JOURNAL OF GEOGRAPHIC INFORMATION AND DECISION ANALYSIS

Journal of Geographic Information and Decision Analysis, vol. 2, no. 2, pp. 139-151, 1998

Interpolation of Rainfall Data with Thin Plate Smoothing Splines - Part I: Two Dimensional Smoothing of Data with Short Range Correlation

Michael F. Hutchinson

Centre for Resource and Environmental Studies, Australian National University, Canberra, ACT 0200, Australia

hutch@cres.anu.edu.au

ABSTRACT **Two dimensional thin plate smoothing splines were used to interpolate 100 daily rainfall values, with the degree of data smoothing determined by minimizing the generalised cross validation. Analyses were performed on the square roots of the rainfall values, permitting robust calibration of spatially distributed standard errors which are correlated with rainfall amount. Initial model calibration was made difficult by apparent short range spatial correlation in the data. This was overcome by removing one point from each of the ten closest pairs of data points, indicating that this correlation had a range of less than 10 km. A companion paper shows that the short range correlation can be associated with topographic effects. The error analyses were confirmed by comparing predictive accuracies on 367 withheld data points. The validity of using the square root transformation was also confirmed.**

KEYWORDS: **thin plate smoothing splines, generalised cross validation, data smoothing, standard errors, short range correlation, correlated errors, rainfall data, square root transformation**

Contents

11. Conclusion
References

---

1. Introduction and methodology

Rainfall, particularly at the daily time scale, typically displays complex spatial patterns. These are often related to topography and prevailing wind direction, but calibration of such factors can be difficult. This paper examines two dimensional analyses of daily rainfall data, using thin plate smoothing splines, as implemented in the ANUSPLIN package (Hutchinson 1997). A companion paper (Hutchinson 1998) uses partial thin plate smoothing splines to examine topographic dependencies of the same data.

Thin plate smoothing splines are commonly applied to smooth multivariate interpolation of irregularly scattered noisy data. Early applications to meteorological data were presented by Wahba and Wendelberger (1980) and Hutchinson and Bischof (1983), following development of the basic methodology by Wahba (1979) and Bates and Wahba (1982). A detailed account of thin plate smoothing splines, and various generalisations, may be found in Wahba (1990). Further applications to climate interpolation have been described by Hutchinson (1991) and further methodological developments, suitable for geographic applications, have been presented by Gu and Wahba (1993) and Mitasova and Mitas (1993).

Comparisons with geostatistical techniques, with which splines share close formal connections, have been presented by Hutchinson (1993) and Hutchinson and Gessler (1994). A feature of thin plate smoothing spline analyses is their operational simplicity, since they do not require separate calibration of spatial covariance structure. Ordinary kriging normally requires prior calibration of a variogram with three parameters - range, nugget and sill (Cressie 1991; Wackernagel 1995). Once the range is defined, predicted values depend only on the ratio of the nugget to the height to sill, but the actual values of both nugget and sill are required to estimate spatially distributed standard errors of the fitted model. The range parameter can be difficult to calibrate, even from simulated data (Dietrich and Osborne 1991).

Thin plate smoothing splines, which are unaffected by uniform scaling of the independent variables by an arbitrary non-zero constant, enjoy the significant practical advantage of having no range parameter (Hutchinson 1993). This makes the associated covariance structure more robustly determined when data are limited. Splines are calibrated by optimising a single parameter, the smoothing parameter, to determine the degree of data smoothing. This is usually done by minimising the generalised cross validation (GCV), first introduced by Craven and Wahba (1979).

The GCV is a direct measure of the predictive error of the fitted surface, calculated by removing each data point in turn, and forming a weighted sum of the square of the discrepancy of each omitted data point from a surface fitted to all other data points. The GCV may be calculated implicitly, and hence efficiently, using the "leaving out one" lemma (Craven and Wahba 1979, lemma 3.1). Minimising the GCV has been found to yield good results with simulated and actual data, although it can yield undersmoothing of data with correlated errors (Diggle and Hutchinson 1989). Applications of GCV to large scale data smoothing problems have been described by Golub and von Matt (1997).

Spatially distributed standard errors for thin plate smoothing splines can be estimated by applying standard geostatistical techniques to the spline model (Wahba 1983, Silverman 1985; Hutchinson 1993). These estimates are scaled by the nugget variance, estimated in the case of splines by analogy with least squares regression analysis (Wahba 1990).

A useful diagnostic associated with thin plate smoothing splines is the signal, or effective degrees of freedom of the model, as estimated by the trace of the influence matrix associated with the fitted spline. Hutchinson and Gessler (1994) present evidence to suggest that this should be no greater than about half of the number of data points. Signals larger than this are indicative of either insufficient data or short range correlation in the data values. Thin plate spline analyses can be extended to incorporate correlation in the data (Diggle and Hutchinson 1989; Hutchinson 1995), but a simple alternative approach is employed in this paper by omitting one point from each of the closest pairs of data points. The spacing of these pairs can give insight into the range of the short range correlation structure.

2. Features of the rainfall data

Three features of the rainfall data set deserve attention. Firstly the small size of the data set places limits on the complexity of proposed spatial analyses, such as the incorporation of topographic effects. This is examined further in the companion paper.

Secondly, from the point of view of two dimensional spline analyses, errors in the positions of the data points should be considered, particularly if short range correlation is evident in the data. Positions have been derived from positions recorded in geographic coordinates, usually recorded to the nearest minute of latitude and longitude. This implies that the standard error in data point position is around 1 km. The two closest data pairs have separations of just 1.1 km.

Finally, the nature of variability of rainfall should be considered. Daily rainfall distributions are typically skewed, with respect to both space and time. The transport mechanisms associated with rainfall imply that distributions in space and time are closely intertwined. It is well known that skewness can be greatly reduced by taking rainfall to a small power. Stidd (1973) suggested the cube root as a universal means of reducing observed rainfall totals to the upper tail of a normal distribution. Hutchinson *et al.* (1993) have shown that this power should be closer to the square root, at least for daily rainfall data across the USA. The square root transformation was therefore applied to all rainfall data values before performing the spatial analyses. Figure 1 shows that the square root transformation significantly reduced the skew in the raw data.

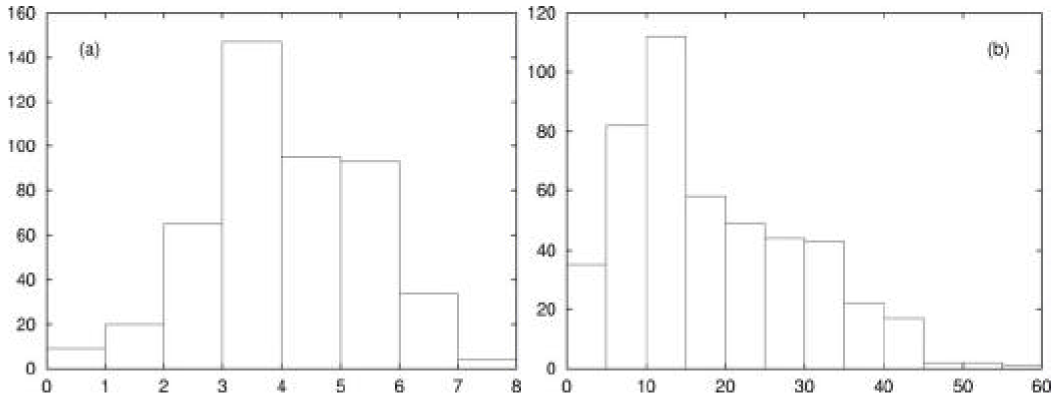

Figure 1 Histograms using all 467 data points of (a) the square root rainfall data, (b) the raw rainfall data.

3. Error analysis of the square root transformation

An immediate consequence of applying the square root transformation to the rainfall data is that estimated errors in the squared interpolated values are positively correlated with the rainfall values. Squaring the interpolated square root values also introduces a small positive bias. This bias is ignored.

The square root transformation permits a simple analysis of standard error. Suppose that the true value of rainfall at a point is $R = Z^2$ and that Z is estimated by X with error e so that

$$X = Z + e. \tag{1}$$

Suppose also that e is normally distributed with zero mean and variance s^2, so that the standard error in X is s . The variance of X^2 is then given by

$$var(X^2) = var(Z^2 + 2Ze + e^2) = 4Z^2s^2 + 2s^4. \tag{2}$$

Replacing Z by the estimate X, an estimate of the standard error in X^2, as an approximately unbiased estimate of R, is then given by

$$SE(X^2) = (4X^2s^2 + 2s^4)^{1/2} = s\ (4X^2 + 2s^2)^{1/2}. \tag{3}$$

Equation (3) was used to calculate spatially distributed standard error estimates of the estimated untransformed rainfall values. This equation shows explicitly how these estimates are positively correlated with rainfall amount. When X is negative, then X is replaced by zero. The second term under the square root in equation (3) is negligible except when X is close to zero or s is relatively large. This gives a simple approximate standard error estimate of $2Xs$, implying that smaller rainfall values should be estimated with better *absolute* accuracy than larger rainfall values. The corresponding relative standard error estimate is then given by

$$RE(X^2) = 2s/X \tag{4}$$

implying that larger rainfall values should be estimated with better *relative* accuracy than smaller rainfall values. This relative error estimate is exactly twice the relative error in the square root rainfall estimate.

4. Analysis of short range correlation

Initial application of the procedure SPLINA to fit a second order thin plate smoothing spline to the data set yielded exact interpolation. This occurred with both transformed and untransformed data. It was reasoned that this was due to short range correlation in the data. This has been overcome by Hutchinson and Gessler (1994), in their analysis of soil electrical conductivity data, by successively removing points from the closest data point pairs, until the closest spacings exceeded the range of the apparent short range correlation. Once sufficiently many close pairs were removed, the procedure applied reasonable data smoothing and the *signal* (the trace of the influence matrix) reduced to less than half the number of data points. The estimated standard deviation of the model error also approached estimates of the nugget standard deviation obtained by standard variogram analysis.

For the present study, the SELNOT procedure was used to successively remove points from the closest pairs in the data set. This procedure is normally used in the ANUSPLIN package to select knots for an approximate thin plate spline procedure, used for analysing larger data sets. Applying the SPLINA procedure to the square root rainfall data, the signal, the estimated standard deviation of the model error and the GCV were monitored as data points were removed and minimum data point separation increased. These statistics are listed in Table 1 for 0, 5, 10 and 15 removed data points. Table 1 also lists the root mean square residual of the fitted surfaces from the withheld 367 data points. The estimated standard deviation of model error and the root mean square residual from the 367 withheld data points are also plotted in Figure2 as a function of number of points omitted.

Table 1 Statistics of second order thin plate smoothing analyses of the square root rainfall data as data points are removed to increase minimum data point separation.

Number of points removed	Minimum data point separation (km)	Signal	Square root of the GCV ($mm^{1/2}$)	Estimate of standard deviation of model error ($mm^{1/2}$)	RMS residual from 367 data points ($mm^{1/2}$)
0	1.1	100	0.425	0.0	0.712
5	4.5	75.2	0.745	0.340	0.670
10	7.3	61.2	0.799	0.452	0.648
15	8.3	60.4	0.835	0.449	0.657

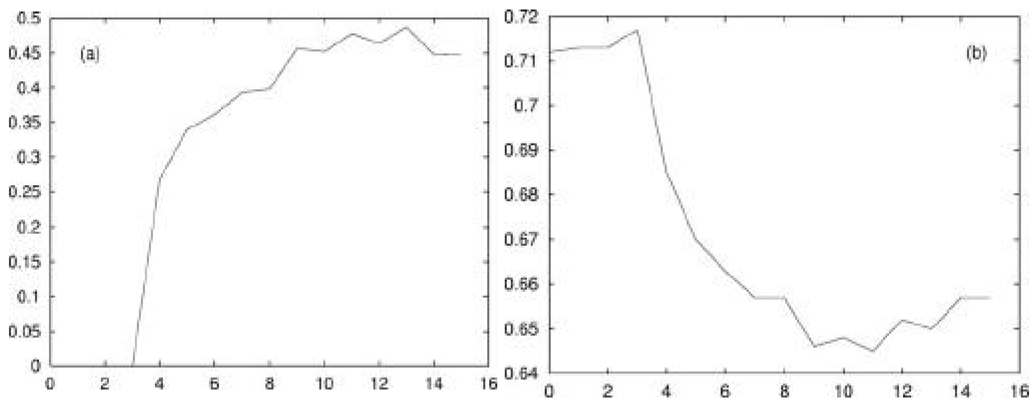

Figure 2 Estimated standard deviation of model error (a) (nugget standard deviation) and (b) RMS validation error from 367 withheld data points, as a function of number of data points removed.

Exact interpolation was avoided after four data points were removed and the minimum data separation exceeded 4.0 km. Estimates of the signal and the standard deviation of model error (nugget standard deviation) stabilised after a total of 10 points were removed and the minimum data point separation was 7.3 km. The signal for this surface was 61.2, acceptably close to the usually recommended upper limit of 45 (half the number of data points). This surface was chosen for detailed analysis and examination of residuals from the withheld 367 data points. The IDs of the 10 points withheld from the spline analysis were, in order of removal, 341, 37, 458, 369, 192, 188, 292, 372, 442, 281. It was concluded that the range of the short range correlation was certainly less than 10 km, perhaps consistent with an exponential model with a range of around 5 km.

Examination of Table 1 and Figure 2 shows that the best RMS validation residuals were obtained by the surfaces with the largest amount of data smoothing. They also show that the GCV was not a reliable estimate of model error in the presence of short range correlation. On the other hand, both the signal and the estimated standard deviation of model error appear to be reliable indicators of the best number of close data points to be removed. The smallest RMS validation residual, obtained after 11 points were removed, with a minimum data spacing of 7.4 km, was associated with the first minimum of the signal and a local maximum of the estimated standard deviation of model error.

In fact, the best indicator of the number of close points to be removed, without the benefit of withheld data, appears to be the relative signal, defined as the ratio of the signal to the number of data points analysed. Each local minimum of the relative signal also occurred with a local minimum of the root mean square validation residual, and the absolute minimum of the relative signal occurred when 11 close data points were removed.

A similar thin plate spline analysis was applied directly to the untransformed data. The appropriateness of the square root transformation was indicated by slightly less systematic behaviour of the statistics as minimum data separation increased, although a clear absolute minimum and maximum respectively of the relative signal and the estimated standard deviation of model error were obtained after 13 data points were removed. More significantly, the RMS validation residuals were greater than the corresponding RMS validation residuals of the squares of the interpolated square root rainfall values by around 10 percent.

A third order, two dimensional, thin plate smoothing spline, in which the roughness penalty defining the interpolation is specified in terms of third order partial derivatives was similarly investigated. It is discussed in the companion paper. This model performed less well than the second order spline analysis, in terms of GCV and RMS validation residuals. This confirmed the author's experience that surface climate data rarely exhibit the high degree of spatial continuity implied by third and higher order thin plate splines.

5. Statistics

Most statistics and maps are presented in terms of the square root analysis and for the ensuing untransformed values, obtained by simply squaring all interpolated square root rainfall values. All rainfall values are given in units of mm and all square root rainfall values are given in units of $mm^{1/2}$.

General statistics are presented in Table 2 and histograms of all 467 estimated values are shown in Figure 3. These indicate close agreement with the means and medians, slight attenuation of the extreme rainfall values and reduction in standard deviation, particularly for the untransformed values. These are consistent with the process of data smoothing. Figure 3 also shows slight smoothing of the histograms of the data shown in Figure 1.

Table 2 General statistics for the withheld 367 data points and for the 367 estimated values.

Data	Minimum	Maximum	Mean	Median	Standard deviation
True square root values ($mm^{1/2}$)	0	7.19	4.07	4.02	1.39
Estimated square root values ($mm^{1/2}$)	0.53	6.73	4.02	3.99	1.24
True untransformed values (mm)	0	51.7	18.5	16.2	11.1
Estimated untransformed values (mm)	0.3	45.4	17.7	15.9	9.8

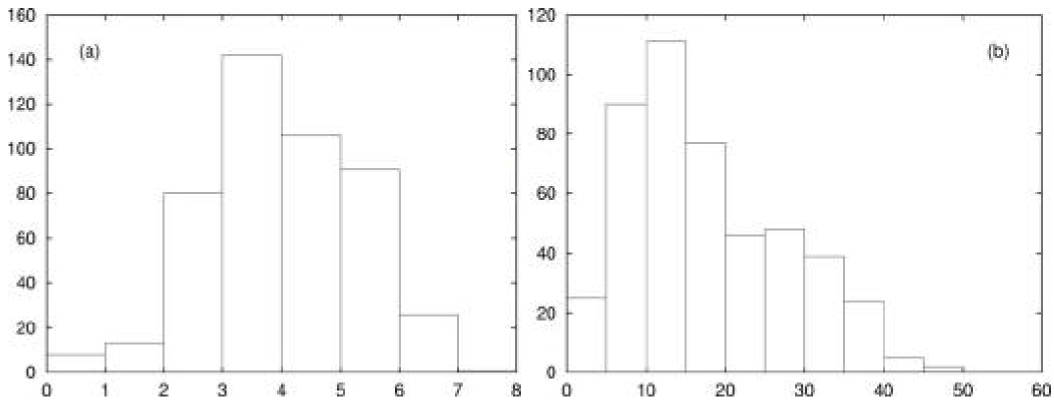

Figure 3 Histograms of (a) interpolated square root values and (b) squared interpolated square root values, for all 467 data points.

Comparative statistics, including the number of agreements with the actual 10 smallest and largest values, are presented in Table 3. Scatter plots of the residuals from the 367 data values are shown in Figure 4 and scatter plots of the 367 estimated values are shown in Figure 5.

The bias in the square root estimates is negligible, being just 1 percent of the mean. The bias in the untransformed estimates is just 5 percent of the mean. There is slight correlation of -0.45 between square root data residuals and actual values in Figure 4(a). This is consistent with two dimensional smoothing of the largest and smallest data values. Figure 5(a) shows uniform scatter of the estimated square root values about the square root data values, with correlation of 0.89.

The untransformed estimated values naturally introduce additional bias, due to the process of squaring the interpolated square root values. Accordingly, there is a greater correlation of -0.47 in the untransformed residual values in Figure 4(b) and Figure 5(b) shows non-uniform scatter about the true values, with correlation of 0.87.

Table 3 Comparative statistics for the 367 square root and untransformed rainfall data. The number of agreements with the 10 smallest and 10 largest points of all 467 data points.

Data	RMS Error	Bias	Mean Absolute Error	Mean Relative Error	Number in 10 smallest	Number in 10 largest
Square root ($mm^{1/2}$)	0.648	-0.05	0.49	0.13	8	3
Untransformed (mm)	5.60	-0.8	3.9	0.28		

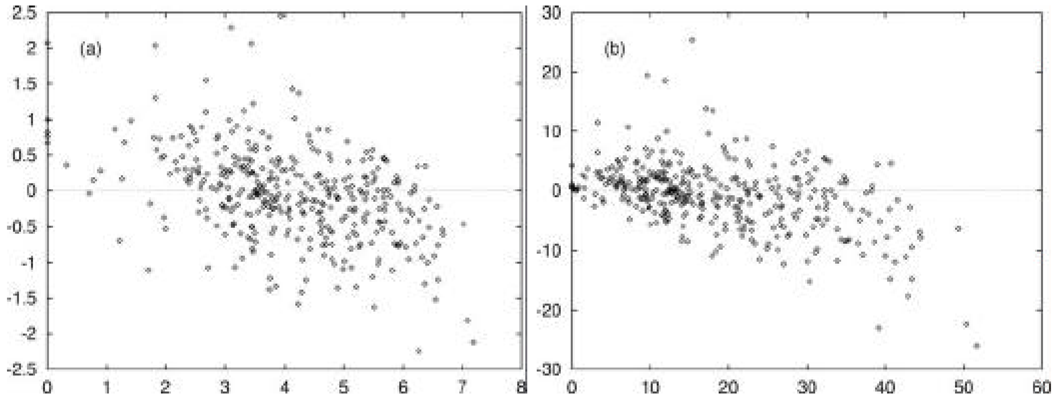

Figure 4 Scatter plot of 367 data residuals versus data values for (a) square root values (correlation = -0.45) and (b) untransformed values (correlation = -0.47).

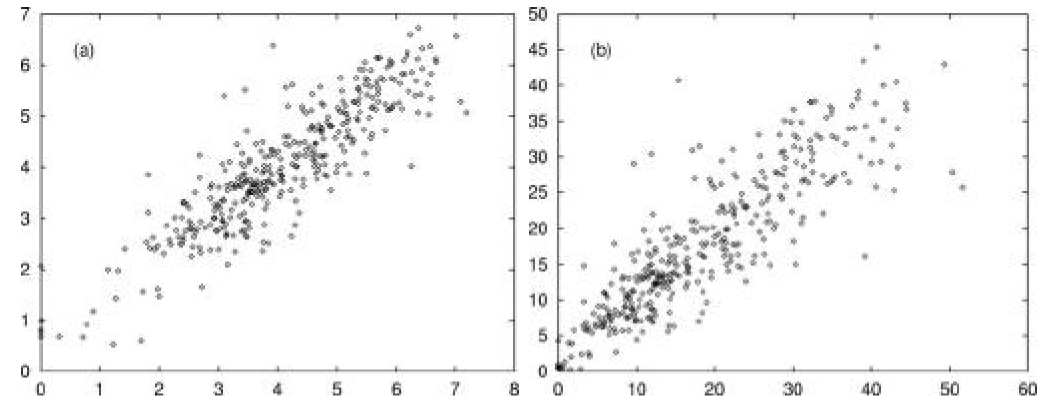

Figure 5 Scatter plot of 367 estimate values versus data values for (a) square root values (correlation = 0.89) and (b) untransformed values (correlation = 0.87).

6. Maps for the square root rainfall values

An isoline plot of the grid of standard errors of the interpolated square root values is shown in Figure 6. This grid is calculated by the ERRGRD program from the surface coefficients and error covariance matrix calculated by SPLINA, according to the method for estimating spatially distributed errors described in Hutchinson (1993). As in Figure 2 of Hutchinson and Gessler (1994), this map shows a smooth trend in estimated standard errors which are closely related to data network density.

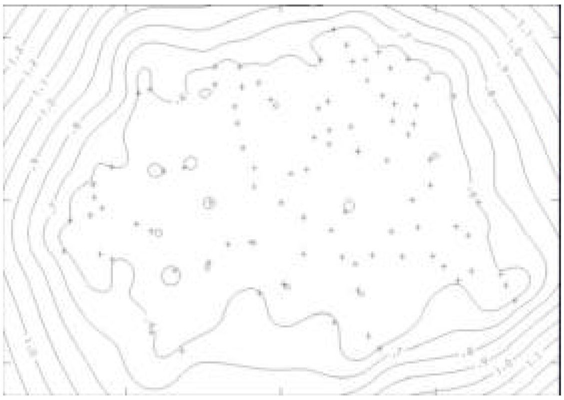

Figure 6 Estimated standard errors ($mm^{1/2}$) for the interpolated square root values overlaid with the positions of the 90 data points used to fit the surface.

Figure 7 Isolines of the interpolated square root values ($mm^{1/2}$).

Figure 7 shows an isoline plot of the grid of interpolated square root values. This grid is calculated by the LAPGRD program from the fitted surface coefficients. The even spacing of these isolines further reflects the appropriateness of the square root transformation.

7. Maps for the untransformed rainfall values

A proportional symbol map of the untransformed data residuals is shown in Figure 8, overlaid with an isoline plot of the grid of standard errors of the untransformed interpolated rainfall values. This grid is calculated from the grids shown in Figure 6 and 7 using equation (3) described above. The grid values, and corresponding isolines, are limited to those grid points for which the estimated standard error in Figure 6 does not exceed 0.9 $mm^{1/2}$. Interpolated values larger than the measured values are indicated by red dots and interpolated values less than the values are indicated by blue dots. The sizes of the residuals are in good agreement with the plotted standard error isolines, which reflect both data network density and the magnitude of the interpolated rainfall.

Figure 9 shows an isoline plot of the grid of untransformed interpolated values, overlaid with the symbols denoting the 10 smallest and 10 largest rainfall data values. This grid was calculated by squaring the interpolated square root values shown in Figure 7, and limiting grid values to the same grid positions as for Figure 8.

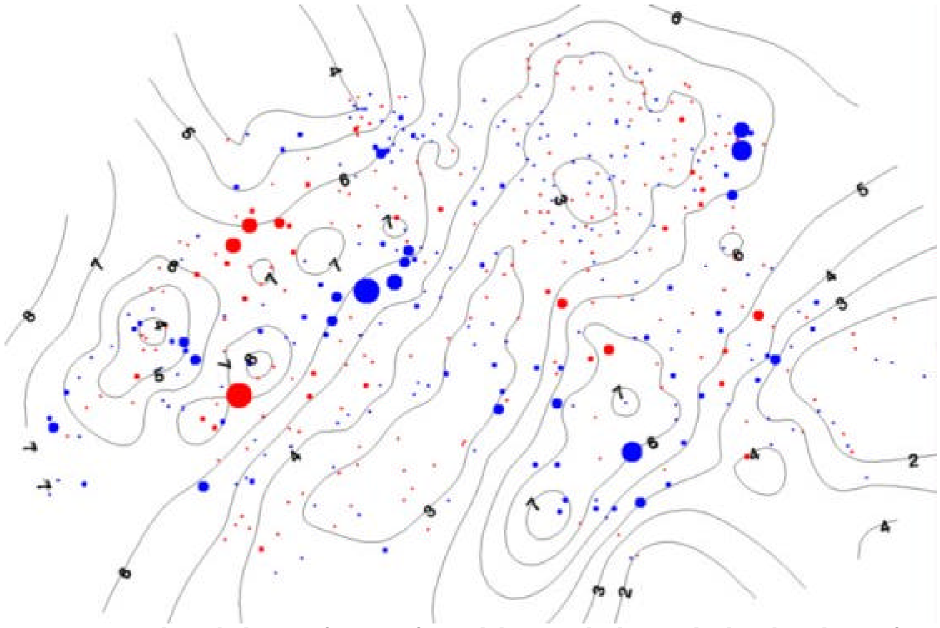

Figure 8 Proportional symbol map of untransformed data residuals, overlaid with isolines of estimated standard errors (mm).

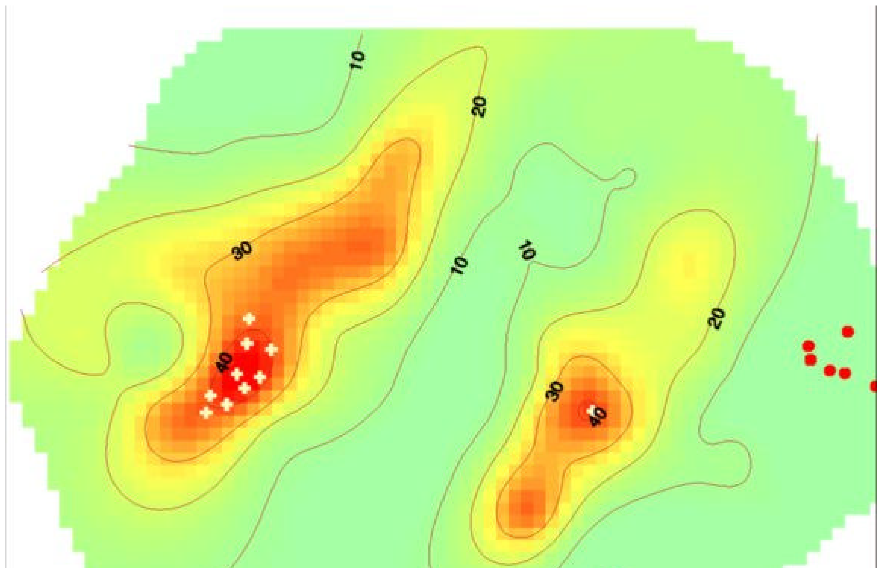

Figure 9 Isolines of the untransformed interpolated values (mm), overlaid with the 10 smallest and 10 largest estimated rainfall data values indicated by red dots and green crosses, respectively.

8. Interpolation accuracy

The accuracy of the interpolated values has already been partly assessed by plotting contours of the standard error estimates in Figure 8 and comparing them graphically with the data residuals. The largest residuals are clearly associated with the largest estimated standard errors. Normalised standard error estimates, obtained by dividing each data residual by the corresponding estimated standard error, showed reasonable agreement with a standard normal distribution. In view of the topographic analyses in the companion paper (Hutchinson 1998), the largest residuals are also associated with topographic aspect effects both north west and south east of The Alps.

The accuracy measures in Table 3 were calculated directly from the differences between the interpolated values and the actual values. The estimated statistics listed in Table 4 were calculated in the same way from the corresponding estimated standard errors. They show reasonable agreement with the actual statistics. In particular they verify, in average terms, the relative standard error formula given by equation (4).

Table 4 Estimates of comparative statistics for the 367 withheld data values.

Data	RMS Error	Mean Abs Err	Mean Rel Err
Square root ($mm^{1/2}$)	0.58	0.57	0.18
Untransformed (mm)	4.8	4.6	0.37

9. Performance

The model performed well in estimating central summary statistics of the withheld data, with biases of just 1 percent and 5 percent of the square root mean and the untransformed mean respectively. The medians were even closer, agreeing to within less than 1 and 2 percent respectively. The root mean square residuals were 16 percent and 31 percent of their respective data means.

Extremes of the withheld data were slightly attenuated, and standard deviations were also slightly reduced, to be expected when data are smoothed. There was strong agreement between the actual and estimated 10 smallest values. The agreement was less strong for the 10 largest values. This is entirely consistent with the standard error analysis given by equations (3, 4), which indicate that smaller values are estimated more accurately than larger values, in absolute terms, even though they are estimated less well in relative terms. Nevertheless, the estimated 10 largest values were located in the same two regions occupied by the actual 10 largest values.

10. Emergency application of the methodology

The methodology is well suited to rapid application in emergency situations. As previously noted, thin plate smoothing splines are relatively simple to use, in part because they do not require separate calibration of spatial covariance structure. The programs in the ANUSPLIN package are computationally efficient, permitting application on modest workstations to data sets containing up to several thousand data points.

The initial square root transformation is simple and likely to be universally appropriate, given the tendency for the square roots of daily rainfall values to be approximately normally distributed. The estimation of the standard errors of the final untransformed interpolated rainfall values is also straightforward.

If closely spaced data points have to be omitted to remove the effects of short range correlation in the data, these may be readily determined using the SELNOT procedure. The criterion for assessing the range of the short range correlation is also simple and appears to be robust. This may not be necessary for larger data sets, since applying the method to the full data set yielded appropriate data smoothing without omitting closely spaced data points.

11. Conclusion

The thin plate spline model, applied to the square roots of the observed rainfalls, has provided point estimates which show good agreement with the withheld data, as measured by both summary comparative statistics and order statistics. The estimated values are consistent with the standard error analysis. The estimated comparative statistics also agree with the actual comparative statistics.

A simple criterion has been indicated for assessing the number of close pairs to be removed to remove the effects of short range correlation in the data. The criterion is to successively remove a point from the closest data pair until a minimum of the relative trace, or equivalently a minimum of the signal to noise ratio, is obtained. This criterion is verified in three dimensional analyses in the companion paper.

The results confirm the validity, and internal consistency, of the square root error model. Moreover, interpolating the square root values yielded more accurate estimates of the withheld 367 data values than directly interpolating the untransformed data values.

Higher order, two dimensional thin plate splines were found to perform less well, confirming experience with surface climate data. The effects of short range correlation in the data were readily identified and removed. The analysis is general and readily applied to emergency situations.

References

Bates, D. and Wahba, G. (1982) Computational methods for generalized cross-validation with large data sets. In: *Treatment of Integral Equations by Numerical Methods*, edited by C.T.Baker and G.F.Miller, pp. 283-296. London: Academic Press.

Cressie, N. A. C. (1991) *Statistics for Spatial Data.* New York: John Wiley & Sons .

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

Dietrich, C. R. and Osborne, M. R. (1991) Estimation of covariance parameters in kriging via restricted maximum likelihood. *Mathematical Geology,* 23; 119-135.

Diggle, P. J. and Hutchinson, M. F. (1989) On spline smoothing with autocorrelated errors. *Australian Journal of Statistics,* 31; 166-182.

Golub, G. and von Matt, U. (1997) Generalized cross-validation for large scale problems. *Journal of Computational and Graphical Statistics,* 6; 1-34.

Gu, C. and Wahba, G. (1993) Semiparametric analysis of variance with tensor product thin plate splines. *Journal of the Royal Statistical Society, Series B*, 55: 353-368.

Hutchinson, M. F. (1991) The application of thin plate splines to continent-wide data assimilation. In: *Data Assimilation Systems,* edited by J. D. Jasper, BMRC Research Report No. 27, pp. 104-113. Melbourne: Bureau of Meteorology.

Hutchinson, M. F. (1993) On thin plate splines and kriging. In: *Computing Science and Statistics, Vol. 25.* Edited by M.E.Tarter and M.D.Lock, pp. 55-62. Berkeley: Interface Foundation of North America.

Hutchinson, M. F. (1995) Interpolation of mean rainfall using thin plate smoothing splines. *International Journal of Geographical Information Systems,* 9; 385-403.

Hutchinson, M. F. (1997) *ANUSPLIN VERSION 3.2,*
http://cres.anu.edu.au/software/anusplin.html.

Hutchinson, M. F. (1998) Interpolation of rainfall data with thin plate smoothing splines: II analysis of topographic dependence. *This Volume*.

Hutchinson, M. F. and Bischof, R. J. (1983) A new method for estimating mean seasonal and annual rainfall for the Hunter Valley, New South Wales. *Australian Meteorological Magazine*, 31; 179-184.

Hutchinson, M. F. and Gessler, P. E. (1994) Splines - more than just a smooth interpolator. *Geoderma*, 62; 45-67.

Hutchinson, M. F., Richardson, C. W. and Dyke, P. T. (1993) Normalization of rainfall across different time steps. *Management of Irrigation and Drainage Systems*, 21-23 July, (1993, Park City, UT, pp. 432-439. Irrigation and Drainage Division, ASCE, US Department of Agriculture.

Mitasova, H. and Mitas, L. (1993) Interpolation by regularized spline with tension: I theory and implementation. *Mathematical Geology*, 25; 641-655.

Silverman, B. W. (1985) Some aspects of the spline smoothing approach to non-parametric regression curve fitting. *Journal Royal Statistical Society Series B*, 47; 1-52.

Stidd, C. K. (1973) Estimating the precipitation climate. *Water Resources Research*, 9; 1235-1241.

Wackernagel, H. (1995) *Multivariate Geostatistics. An Introduction with Applications*. Berlin: Springer Verlag.

Wahba, G. (1979) How to smooth curves and surfaces with splines and cross-validation. *Proceedings of the 24 th Design of Experiments Conference*, US Army Research Office, Report 79-2, pp. 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. (1990) *Spline Models for Observational Data*. CBMS-NSF Regional Conference Series in Applied Mathematics. Philadelphia: Society for Industrial and Applied Mathematics.

Wahba, G. and Wendelberger, J. (1980) Some new mathematical methods for variational objective analysis using splines and cross validation. *Monthly Weather Review*, 108; 1122-1143.

JGIDA vol. 2, no. 2

JGIDA Home