Johnson ME. 1984b. Monthly estimates of wind speed and wind run for Australia. *Journal of Climatology* 4: 311-324.

Hutchinson MF. and de Hoog FR. 1985. Smoothing noisy data with spline functions. *Numerische Mathematik* 47: 99-106.

Hutchinson MF. Nix HA. and McMahon JP. 1992. Climate constraints on cropping systems. In: Pearson CJ. (ed), *Ecosystems of the World, 18 Field Crop Ecosystems.* Amsterdam: Elsevier: 37-58.

Hutchinson MF. and Gessler PE. 1994. Splines – more than just a smooth interpolator. *Geoderma* 62: 45-67.

Hutchinson MF. Nix HA, McMahon JP. and Ord KD. 1996a. The development of a topographic and climate database for Africa. In: Proceedings of the Third International Conference/Workshop on Integrating GIS and Environmental Modeling, NCGIA, Santa Barbara, California. http://www.ncgia.ucsb.edu/conf/SANTA_FE_CD-ROM/santa_fe.html

Hutchinson MF., Belbin L., Nicholls AO., Nix HA., McMahon J.P. and Ord KD. 1996b. *Rapid Assessment of Biodiversity, Volume Two, Spatial Modelling Tools*. The Australian BioRap Consortium, Australian National University, 142pp.

Kesteven JL. and Hutchinson MF. 1996. Spatial modelling of climatic variables on a continental scale. In: Proceedings of the Third International Conference/Workshop on Integrating GIS and Environmental Modeling, NCGIA, Santa Barbara, California. http://www.ncgia.ucsb.edu/conf/SANTA_FE_CD-ROM/santa_fe.html

Kohn,R., Ansley, CF. and Tharm, D. 1991. The performance of cross-validation and maximum likelihood estimators of spline smoothing parameters. *Journal American Statistical Association* 86: 1042-1049.

Price DT., McKenney DW., Nalder IA., Hutchinson MF. and Kesteven JL. 2000. A comparison of two statistical methods for spatial interpolation of Canadian monthly mean climate date. *Agricultural and Forest Meteorology* 101: 81-94.

Schimek, MG. (ed) 2000. *Smoothing and Regression: approaches, computation and application*. John Wiley & Sons, New York.

Silverman BW. 1985. Some aspects of the spline smoothing approach to nonparametric regression curve fitting (with discussion). *Journal Royal Statistical Society Series B* 47: 1-52.

Wahba G. 1979. How to smooth curves and surfaces with splines and cross-validation. *Proc. 24th Conference on the Design of Experiments*. US Army Research Office 79-2, Research Triangle Park, NC: 167-192.

Wahba G. 1983. Bayesian confidence intervals for the cross-validated smoothing spline. *Journal Royal Statistical Society Series B* 45: 133-150.

Wahba G. 1985. A comparison of GCV and GML for choosing the smoothing parameter in the generalized spline smoothing problem. *Annals of Statistics* 13: 1378-1402.

Wahba G. 1990. *Spline Models for Observational Data*. CBMS-NSF Regional Conference Series in Applied Mathematics 59, SIAM, Philadelphia, Pennsylvania.