Department of Statistics

University of Wisconsin
1210 W. Dayton St.
Madison, WI 53706

TECHNICAL REPORT NO. 597

February 1980

SPLINE BASES, REGULARIZATION, AND GENERALIZED CROSS VALIDATION FOR SOLVING APPROXIMATION PROBLEMS WITH LARGE QUANTITIES OF NOISY DATA

by

Grace Wahba

This paper prepared for inclusion in Proceedings of the International Conference on Approximation Theory in Honor of George Lorenz, held January 8-11, 1980, at The University of Texas at Austin. Ward Cheney, Editor. To appear Academic Press.

Abstract

We consider the problem of estimating a smooth function $f(t)$, t in a bounded region in Euclidean d space, given observations $z_i = L_i f + e_i$, $i = 1,2,\ldots,n$, where the L_i are linear functionals and the e_i are random errors. A large number of practical problems can be placed in this context. We estimate f as the minimizer of $\frac{1}{n}\sum_{i=1}^{n}(L_i f - z_i)^2 + \lambda J_m(f)$, where, for $d = 2$, $J_m(f) = \iint \sum_{\nu=0}^{m} \binom{m}{\nu}\left(\frac{\partial^m f}{\partial x^\nu \partial y^{m-\nu}}\right)^2 dxdy$. The parameters λ and m are chosen from the data by the method of generalized cross validation. We cite a number of results concerning the theoretical properties of this estimate of f and present some numerical results concerning the smoothing of 500 millibar heights ($L_i f = f(x_i, y_i)$), with $n = 120$. The numerical method used is not very practical for n much larger than around 130. We propose an approximate method for computing the estimate of f which should be suitable for n several times as large.

SPLINE BASES, REGULARIZATION, AND GENERALIZED CROSS VALIDATION FOR SOLVING APPROXIMATION PROBLEMS WITH LARGE QUANTITIES OF NOISY DATA

Grace Wahba[1]

Department of Statistics
University of Wisconsin-Madison
Madison, Wisconsin

We consider the problem of estimating a smooth function $f(t)$, $t \in \Omega \subset R^d$ given noisy observations of linear functionals of f. In particular, one observes

$$z_i = L_i f + \varepsilon_i, \quad i = 1,2,\ldots,n,$$

where the ε_i are independent, zero mean random variables with common unknown variance σ^2. The L_i are assumed to be continuous linear functionals on a family $\mathcal{H}_m$, $m \geq m_o$ of reproducing kernel Hilbert spaces (rkhs) where m indexes the highest square integrable derivative (not necessarily an integer) possessed by functions in $\mathcal{H}_m$. In this setup, one seeks a regularized estimate $f_{n,\lambda}$ of f by solving the minimization problem: Find $f \in \mathcal{H}_m$ to minimize

$$\frac{1}{n}\sum_{i=1}^{n}(L_i f - z_i)^2 + \lambda J_m(f) \qquad (*)$$

where $J_m(\cdot)$ is a norm or seminorm in $\mathcal{H}_m$. The parameter λ controls the tradeoff between the infidelity

$$\frac{1}{n}\sum_{i=1}^{n}(L_i f - z_i)^2$$

of the approximate solution to the data, and the roughness $J_m(f_{n,\lambda})$ of

[1]This work supported by USARO Grant DAAG29-77-G-0207.

the solution. When a reproducing kernel for H_m is known, explicit solutions to this problem are well known. The method of generalized cross validation (GCV) has been shown to be an effective tool for choosing λ (as well as m) from the data. For applications to the solution of ill posed problems ($L_i f = \int K(t_i,s)f(s)ds$) see Wahba (1980) and references cited there, for applications to estimation of the derivative in one dimension, see Craven and Wahba (1979).

In this note we primarily consider the case $L_i f = f(t_i)$, $t_i \varepsilon R^2$, ($t_i = (x_i, y_i)$) and

$$J_2(f) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (f_{xx}{}^2 + 2f_{xy}{}^2 + f_{yy}{}^2)dxdy$$

or, more generally,

$$J_m(f) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \sum_{\nu=0}^{m} \binom{m}{\nu}\left(\frac{\partial^m}{\partial x^{\nu}\partial y^{m-\nu}} f(x,y)\right)^2 dxdy.$$

Then the minimizer, $f_{n,\lambda}$ of (*) is one of the "thin plate splines", so called because $J_2(f)$ is the bending energy of a thin plate.

Duchon (1976a) has given explicit formulae for $f_{n,\lambda}$ for the general d dimensional situation for m (integer or not) with $2m-d > 0$. (This condition is necessary for H_m to be an rkhs.) See also Meinguet (1978). Wahba (1979a) and Wahba and Wendelberger (1979) have shown how to use GCV to choose λ and m from noisy data, and have given a computational algorithm. Utreras Dias (1979) and Paihua Montes (1979), who along with Duchon are former students of P.J. Laurent, have also given computational procedures, and our work has benefited from the availability of their work. Wahba (1979b) has conjectured convergence results for the integrated mean square error (IMSE)

$$E \int_{\Omega} (f_{n,\lambda^*}(t) - f(t))^2 dt$$

in the noisy data case and has provided supporting numerical evidence for the conjectured rates. Here λ^* must tend to zero at the correct rate, and an estimate of the λ which minimizes the IMSE is provided by the GCV method.

Thin plate spline interpolating and smoothing methods (which involve minimization of a seminorm in an rkhs) are equivalent to a certain form of minimum variance unbiassed estimation on the homogeneous random fields of Matheron (1973). This was demonstrated in 1971 in Kimeldorf and Wahba (1971) in the one dimensional case for more general seminorms, and has been generalized by Duchon (1976b). Two dimensional interpolation and smoothing methods known as "kriging" in the mining industry are also based on minimum variance unbiassed estimation on Matheron's homogeneous random fields (see Delfiner (1978)). By comparing the form of the estimates in Delfiner (1978) and Duchon (1976a) one can see that, loosely speaking, the kriging estimates solve minimization problems in H_m for $m = 3/2, 5/2, \ldots$. The main difference between kriging as described by Delfiner and the thin plate splines we use below is in the choice of the free parameters (here λ and m) and how they are estimated. Delfiner uses (ordinary) cross validation as part of his estimation procedure.

Returning to interpolation case, where $f(t_i)$ is known exactly, Micchelli and Wahba (1979) have given lower bounds for best attainable convergence rates of $||f - f_{o,n}||_m$ for f "very smooth" in general rkhs, in terms of the rate of decay of the eigenvalues of the relevant reproducing kernel, and it appears to be establishable from Duchon (1978) that the best possible rates are achieved with thin plate spline interpolation for uniform rectangular or triangular arrays of points $t_1, t_2, \ldots, t_n$. This is in sharp contrast to interpolation in rkhs in two dimensions which are tensor products of two one dimensional rkhs.

In this latter case, uniform rectangular arrays of points do not lead to the best achieveable rates and better designs (ratewise) can be found. See Micchelli and Wahba (1979) and the examples in Delvos and Posdorf (1978) and Wahba (1979c).

In the remainder of this note we will discuss our computational experience and numerical results in smoothing 500 millibar height data from an irregular array of 120 North American weather stations, using thin plate splines and GCV. The numerical procedures we propose are very satisfactory on the University of Wisconsin-Madison UNIVAC 1110 for $n \leq 130$ or so. After describing our present results, we will propose an approximate method which will allow the calculations to be made easily for much larger data sets. The numerical details for the $n \leq 130$ case, and further numerical results may be found in Wahba (1979a) and Wahba and Wendelberger (1979). In summary, for $t = (x,y)$,

$$f_{n,\lambda}(t) = \sum_{i=1}^{n} c_i E_m(t - t_i) + \sum_{\nu=1}^{M} d_\nu \phi_\nu(t)$$

where $E_m(t) = |t|^{2m-2}\log|t|$, $|t| = \sqrt{x^2 + y^2}$; $\phi_1(t),\ldots,\phi_M(t) = 1, x, y, x^2, xy, \ldots, y^{m-1}$ where $M = \binom{m+1}{2}$ is the number of polynomials of total degree $\leq m - 1$. The vectors $c = (c_1,\ldots,c_n)'$ and $d = (d_1,\ldots,d_m)'$ are solutions to $(K + n\lambda I)c + Td = z$, $z = (z_1,\ldots,z_n)'$, $T'c = 0$, where K is the $n\times n$ matrix with ij^{th} entry $E_m(t_i - t_j)$, and T is the $n\times M$ matrix with $j\nu^{th}$ entry $\phi_\nu(t_j)$. To have a unique solution we need that the rank of T be M. Let U be any $n\times n - M$ matrix whose $n - M$ columns are orthogonal unit vectors orthogonal to the M columns of T, and let $I - A(\lambda) = n\lambda U(U'KU + n\lambda I)^{-1}U'$. The GCV estimate of λ is the minimizer of $V(\lambda) = \frac{1}{n}||(I - A(\lambda))z||^2/(\frac{1}{n}\mathrm{Tr}(I - A(\lambda)))$. See Wahba (1979a), Craven and Wahba (1979). It is known that $U'KU$ is strictly positive definite.

$V(\lambda)$ is easily computed for many values of λ and c and d are easily computed once the eigenvalue decomposition $\Gamma D \Gamma'$ of the $n - M \times n - M$ matrix $U'KU$ is computed. We get this decomposition using double precision EISPACK. The condition number of the interpolation problem is the condition number of D.

Figure 1 shows the location of 120 North American weather stations. Using a mathematical model of 500 mb. heights provided by Koehler (1979) a model f was generated whose contours are given by the dashed lines in Figure 2. Data z was obtained by evaluating f at the weather stations and adding a pseudo random normal error with realistic standard deviation 10 meters. The value $\hat{\lambda}$ minimizing $V(\lambda)$ for $m = 2,3,\ldots,7$ was obtained by global search and $m = 5$ was selected because $V(\hat{\lambda})$ was smallest for $m = 5$. The resulting $f_{n,\hat{\lambda}}$ with $m = 5$ satisfied

$$\sum_{i=1}^{n} (f_{n,\hat{\lambda}}(t_i) - f(t_i))^2 \simeq 1.01 \min_{\lambda} \sum_{i=1}^{n} (f_{n,\lambda}(t_i) - f(t_i))^2$$

and so $\hat{\lambda}$ was an excellent choice of λ in this (typical!) example. The contours of $f_{n,\lambda}$ are given in Figure 2 by solid lines.

The condition number of D was between 10^6 and 10^{10} depending on m, for these 120 data points. The computation of an interpolant ($\lambda = 0$) will become unstable with increasing n as the condition number approaches the range of double precision accuracy. The presence of a sufficiently positive λ eliminates this problem. However, when this condition number is large it is likely that smoothing problems with very large n can be satisfactorily solved in some restricted subspaces of H_m.

Let $X_n = \text{span}\,\{\sum_{i=1}^{n} u_{ji} E_m(t - t_i),\ j = 1,2,\ldots,n - M\}$ where $u_j = (u_{ji},\ldots,u_{jn})'$ is the j^{th} column of U. X_n is a proper $n - M$ dimensional subspace of M. If the Gram matrix $U'KU$ of this basis for X_n is machine rank deficient, it appears reasonable to minimize (*) in the union of an $N \ll n - M$ dimensional subspace of X_n and span $\{\phi_\nu\}_{\nu=1}^{M}$,

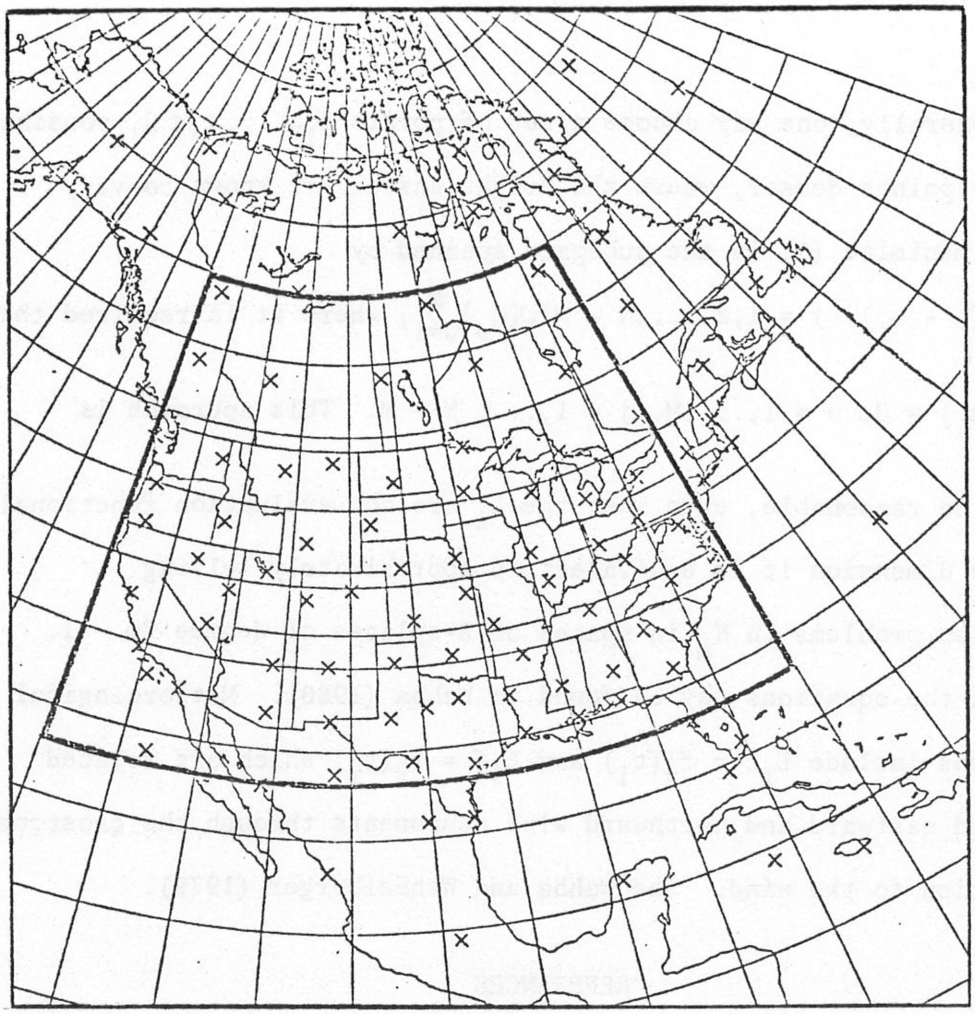

Figure 1. Location of 120 North American Weather Stations

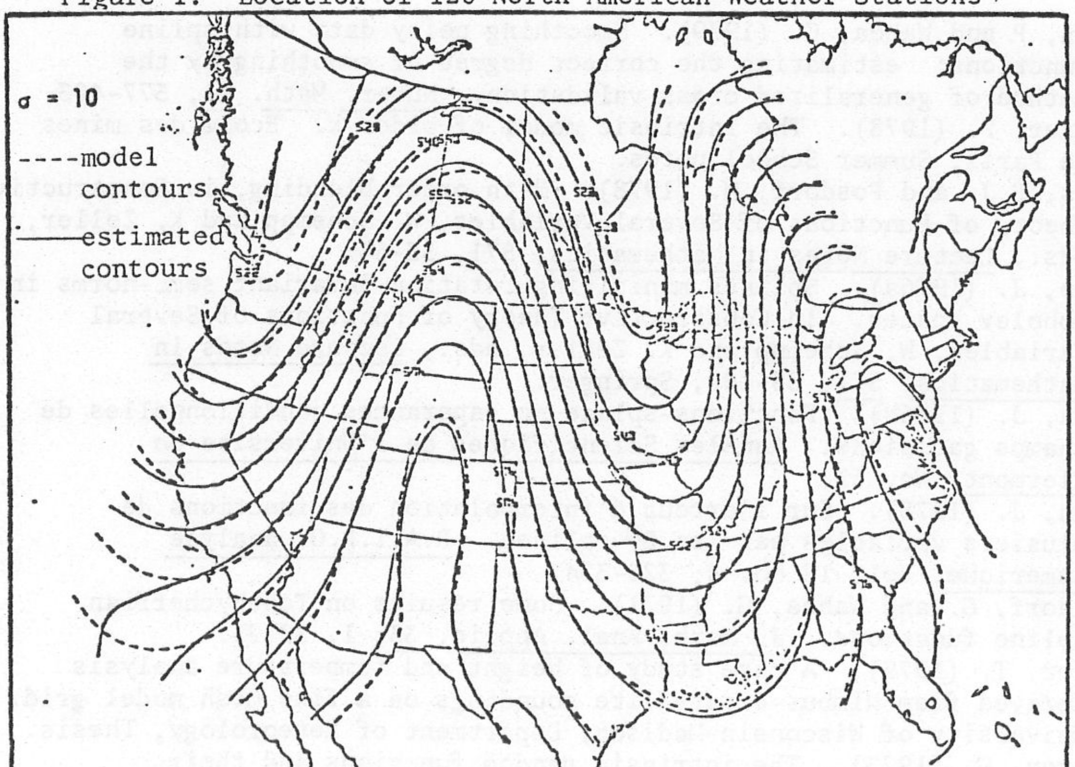

Figure 2. Model and Estimated 500 Millibar Contours

or more generally, one may choose a set of points $s_1,\ldots,s_N \varepsilon \Omega$, possibly making the points denser, where the "right answer" is known to vary most, and minimize (*) in the subspace spanned by $\{\sum_{i=1}^{N} \tilde{u}_{ji} E_m(t - s_i),\ j = 1,2,\ldots,N - M\} \cup \{\phi_\nu\}_{\nu=1}^{M}$, where it is required that $\sum_{i=1}^{n} \tilde{u}_{ji}\phi_\nu(s_i) = 0,\ \nu = 1,\ldots,M,\ j = 1,\ldots, N - M$. This approach is likely to be reasonable, even when the L_i are not evaluation functionals, and in one dimension it is equivalent to approximately solving optimization problems in H_m in spaces of B-splines of degree 2m - 1. Details of the equations may be found in Wahba (1980). Meteorological applications include $L_i f = f_y(t_i)$ and $L_i f = f_x(t_i)$ which are related to measured eastward and northward wind components through the geostrophic approximation to the wind. See Wahba and Wendelberger (1979).

REFERENCES

Craven, P. and Wahba, G. (1979). Smoothing noisy data with spline functions: estimating the correct degree of smoothing by the method of generalized cross validation. Numer. Math. 31, 377-403.

Delfiner, P. (1978). The intrinsic model of order k. Ecole des mines de Paris, Summer School notes.

Delvos, F.J. and Posdorf, H. (1978). N-th order blending, in Constructive Theory of Functions of Several Variables, W. Schempp and K. Zeller, eds., Lecture Notes in Mathematics, 571, 53-64.

Duchon, J. (1976a). Splines minimizing rotation-invariant semi-norms in Sobolev spaces. In Constructive Theory of Functions of Several Variables, W. Schempp and K. Zeller, eds., Lecture Notes in Mathematics, 571, 85-100, Springer.

Duchon, J. (1976b). Fonctions-spline et esperances conditionnelles de champs gaussiens. Annales Scientifiques de l'Universite de Clermont, No. 61.

Duchon, J. (1978). Sur l'erreur d'interpolation des fonctions de plusiers variables par les D^m-splines. R.A.I.R.O. analyse numerique, Vol. 12 No. 4, 325-334.

Kimeldorf, G. and Wahba, G. (1971). Some results on Tchebycheffian spline functions. J. Math. Anal. Applic. 33, 1, 82-95.

Koehler, T. (1979). A case study of height and temperature analysis derived from Nimbus-6 satellite soundings on a fine mesh model grid. University of Wisconsin-Madison, Department of Meteorology, Thesis.

Matheron, G. (1973). The intrinsic random functions and their applications. Adv. Appl. Prob. 5, 439-468.

Meinguet, J. (1978). An intrinsic approach to multivariate spline interpolation at arbitrary points. Proceedings of the NATO Advanced Study Institute on Polynomial and Spline Approximation, Calgary, B. Sahney, ed. 163-190. D. Reidell Publishing Co.

Micchelli, G. and Wahba, G. (1979). Design problems for optimal surface interpolation. University of Wisconsin-Madison, Department of Statistics Technical Report No. 565.

Montes, L. Paihua (1979). Quelques methodes numeriques pour le calcul de fonctions splines a une et plusiers variables. Thesis, University Scientifique et Medicale de Grenoble.

Utreras Dias, F. (1979). Cross validation techniques for smoothing spline functions in one or two dimensions. In Smoothing Techniques for Curve Estimation, T. Gasser and M. Rosenblatt, eds. Lecture Notes in Mathematics, No. 757, Springer-Verlag, Berlin.

Wahba, G. (1977). Practical approximate solutions to linear operator equations when the data are noisy, <u>SIAM J. Numerical Analysis, 14</u>, 4, 651-667.

Wahba, G. (1979a). How to smooth curves and surfaces with splines and cross-validation. In "Proceedings of the 24th Conference on the Design of Experiments" U.S. Army Research Office, Report 79-2, also University of Wisconsin-Madison Statistics Department Technical Report No. 555.

Wahba, G. (1979b). Convergence rates of "Thin Plate" smoothing techniques for curve estimation, T. Gasser and M. Rosenblatt, eds. Lecture notes in Mathematics, No. 757, Springer-Verlag, Berlin.

Wahba, G. (1979c). Interpolating surfaces: high order convergence rates and their associated designs, with application to X-ray image reconstruction. University of Wisconsin-Madison, Department of Statistics Technical Report No. 523.

Wahba, G. and Wendelberger, J. (1979). Some new mathematical methods for variational objective analysis using splines and cross validation. University of Wisconsin-Madison Statistics Department Technical Report No. 578.

Wahba, G. (1980). Ill-posed problems: Numerical and statistical methods for mildly, moderately, and severely ill posed problems with noisy data. University of Wisconsin-Madison Statistics Department Technical Report No. 595. To appear in Proceedings of the International Symposium on Ill-Posed Problems, M.Z. Nashed, ed.

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER Technical Report No. 597	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) SPLINE BASES, REGULARIZATION, AND GENERALIZED CROSS VALIDATION FOR SOLVING APPROXIMATION PROBLEMS WITH LARGE QUANTITIES OF NOISY DATA		5. TYPE OF REPORT & PERIOD COVERED Scientific Interim 6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) Grace Wahba		8. CONTRACT OR GRANT NUMBER(s) DAAG29-77-0207
9. PERFORMING ORGANIZATION NAME AND ADDRESS Department of Statistics University of Wisconsin Madison, WI 53706		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
11. CONTROLLING OFFICE NAME AND ADDRESS U.S. Army Research Office P.O. Box 12211 Research Triangle Park, N.C.		12. REPORT DATE February 1980 13. NUMBER OF PAGES 9
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)		15. SECURITY CLASS. (of this report) Unclassified 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

This document has been approved for public release and sale; its distribution is unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Thin plate spline bases, generalized cross validation, smooth estimation of functions of several variables

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

(See reverse side)

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE
S/N 0102-LF-014-6601

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

We consider the problem of estimating a smooth function $f(t)$, t in a bounded region in Euclidean d space, given observations $z_i = L_i f + e_i$, $i = 1,2,\ldots,n$, where the L_i are linear functionals and the e_i are random errors. A large number of practical problems can be placed in this context. We estimate f as the minimizer of $\frac{1}{n}\sum_{i=1}^{n}(L_i f - z_i)^2 + \lambda J_m(f)$, where, for $d = 2$, $J_m(f) = \iint \sum_{\nu=0}^{m} \binom{m}{\nu} \left(\frac{\partial^m f}{\partial x^\nu \partial y^{m-\nu}}\right)^2 dxdy$. The parameters λ and m are chosen from the data by the method of generalized cross validation. We cite a number of results concerning the theoretical properties of this estimate of f and present some numerical results concerning the smoothing of 500 millibar heights ($L_i f = f(x_i, y_i)$), with $n = 120$. The numerical method used is not very practical for n much larger than around 130. We propose an approximate method for computing the estimate of f which should be suitable for n several times as large.