DEPARTMENT OF STATISTICS
UNIVERSITY OF WISCONSIN
MADISON, WISCONSIN

MAY 7 1981

DEPARTMENT OF STATISTICS

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

TECHNICAL REPORT NO. 638

March 1981

NUMERICAL EXPERIMENTS WITH THE THIN PLATE HISTOSPLINE

Grace Wahba
University of Wisconsin

MAY 7 1981

This research was supported by the U.S. Army under contract no. DAAG-29-80-K0042.

NUMERICAL EXPERIMENTS WITH THE THIN PLATE HISTOSPLINE

Grace Wahba

Department of Statistics
University of Wisconsin-Madison
Madison, Wisconsin 53706, USA

Key words and Phrases: thin plate histosplines; smoothing of data aggregated by area; contour plots for standardized mortality ratios.

ABSTRACT

The thin plate volume matching and volume smoothing histosplines are described. These histosplines are suitable for estimating densities or incidence rates as a function of position on the plane when data is aggregated by area, for example by county. We give a numerical algorithm for the volume matching histospline and for the volume smoothing histospline using generalized cross validation to estimate the smoothing parameter. Some numerical experiments were performed using synthetic data, population data and SMR's (standardized mortality ratios) aggregated by county over the state of Wisconsin. The method turns out to be not particularly suited for obtaining population density maps where the population density can vary by two orders of magnitude, because the histospline can be negative in unpleasant ways. However the fitting of SMR's, which are all about the same order of magnitude, results in some esthetically pleasing pictures which may be used to search visually for geographic patterns. A number of open questions remain.

1. INTRODUCTION

Tobler (1979), in a recent J.A.S.A. paper, considered the problem of obtaining a smooth surface representing spatial density from observations on average densities over irregularly shaped geographical regions. He was interested in describing the population density in an area Ω given the average population density in each of n subareas $\{\Omega_i\}_{i=}^{n}$, $\underset{i}{U}\Omega_i = \Omega$. As an example, Ω is the union of the contiguous 48 states. The problem is to obtain a smooth, non-negative function f with the volume matching property:

$$\frac{1}{|\Omega_i|}\int_{\Omega_i} f(x,y)dxdy = g_i, \quad i = 1,2,\ldots,n,$$

where $|\Omega_i|$ is the area of Ω_i, and g_i is the observed average value of f over Ω_i. Tobler suggests, (among other things), seeking f to minimize

$$J_1^{\Omega}(f) = \int_{\Omega}(f_x^2+f_y^2)dxdy$$

subject to

$$\frac{1}{|\Omega_i|}\int_{\Omega_i} f(x,y)dxdy = g_i, \quad i = 1,2,\ldots,n,$$

and

$$f(x,y) \geq 0, \ (x,y)\in\Omega.$$

See Tobler (1979) and the comments by Dyn, Wahba and Wong (1979). Tobler proposed an iterative algorithm of finite difference type for calculating an approximate solution to this and some related problems, and presented two population density maps, the population of Ann Arbor given census tract data, and the population of the contiguous United States, given state data. Tobler's work is one of two generalizations of the one dimensional histospline given by Boneva, Kendall and Stevanov (BKS), (1971), and later analyzed by Wahba (1975), and Tobler's J_1^{Ω} above leads to the thin plate histospline, which we will be discussing further

in this paper. The thin plate histosplines are the solution of the problem of "logarithmic potential theory" mentioned by BKS on p. 33 of their paper. The other two dimensional generalization of the histospline, namely, the tensor product histospline, was studied by Schoenberg (1973), and Kuhn (1975), for the special case where the subareas result from a rectangular partition. We will not discuss the tensor product histospline here.

There are a number of variations on the solutions proposed by Tobler, of varying degrees of analytical and numerical complexity. To place them in perspective, we shall describe some of them here. We shall ultimately be concerned with a computationally relatively simple member of this group. Some of the variations generalize to the number of dimensions d greater than 2. (See Dyn and Wahba (1979), Wahba and Wendelberger (1980), but we shall only consider d = 2 here.

Let Ω be a closed bounded region of the plane with a sufficiently "nice" boundary $\partial\Omega$ and let $H_m(\Omega)$ be the Sobolev space (see Adams (1975)) of square integrable functions on Ω with

$$J_m^{\Omega}(f) = \sum_{\nu=0}^{m} \binom{m}{\nu} \int_{\Omega} \left(\frac{\partial^m}{\partial x^{\nu} \partial y^{m-\nu}} f(x,y) \right)^2 dxdy < \infty. \qquad m = 1,2,\ldots$$

$J_m^{\Omega}(\cdot)$ is a semi-norm on $H_m(\Omega)$ with null space the $M = \binom{m+1}{2}$ dimensional space spanned by the polynomials $(1,x,y,\ldots)$ of total degree $\leq m-1$. Let $f_{n,m}^{\Omega}$ be the solution to the volume matching problem: Find $f \in H_m(\Omega)$ to minimize $J_m^{\Omega}(f)$ subject to

$$\frac{1}{|\Omega_i|} \int_{\Omega_i} f(x,y)dxdy = g_i, \quad i = 1,2,\ldots,n.$$

The solution $f_{n,m}$ has been shown to exist uniquely provided the $n \times M$ matrix T with $j\nu$th entry $\tau_{j\nu}$ defined by

$$\tau_{j\nu} = \int_{\Omega_j} \phi_{\nu}(x,y)dxdy, \quad j = 1,2,\ldots,n, \ \nu = 1,2,\ldots,M \qquad (1.1)$$

where $\{\phi_\nu\}_{\nu=1}^{M}$ span the null space of J_m^{Ω}, is of rank M. $f_{n,m}^{\Omega}$ has been characterized as the solution to the boundary value problem

$$\Delta^m f = \sum_{i=1}^{n} a_i I_{\Omega_i},$$

subject to Neumann boundary conditions on $\partial\Omega$. Here Δ is the Laplacian operator ($\Delta f = f_{xx}+f_{yy}$), $I_{\Omega_i}(x,y) = 1$, $(x,y)\in\Omega_i$, $= 0$, $(x,y)\notin\Omega_i$, and the $\{a_i\}$ are constants determined by the data. See Dyn and Wahba (1979), Dyn, Wahba and Wong (1979), Dyn and Wong (1981), Wong (1980).

The results readily generalize to the volume smoothing problem: Find $f\in H_m(\Omega)$ to minimize

$$\frac{1}{n}\sum_{i=1}^{n}\left(z_i - \frac{1}{|\Omega_i|}\int_{\Omega_i} f(x,y)dxdy\right)^2 + \lambda J_m^{\Omega}(f), \qquad (1.2)$$

where here the z_i are supposed to be imperfectly measured estimates of the average of f over Ω_i, and λ is a positive smoothing parameter. See Dyn and Wahba (1980). For $m \geq 2$, $H_m(\Omega)$ is a reproducing kernel space (that is, the evaluation functionals are bounded, see Adams (1975)), and so the set of non-negative functions in $H_m(\Omega)$ is closed and convex. Existence and uniqueness of a solution to the volume matching problem in $H_m(\Omega)$, $m \geq 2$, subject to the positivity constraints

$$f(x,y) \geq 0 \qquad ,(x,y)\in\Omega, \qquad (1.3)$$

then follows because $J_m^{\Omega}(f)$ is <u>strongly</u> convex over the set of functions satisfying the volume matching constraints (provided T is of rank M), see Laurent (1980), Wong (1980). Similarly the minimizer of (1.2) subject to the non-negativity constraints (1.3) exists uniquely because (1.2) is a strongly convex functional on $H_m(\Omega)$ if T is of rank M. The problem of establishing the existence and uniqueness of a solution to the volume matching or volume smoothing problem with non-negativity constraints in

in $H_1(\Omega)$ is more difficult, since $H_1(\Omega)$ is not a reproducing kernel space and the set of non-negative functions has to be defined in the distributional sense. However Dyn and Wong (1981), and Wong (1980) have established existence, uniqueness, and a characterization of the non-negatively constrained volume matching and volume smoothing histospline in $H_1(\Omega)$, and Wong has proposed a class of algorithms for computing it.

The difficulties of solving a boundary value problem numerically can be circumvented if one replaces $J_m^{\Omega}(f)$ by $J_m(f)$ defined by

$$J_m(f) = \sum_{\nu=0}^{m} \binom{m}{\nu} \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \left(\frac{\partial^m}{\partial x^{\nu} \partial y^{m-\nu}} f(x,y)\right)^2 dxdy \qquad (1.4)$$

in the volume matching and volume smoothing problem, and replaces $H_m(\Omega)$ by $H_m(R^2)$, where $H_m(R^2)$ is the (Beppo Levi) space of (Schwartz) distributions whose partial derivatives (in the distributional sense) of total order m are square integrable on R^2, see Meinguet (1978). For $m > 1$ the elements of $H_m(R^2)$ are functions in the ordinary sense.

In this paper we give an algorithm for and report on some numerical experience with the solution to the volume matching problem: Find $f \in H_m(R^2)$ to minimize $J_m(f)$ of (1.4) subject to

$$\frac{1}{|\Omega_i|} \int_{\Omega_i} f(x,y)dxdy = g_i, \quad i = 1,2,\ldots,n \qquad (1.5)$$

and the volume smoothing problem: Find $f \in H_m(R^2)$ to minimize

$$\frac{1}{n} \sum_{i=1}^{n} \left(z_i - \frac{1}{|\Omega_i|} \int_{\Omega_i} f(x,y)dxdy\right)^2 + \lambda J_m(f). \qquad (1.6)$$

We consider $m = 2,3,4$, but most of the experimental work is with $m = 2$. In this paper we do not impose positivity constraints.

With J_m^{Ω} replaced by J_m, an explicit representation of $f_{n,m}$ and $f_{n,m,\lambda}$, the solutions to the volume matching and volume smoothing problems in $H_m(R^2)$ can be given and are found in

Wahba and Wendelberger (1980), based on earlier work by Duchon (1977) and Meinguet (1978). We give an algorithm here for computing $f_{n,m}$ and $f_{n,m,\lambda}$. The algorithm is sufficiently simple so that in the volume smoothing problem λ may be chosen by the method of generalized cross validation (GCV), see Craven and Wahba (1979), Wahba and Wendelberger (1980). The basic design of the algorithm follows the approach of Wendelberger (1981) who incorporated some suggestions of G. Golub. An earlier algorithm is due to Paihua Montes (1978). The program was coded by A. Kirsch, who benefited from some advice of D. Bates.

After describing the numerical method, we then try the method out on several sets of synthetic data as well as population data and four sets of revised standardized mortality ratios (SMR's) for different types of cancer, reported by county in the State of Wisconsin. The method turns out to be unsatisfactory for the population data because population density in Wisconsin by county varies by over two orders of magnitude, and the estimation goes negative here in an unpleasant way. The SMR data, however varies much less, and the estimates are non-negative or only negligibly negative. It is hoped that the resultant density maps can provide a useful visual picture of relative SMR's which may be used to screen for possible geographic patterns in the SMR's. Questions of significance of observed "bumps", "ridges" and other patterns remains completely unresolved here, and leads to a number of interesting open questions, some of which are discussed at the end.

2. THE ALGORITHM FOR THE VOLUME MATCHING PROBLEM IN $H_m(R^2)$.

The solution $f_{n,m,\lambda}$ to the volume smoothing problem of (1.6) is given by

$$f_{n,m,\lambda}(x,y) = \sum_{i=1}^{n} c_i \int_{\Omega_i} E_m(|t-s|)ds + \sum_{\ell=1}^{m} d_\nu \phi_\nu(t) \qquad (2.1)$$

where $t = (x,y)$,

$$E_m(|t|) = |t|^{2m-2}\log|t|, \quad |t| = \sqrt{x^2+y^2}$$

and the $\{\phi_\nu\}_{\nu=1}^{M}$ are the M polynomials which span the null space of J_m, for example, if m = 3, then M = 6 and

$$\phi_1(t) = 1 \quad \phi_2(t) = x \quad \phi_3(t) = y$$

$$\phi_4(t) = x^2 \quad \phi_5(t) = xy \quad \phi_6(t) = y^2$$

and $c = (c_1,\ldots,c_n)'$ and $d = (d_1,\ldots,d_M)'$ are the (unique) solutions to the equations

$$\begin{aligned} (K+n\lambda I)c + Td &= g, \\ T'c &= 0 \end{aligned} \tag{2.2}$$

where $g = (g_1,\ldots,g_n)'$, and K is the n×n matrix with jkth entry

$$K_{jk} = \int_{\Omega_i}\int_{\Omega_k} E_m(|t-s|)dtds$$

and T is as in (1.1). This result is a special case of Wahba and Wendelberger (1980), see also Dyn and Wahba (1979).

The solution to the volume-matching problem of (1.5) is obtained by setting $\lambda = 0$ in (2.2).

We remark that since $E_m(|\cdot|)$ is a fundamental solution of the m-iterated Laplacian, we also have here

$$\Delta^m f_{n,m,\lambda} = \sum_{i=1}^{n} a_i I_{\Omega_i},$$

for some $\{a_i\}$.

The GCV estimate of λ is the minimizer of

$$V(\lambda) = \frac{\frac{1}{n}||(I-A(\lambda))z||^2}{(\frac{1}{n}Tr(I-A(\lambda)))^2} \tag{2.3}$$

where $A(\lambda)$ is the n×n influence matrix satisfying

$$\begin{pmatrix} \int_{\Omega_1} f_{n,m,\lambda}(t)dt \\ \vdots \\ \int_{\Omega_n} f_{n,m,\lambda}(t)dt \end{pmatrix} = A(\lambda)z.$$

See Craven and Wahba (1979), Wahba (1977).

The matrix $I-A(\lambda)$ has the representation

$$I - A(\lambda) = n\lambda C(C'KC+n\lambda I)^{-1}C'$$

where $C_{n\times(n-M)}$ is any matrix with n-M orthogonal columns which are orthogonal to the M columns of T, and

$$c = C(C'KC+n\lambda I)^{-1}C'z$$

See Wahba (1979b). It can be shown (see Duchon (1975)) that the n-M×n-M matrix C'KC is positive definite even though K is not. Analytical formulas for $\xi_i(t) = \int_{\Omega_i} E_m(|t-s|)ds$, $\tau_{i\nu}$ and K_{jk} will not generally be available. To preserve the positive definiteness numerically, we discretize the problem at the data functional stage rather than merely applying quadrature formulae to obtain $\xi_i(t)$, T and K. We believe that this is the correct point to discretize, if discretization is necessary. (See, for example, Chambless (1980)). A description of this discretization follows.

Let $\{t_k\}_{k=1}^{N}$ be a fine rectangular grid of points in Ω. Figure 2.1 shows the arrangement of the $\{t_k\}$ (dots) and the Ω_i (squares), for the first test case.

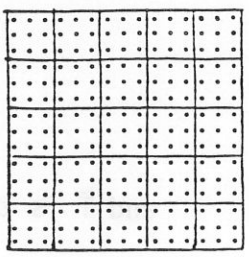

Figure 2.1

Next we approximate the linear functional L_i defined by

$$L_i f \equiv \int_{\Omega_i} f(x,y)dxdy \tag{2.4}$$

by the linear functional $\tilde{L}_i$ defined by

$$\tilde{L}_i f = \frac{|\Omega_i|}{N_i} \sum_{t_k \varepsilon \Omega_i} f(t_k) \quad , \; i = 1,2,\ldots,n \; . \tag{2.5}$$

The solution to the problem: Minimize

$$\frac{1}{n}\sum_{i=1}^{n} (\tilde{L}_i f - z_i)^2 + \lambda J_m(f) \tag{2.6}$$

is

$$\tilde{f}_{n,m,\lambda}(t) = \sum_{i=1}^{n} \tilde{c}_i \tilde{\xi}_i(t) + \sum_{\nu=1}^{M} \tilde{d}_\nu \phi_\nu(t) \tag{2.7}$$

where

$$\tilde{\xi}_i(t) = \frac{|\Omega_i|}{N_i} \sum_{t_k \varepsilon \Omega_k} E_m(|t-t_k|) \tag{2.8}$$

and $\tilde{c} = (\tilde{c}_1,\ldots,\tilde{c}_n)'$ and $\tilde{d} = (\tilde{d}_1,\ldots,\tilde{d}_M)'$ satisfy

$$(\tilde{K}+n\lambda I)\tilde{c} + \tilde{T}\tilde{d} = z \tag{2.9a}$$

$$\tilde{T}'\tilde{c} = 0 \tag{2.9b}$$

where $\tilde{K}$ is the $n\times n$ matrix with jkth entry $\tilde{K}_{jk}$:

$$\tilde{K}_{jk} = \frac{|\Omega_j|}{N_j}\frac{|\Omega_k|}{N_k} \sum_{\substack{t_\ell \varepsilon \Omega_j \\ t_m \varepsilon \Omega_k}} E_m(|t_\ell - t_m|) \tag{2.10}$$

and $\tilde{T}$ is the $n\times M$ matrix with $\tau_{j\nu}$th entry

$$\frac{|\Omega_j|}{N_j} \sum_{t_\ell \varepsilon \Omega_j} \phi_\nu(t_\ell). \tag{2.11}$$

Now $\tilde{K}$ and $\tilde{T}$ can be evaluated to a high degree of accuracy, and it can be shown that, if $\tilde{C}$ is any $n \times n-M$ matrix with $n-M$ orthogonal columns all orthogonal to the columns of T, then $\tilde{C}'\tilde{K}\tilde{C}$ will be positive definite. The cross validation function $\tilde{V}(\lambda)$ appropriate to this problem is now

$$\tilde{V}(\lambda) = \frac{\frac{1}{n}||(I-\tilde{A}(\lambda))z||^2}{(\frac{1}{n}Tr(I-\tilde{A}(\lambda)))^2} \tag{2.12}$$

where

$$I-\tilde{A}(\lambda) = n\lambda\tilde{C}(\tilde{C}'\tilde{K}\tilde{C}+n\lambda I)^{-1}\tilde{C}'. \tag{2.13}$$

We now show how $\tilde{V}(\lambda)$, $\tilde{c}$ and $\tilde{d}$ may be computed. In what follows we will drop all the "~", it being understood that the computations use (2.7) - (2.13).

1. Use the Q-R decomposition in LINPACK (Dongarra et al. (1979)), to obtain

$$T_{n\times M} = Q_{n\times n}R_{n\times M}$$

where Q is orthogonal and R is zero except in the top $n\times M$ block.

$$Q = \left(\underbrace{Q_1}_{M} \;\vdots\; \underbrace{Q_2}_{n-M}\right)\Big\}n \qquad R = \left(\begin{array}{c} R_1 \\ \hdashline 0 \end{array}\right)\begin{array}{l}\}M \\ \}n-M\end{array}$$

$$\underbrace{}_{M}$$

Remark: The columns of Q_2 are orthogonal to the columns of T and C is taken as Q_2'.

2. Let $B = Q'KQ$,

$$B = \left(\begin{array}{c|c} B_{11} & B_{12} \\ \hline B_{21} & B_{22} \end{array}\right) \begin{matrix} \}M \\ \}n-M \end{matrix}$$

and find the eigenvector-eigenvalue decomposition of B_{22} using EISPACK (Smith et al. (1976))

$$B_{22} = \Gamma \begin{pmatrix} b_1 & & 0 \\ & \ddots & \\ 0 & & b_{n-M} \end{pmatrix} \Gamma'.$$

Remark: B_{22} is positive definite.

3. Let

$$\omega = \Gamma' Q_2 z. \tag{2.14}$$

Then $V(\lambda)$, c and d are obtained from:

$$V(\lambda) = n \sum_{i=1}^{n-M} \frac{\omega_i^2}{(b_i+n\lambda)^2} \Bigg/ \left(\sum_{i=1}^{n-M} \frac{1}{(b_i+n\lambda)} \right)^2 \tag{2.15}$$

λ is chosen as the minimizer of $V(\lambda)$.

$$c = Q_2 \Gamma \begin{pmatrix} b_1+n\lambda & & 0 \\ & \ddots & \\ 0 & & b_{n-M}+n\lambda \end{pmatrix}^{-1} \omega$$

$$R_1 d = Q_1' z - B_{12} \Gamma \begin{pmatrix} b_1+n\lambda & & 0 \\ & \ddots & \\ 0 & & b_{n-M}+n\lambda \end{pmatrix}^{-1} \omega \quad .$$

To avoid over and under flow, distance units should be taken so that Ω sits snugly inside the unit square centered at (0,0), and $E_m(|t_i-t_j|)$ should be set equal to 0 for i = j (rather than attempting to compute it!)

It is noted that the expensive part of this program depends on m and the Ω_i, but not the data. In the cancer data examples below, where SMR's for various types of cancer are obtained for each of 72 counties (n=72) in Wisconsin (=Ω), $\tilde{K}$, $\tilde{T}$, $\Gamma' Q_2$, $\{b_i\}$, R_1, Q_1' and $B_{12}\Gamma$ are computed only once and stored. Once these are given, evaluation of $V(\lambda)$, c and d for different data is cheap.

3. NUMERICAL RESULTS

To see what the thin plate histospline will do on synthetic as well as real data several types of solutions were computed.

Given $\Omega = \bigcup_{i=1}^{n} \Omega_i$ one may define the ith "deltaspline" δ_i as the solution to the problem: Find $f \in H_m(R^2)$ to minimize $J_m(f)$ subject to

$$\frac{1}{|\Omega_k|} \int_{\Omega_k} f(x,y)dxdy = 1, \quad k = i$$
$$= 0, \quad k \neq i$$

Then the solution $f_{n,m}$ to the problem: find $f \in H_m(R^2)$ to minimize $J_m(f)$ subject to

$$\frac{1}{|\Omega_i|} \int_{\Omega_i} f(x,y)dxdy = g_i, \qquad = 1,2,\ldots,n$$

is given by

$$f_{n,m}(x,y) = \sum_{i=1}^{n} \delta_i(x,y) g_i .$$

Two sets of Ω, $\{\Omega_i\}_{i=1}^{n}$ $\{t_k\}_{k=1}^{N}$ were considered. For the first set, Ω is the unit square, n = 25, $\{\Omega_i\}$ is a 5×5 square partition, and the $\{t_k\}$ are a 15×15 square array as in Figure 2.1. For the second set Ω is the State of Wisconsin, $\{\Omega_i\}$ are the 72 counties and $\{t_k\}$ were obtained by laying out a rectangular 85×90 grid of points equally spaced in x and y just covering Ω and taking that subset falling inside Ω.

Figure 3.1 plots the "data" for the deltaspline corresponding to the central square in the 5×5 partition. It is a function whose value is 1 on the central square and 0 elsewhere. Figures 3.2, 3.3, and 3.4 give the deltaspline for this "data" for m = 2, 3 and 4. It can be seen that as m increases the minimum value taken on by the deltaspline becomes increasingly large negative. Large m is probably not appropriate when n is small, as it is here. The remaining experiments use m = 2.

Figure 3.5 is a map of the State of Wisconsin with superimposed contours for the surface formed by the sum of two deltasplines, for Oneida and Columbia counties. (Oneida contains the northern peak and Columbia the southern.) The height of the Oneida and Columbia input values were .00090 and .00129 respectively. It can be seen that the heights of the two peaks of the two-deltaspline surface are slightly above .0012 and .0016, respectively. South of Oneida county there is a negative valley of a depth about one eighth of the peak height over Oneida.

Figure 3.6 gives a map of the State of Wisconsin with the 1970 population density g_i, in people/square mile, indicated for each county. (The data has been rounded to the nearest integer for the plot but all available figures were used in the program.) Figure 3.7 gives a contour plot of the volume matching histospline which minimizes $J_2(f)$ subject to

$$\frac{1}{N_i} \sum_{t_k \in \Omega_i} f(t_k) = g_i, \qquad i = 1,\ldots,72. \tag{3.1}$$

where the left side of (3.1) is the (quadrature) approximation to

$$\frac{1}{|\Omega_i|} \int_{\Omega_i} f(x,y)dxdy.$$

It can be seen that this contour map has a sharp peak at the edge of Milwaukee county greater than 6000 people/square mile and a negative valley of maximum depth around 500 people/square mile

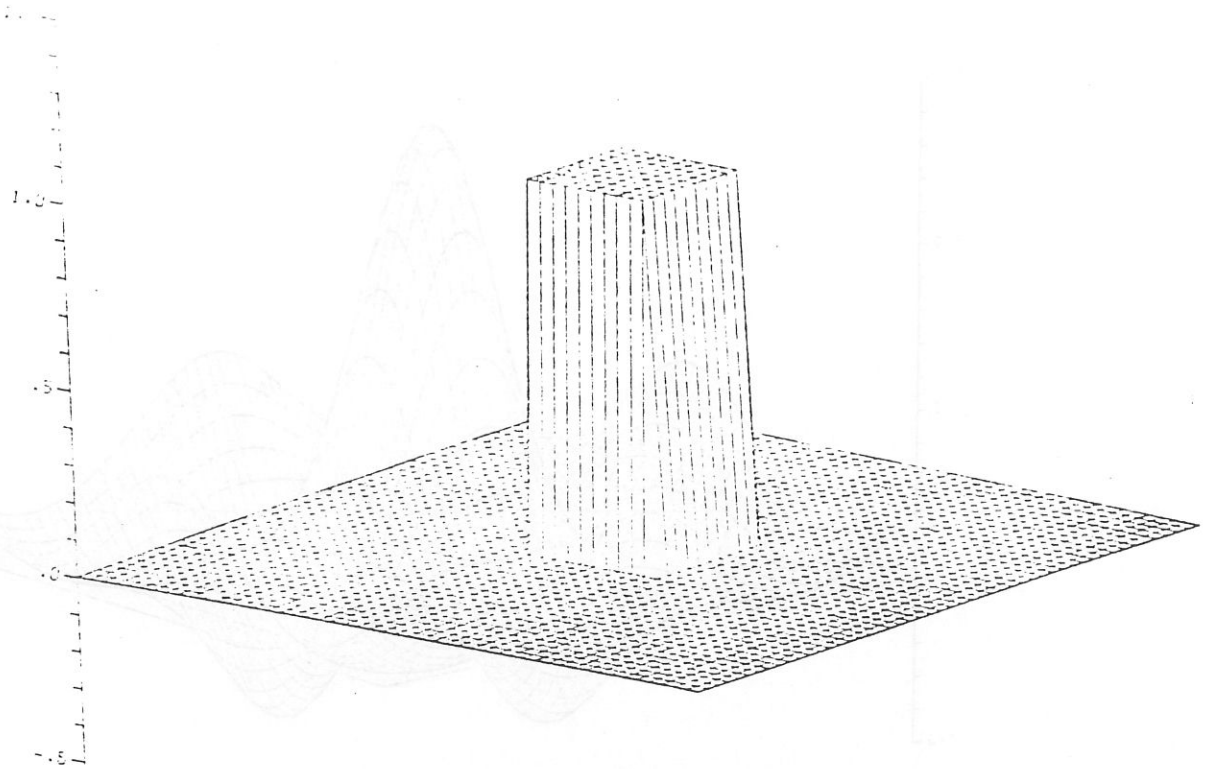

Figure 3.1. The "Data" for the Deltspline

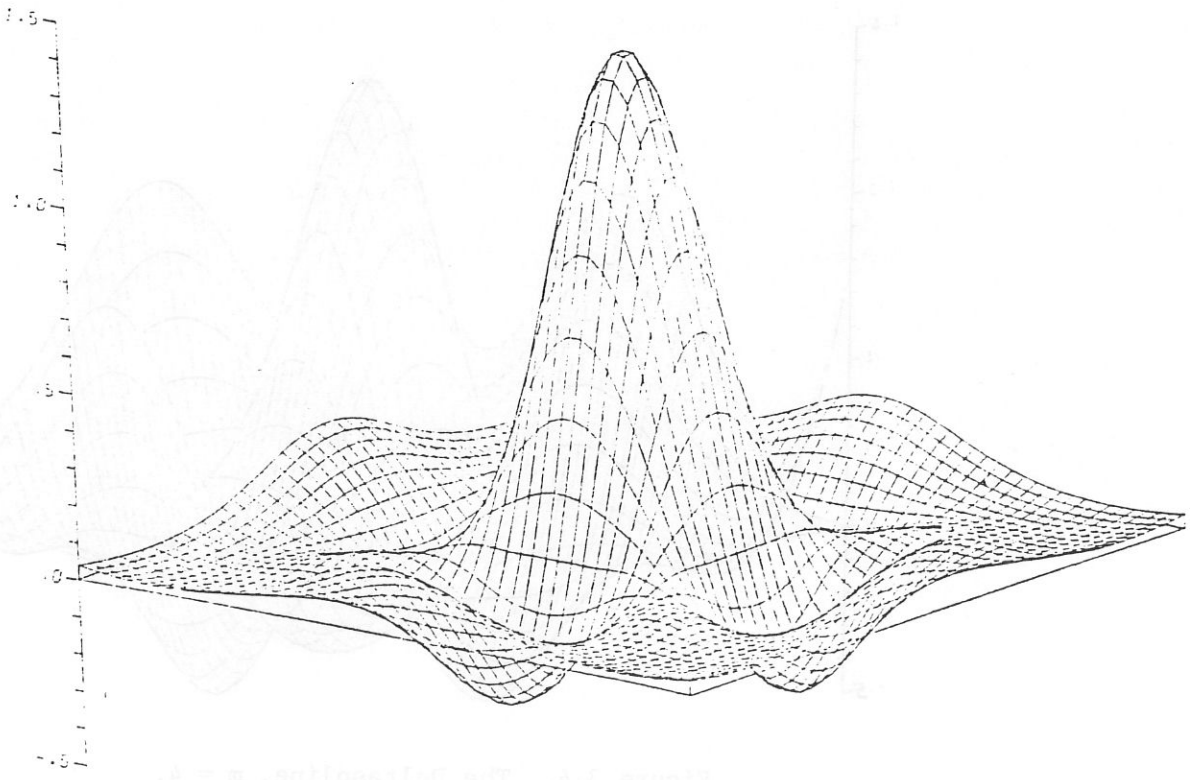

Figure 3.2. The Deltaspline, m = 2.

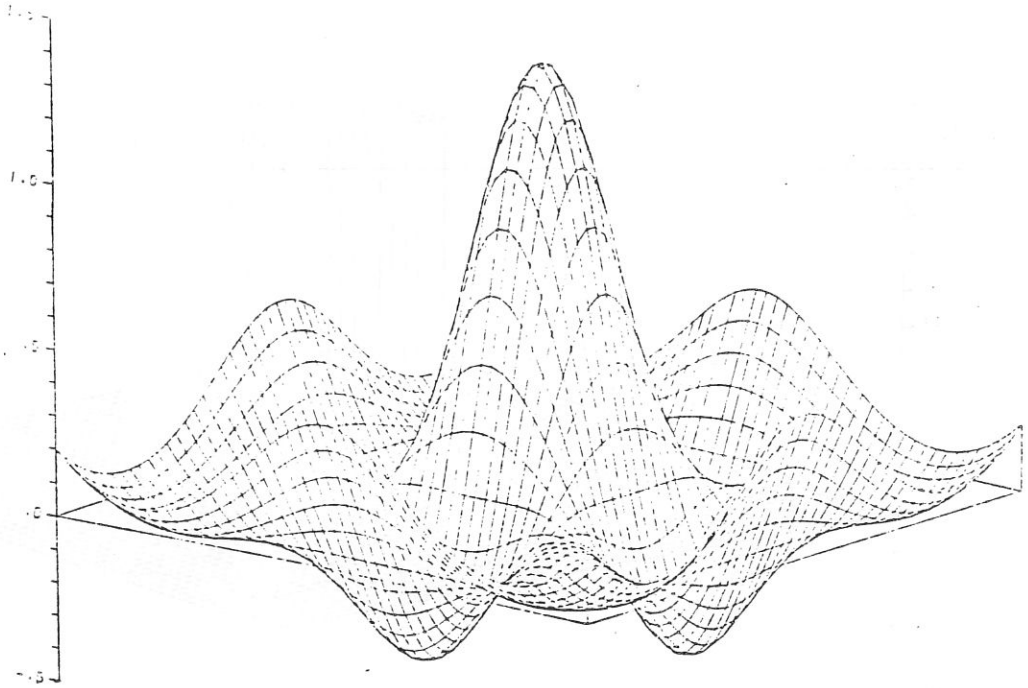

Figure 3.3. The Deltaspline, m = 3.

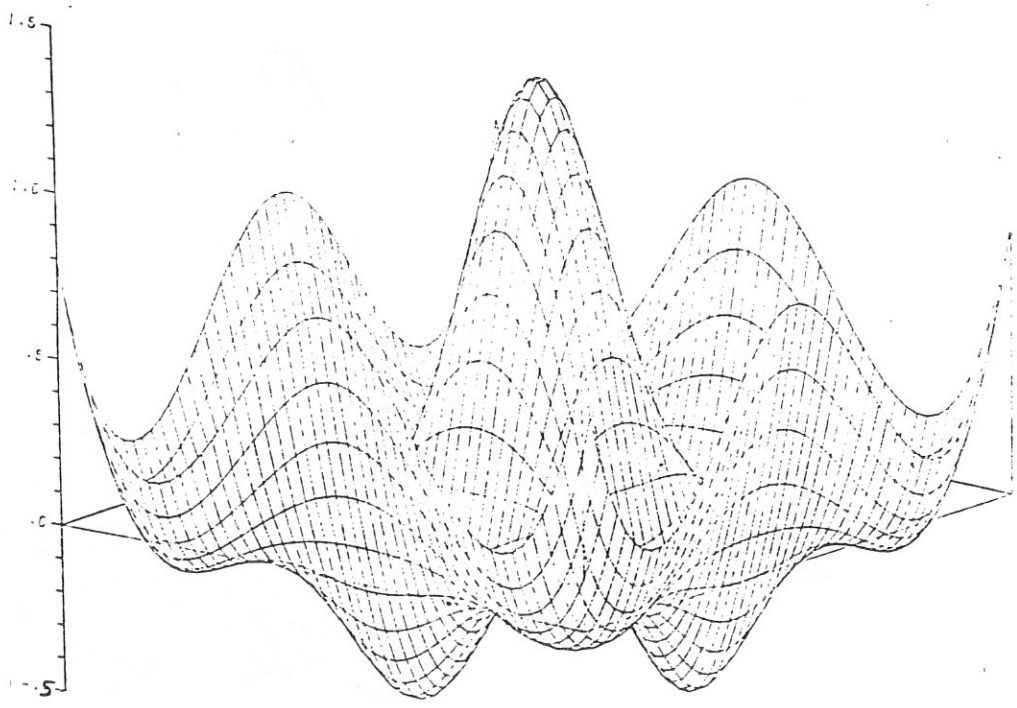

Figure 3.4. The Deltaspline, m = 4.

Figure 3.5. Two Delta Splines, For Oneida and Columbia Counties
Contour Interval, .0002.

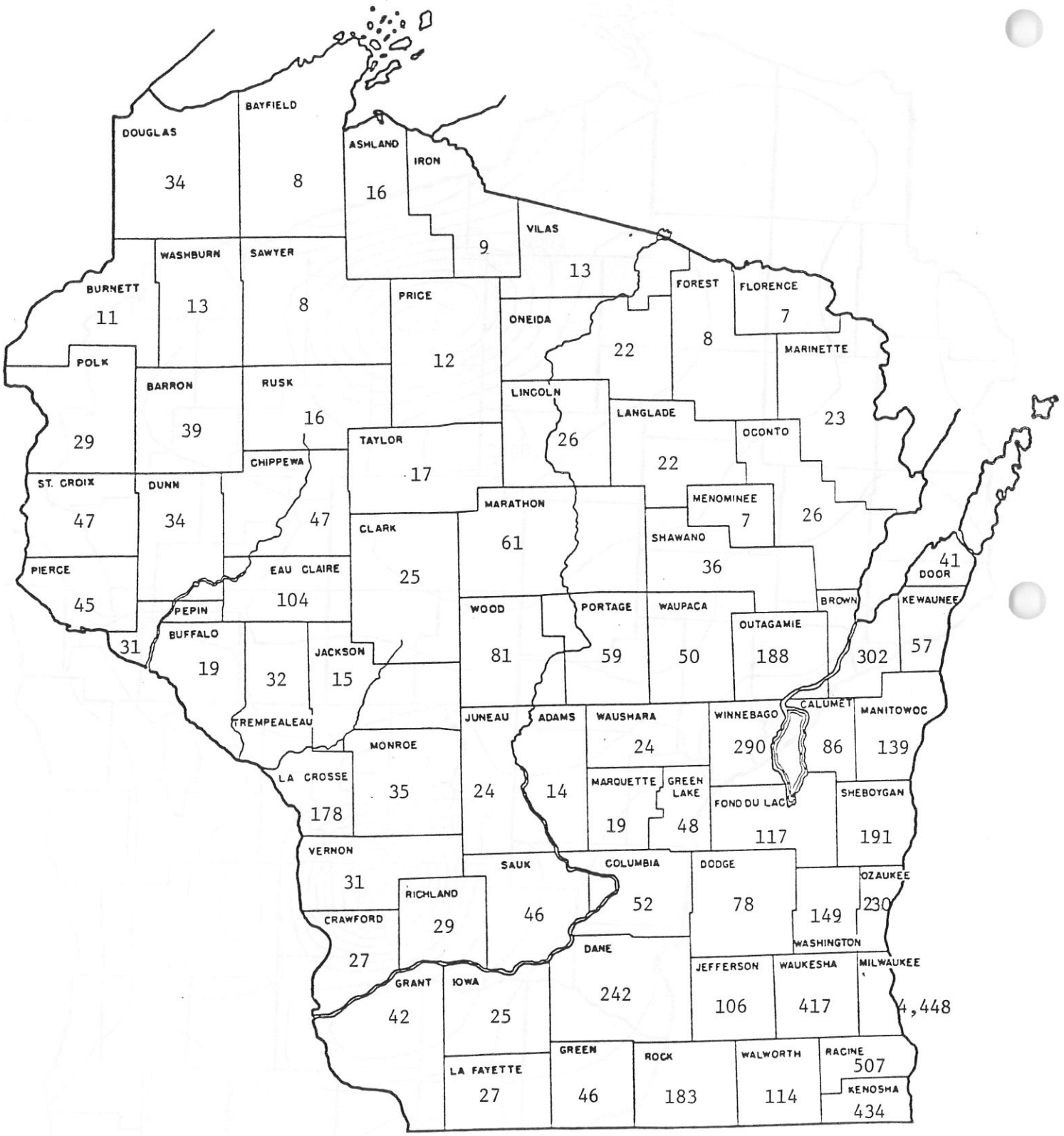

Figure 3.6. 1970 Wisconsin Population Density by County, People/Sq. Mi.

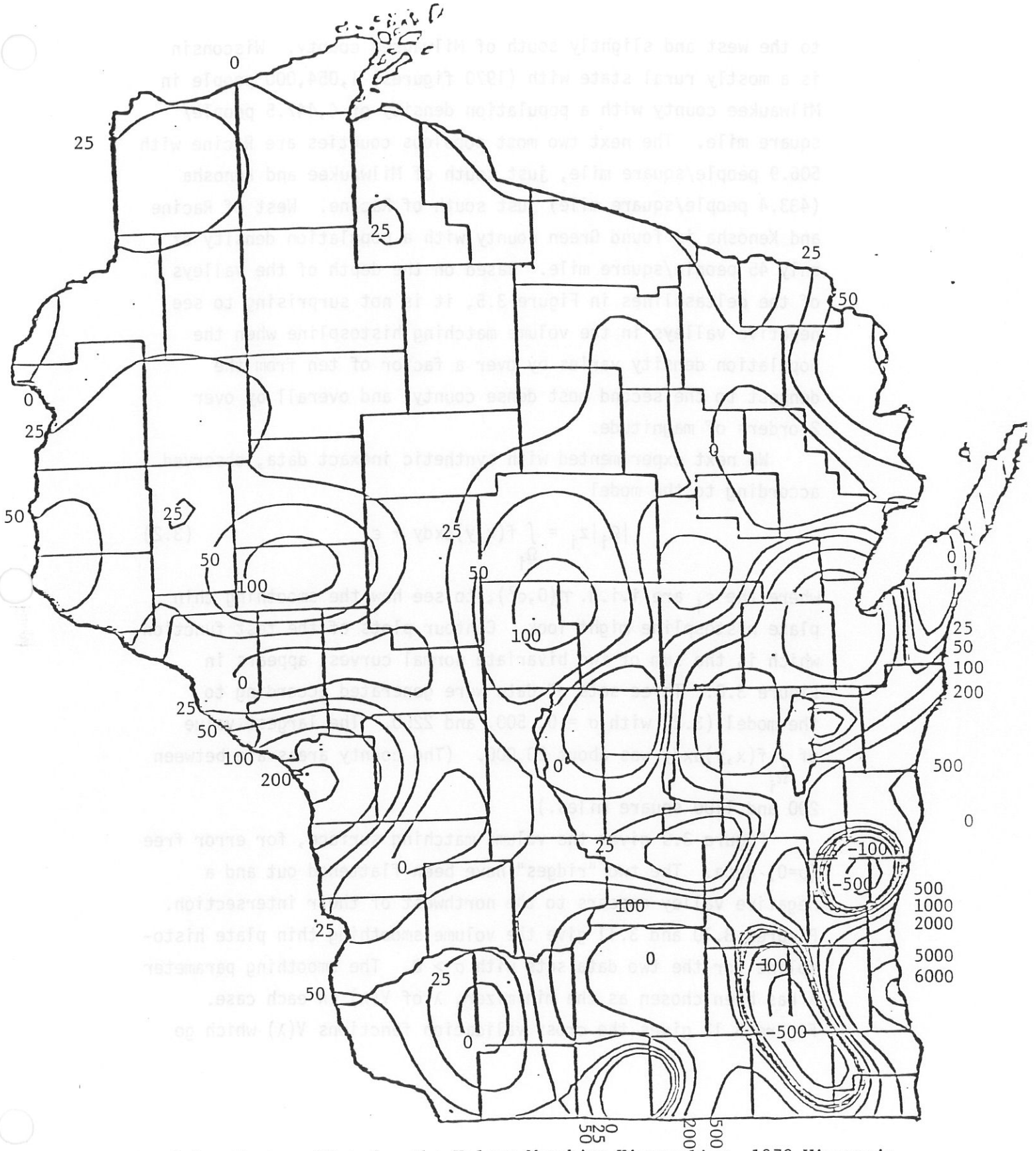

Figure 3.7. Contour Plot for the Volume Matching-Histospline, 1970 Wisconsin Population Density, People/Sq. Mi. Contour Levels: 6,000, 5000, 2000, 1000, 500, 200, 100, 50, 25, 0, -100, -500.

to the west and slightly south of Milwaukee county. Wisconsin is a mostly rural state with (1970 figures) 1,054,000 people in Milwaukee county with a population density of 4,447.5 people/square mile. The next two most populous counties are Racine with 506.9 people/square mile, just south of Milwaukee and Kenosha (433.4 people/square mile) just south of Racine. West of Racine and Kenosha is found Green county with a population density of only 45 people/square mile. Based on the depth of the valleys of the deltasplines in Figure 3.5, it is not surprising to see negative valleys in the volume matching histospline when the population density varies by over a factor of ten from the densest to the second most dense county, and overall by over 2 orders of magnitude.

We next experimented with synthetic inexact data, observed according to the model

$$|\Omega_i| z_i = \int_{\Omega_i} f(x,y)dxdy + \varepsilon_i \qquad (3.2)$$

where the ε_i are i.i.d. $\eta(0,\sigma^2)$, to see how the smoothing thin plate histospline might look. Contour plots of the test function, which is the sum of two bivariate normal curves, appears in Figure 3.8. Three sets of data were generated according to the model (3.2) with σ = 0, 500, and 2250. The largest value of $\int_{\Omega_i} f(x,y)dxdy$ was about 20,000. (The county areas are between 200 and 1100 square miles.)

Figure 3.9 gives the volume matching surface, for error free (σ=0) data. The two "ridges" have been flattened out and a negative valley appears to the northwest of their intersection. Figures 3.10 and 3.11 give the volume smoothing thin plate histospline for the two data sets with $\sigma > 0$. The smoothing parameter λ has been chosen as the minimizer, $\hat{\lambda}$ of $V(\lambda)$ in each case. Figure 3.12 gives the cross validation functions $V(\lambda)$ which go

Figure 3.8. Synthetic Surface for Test of the Smoothing Thin Plate Histospline. Contour Interval: 2.5.

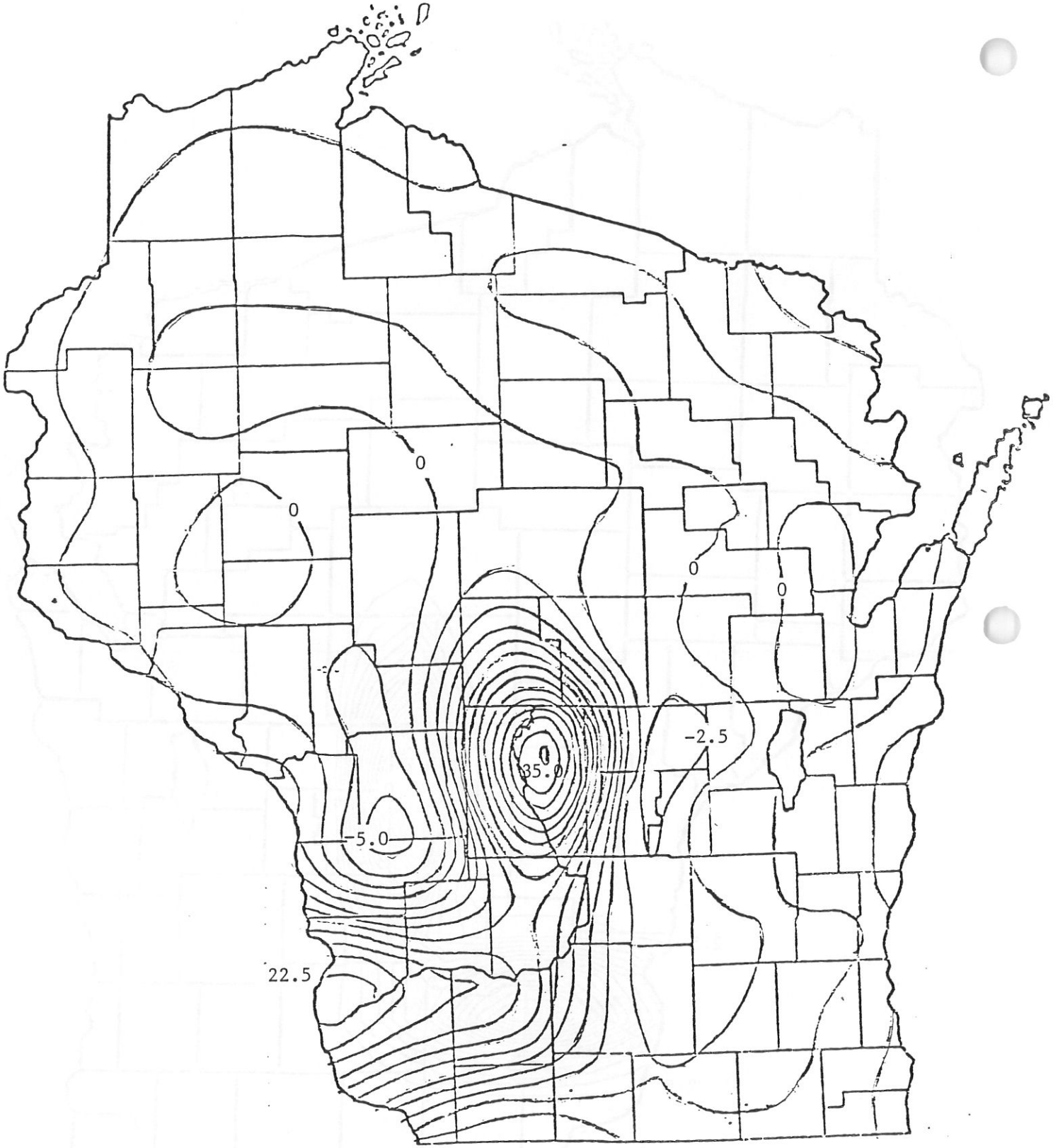

Figure 3.9. Volume Matching Surface for Exact Data, Contour Interval: 2.5.

Figure 3.10. Volume Smoothing Surface, S.D. of Data = 500.
Contour Interval: 2.5.

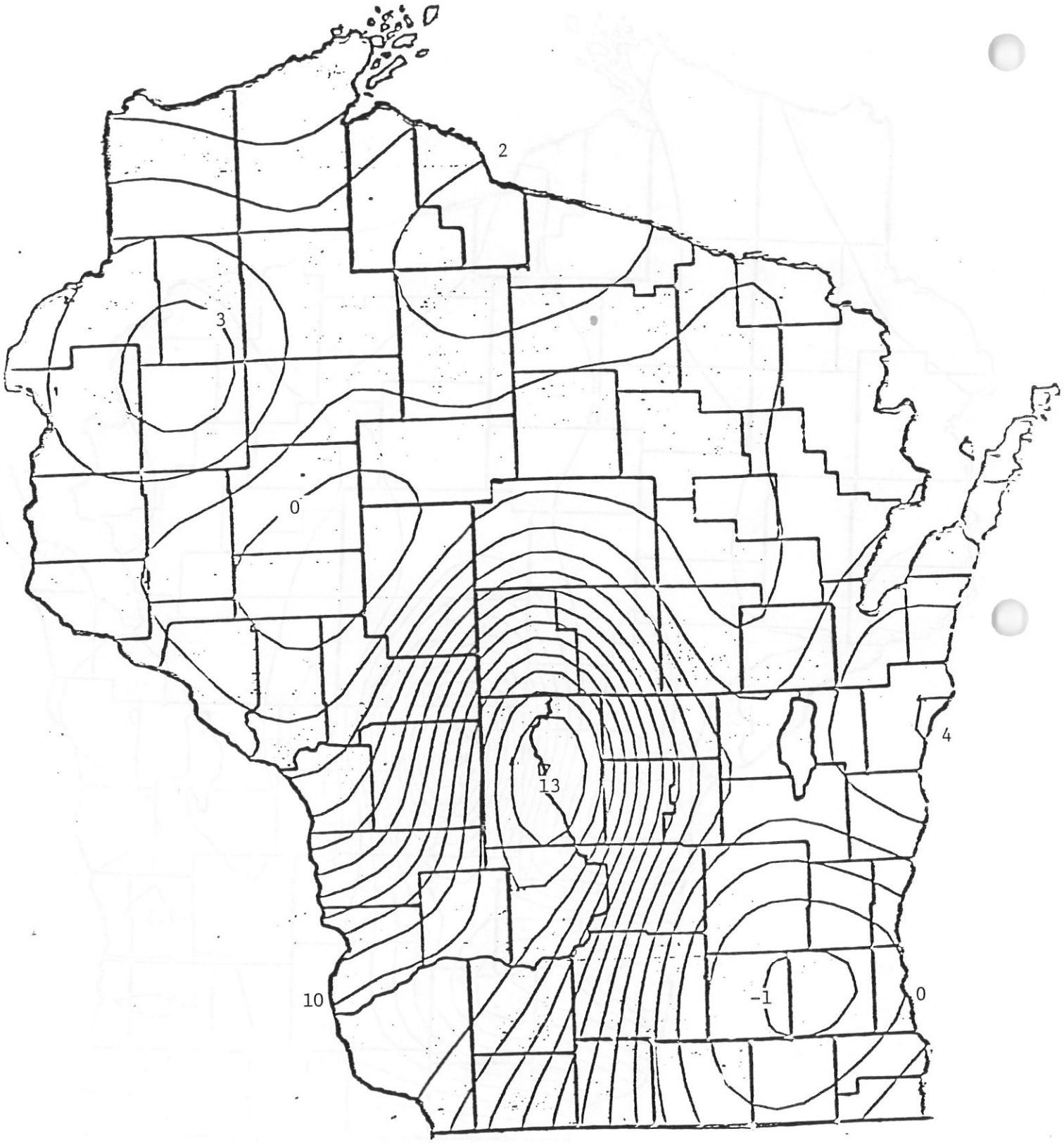

Figure 3.11. Volume Smoothing Surface, S.D. of Data = 2250.
Contour Interval: 1.0.

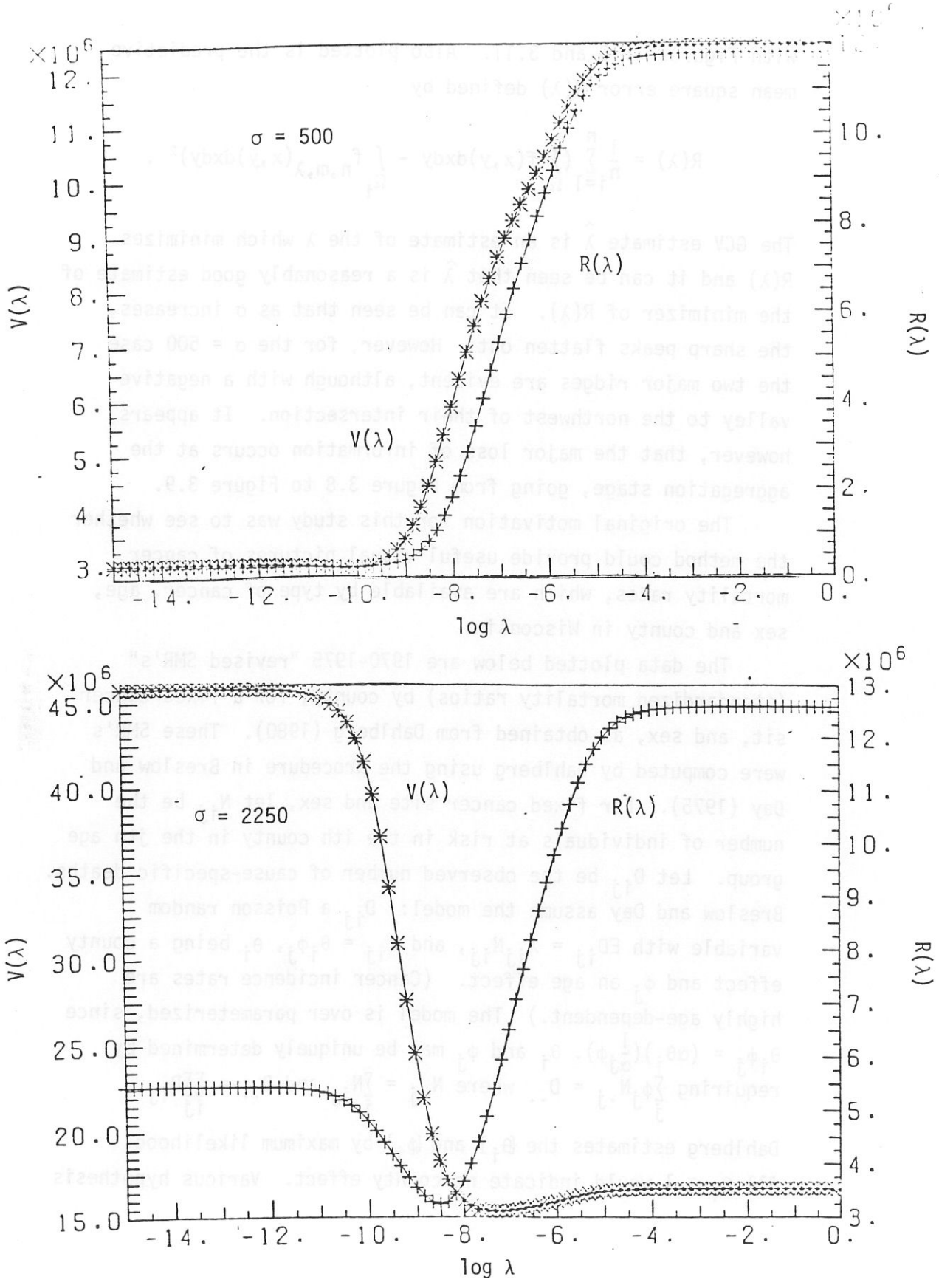

Figure 3.12. Cross Validation Functions and Predictive Mean Square Error Functions for Figures 3.10 and 3.11.

with Figures 3.10 and 3.11. Also plotted is the predictive mean square error $R(\lambda)$ defined by

$$R(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\left(\int_{\Omega_i} f(x,y)dxdy - \int_{\Omega_i} f_{n,m,\lambda}(x,y)dxdy\right)^2 .$$

The GCV estimate $\hat{\lambda}$ is an estimate of the λ which minimizes $R(\lambda)$ and it can be seen that $\hat{\lambda}$ is a reasonably good estimate of the minimizer of $R(\lambda)$. It can be seen that as σ increases, the sharp peaks flatten out. However, for the $\sigma = 500$ case the two major ridges are evident, although with a negative valley to the northwest of their intersection. It appears, however, that the major loss of information occurs at the aggregation stage, going from Figure 3.8 to Figure 3.9.

The original motivation for this study was to see whether the method could provide useful visual pictures of cancer mortality rates, which are available by type of cancer, age, sex and county in Wisconsin.

The data plotted below are 1970-1975 "revised SMR's" (standardized mortality ratios) by county, for a fixed cancer sit, and sex, as obtained from Dahlberg (1980). These SMR's were computed by Dahlberg using the procedure in Breslow and Day (1975). For fixed cancer site and sex, let N_{ij} be the number of individuals at risk in the ith county in the jth age group. Let D_{ij} be the observed number of cause-specific deaths. Breslow and Day assume the model: D_{ij} a Poisson random variable with $ED_{ij} = \lambda_{ij}N_{ij}$, and $\lambda_{ij} = \theta_i\phi_j$, θ_i being a county effect and ϕ_j an age effect. (Cancer incidence rates are highly age-dependent.) The model is over parameterized, since $\theta_i\phi_j = (\alpha\theta_i)(\frac{1}{\alpha_j}\phi)$. θ_i and ϕ_j may be uniquely determined by requiring $\sum_j \phi_j N_{\cdot j} = D_{\cdot\cdot}$ where $N_{\cdot j} = \sum_i N_{ij}$ and $D_{\cdot\cdot} = \sum\sum_{ij} D_{ij}$. Dahlberg estimates the $\{\theta_i\}$ and $\{\phi_j\}$ by maximum likelihood. All $\theta_i \equiv 1$ would indicate no county effect. Various hypothesis

test concerning the θ_i were conducted by Dahlberg, who found significant differences between counties for several data sets. Here we consider that the revised SMR's represent estimates of a county wide average SMR, and compute $f_{n,m,\lambda}$ as the minimizer of

$$\frac{1}{n}\sum_{i=1}^{n}(\theta_i - \frac{1}{|\Omega_i|}\int_{\Omega_i} f(x,y)dxdy)^2 + \lambda J_2(f).$$

If $\lambda = 0$, then $f_{n,m}$ is the minimizer of $J_2(f)$ subject to

$$\frac{1}{|\Omega_i|}\int_{\Omega_i} f(x,y)dxdy = \theta_i, \quad i = 1,2,\ldots,n. \qquad (3.3)$$

(Numerically the left hand side of (3.3) is replaced by

$$\frac{1}{N_i}\sum_{\varepsilon\Omega_i} f(t_k).)$$

Figure 3.13 gives the 1970-1975 female lung cancer revised SMR's by county from Dahlberg (1980). Figure 3.14 gives the volume matching histospline for the data of Figure 3.13. Figure 3.15 gives the volume smoothing histospline for the data of Figure 3.14. The smoothing parameter λ has been selected by the GCV method. Figure 3.16 gives the revised SMR's for 1970-1975 Male Rectal Cancer, and Figure 3.17 gives the volume matching histospline. For this data, GCV estimate of λ was essentially zero, so that only the volume matching histospline is given. Figures 3.18 and 3.19 give the volume matching histosplines for male lung cancer and male pancreatic cancer respectively. In both the cases the GCV estimate of λ was also 0.

We do not, at the present time have specific recommendations as to how to interpret the disease incidence contour plots. The difficulties concerning the interpretation of the individual revised SMR's themselves, considering the variables involved in collecting data of this type are formidable. See Dahlberg (1980). However, we believe that these maps have the potential

Fig. 3.13. 1970-1975 Female Lung Cancer, Revised SMR's by County

Figure 3.14. Female Lung Cancer Revised SMR's. Volume Matching Histospline. Contour Interval: 0.25.

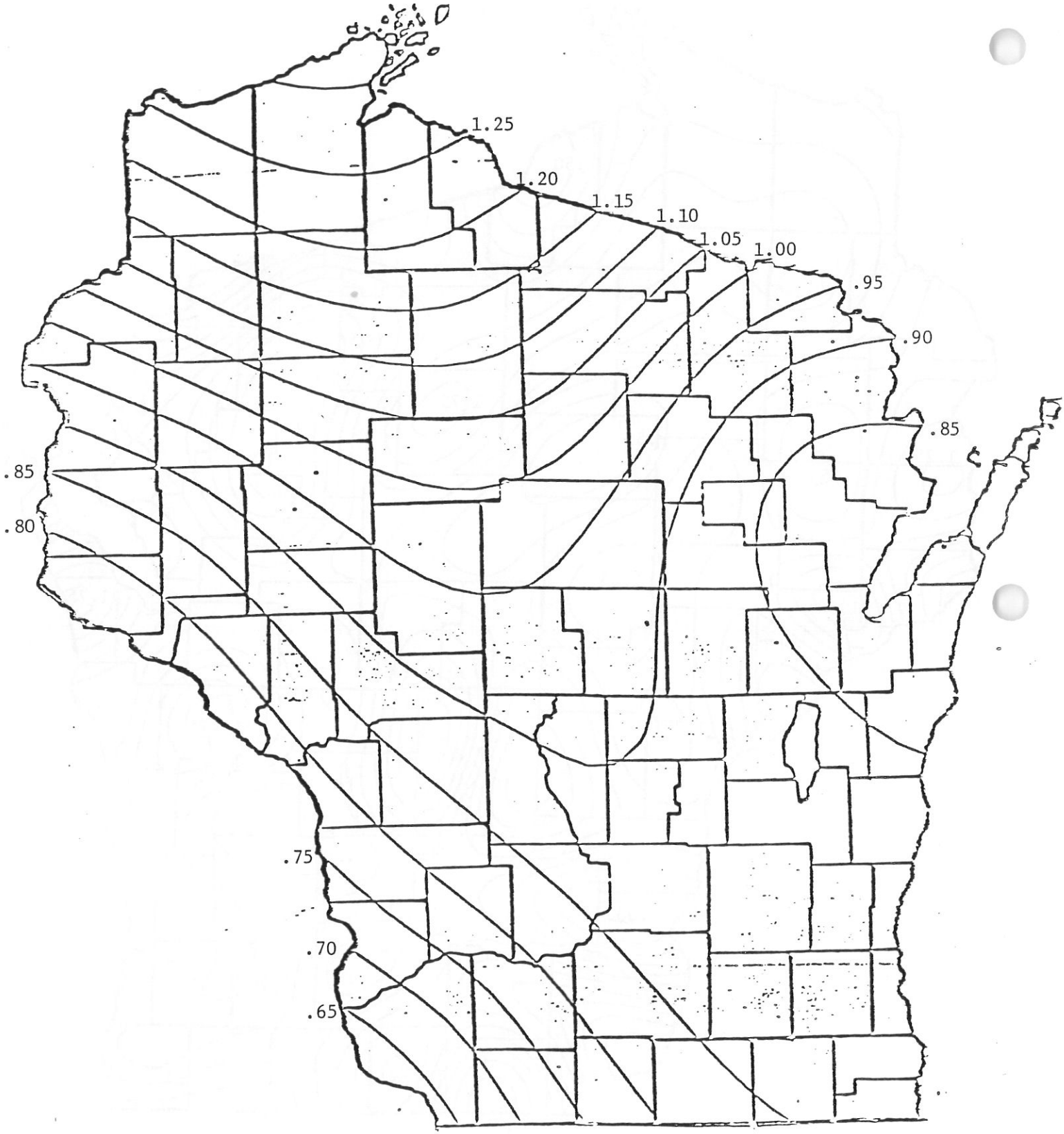

Figure 3.15. 1979-1975 Female Lung Cancer, Histospline Smoothed by GCV. Contour Interval: 0.05.

Figure 3.16. 1970-1975 Male Rectal Cancer, Revised SMR's by County.

Figure 3.17. 1970–1975 male rectal cancer, Volume Matching Histospline, $\lambda = \hat{\lambda} = 0$. Contour Interval: 0.25.

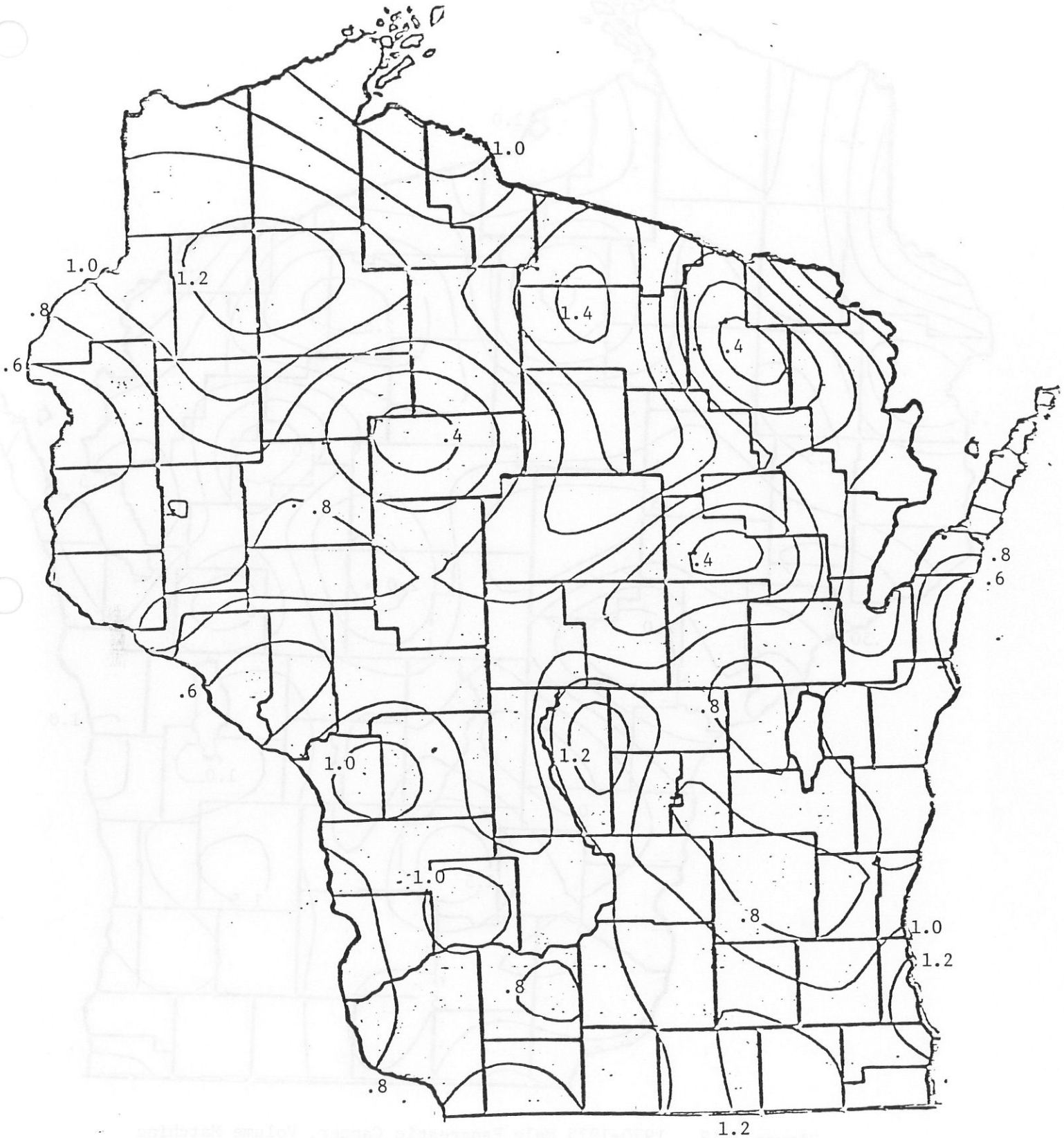

Figure 3.18. 1970-1975 Male Lung Cancer, Volume Matching Histospline, $\lambda = \hat{\lambda} = 0$. Contour interval: 0.2.

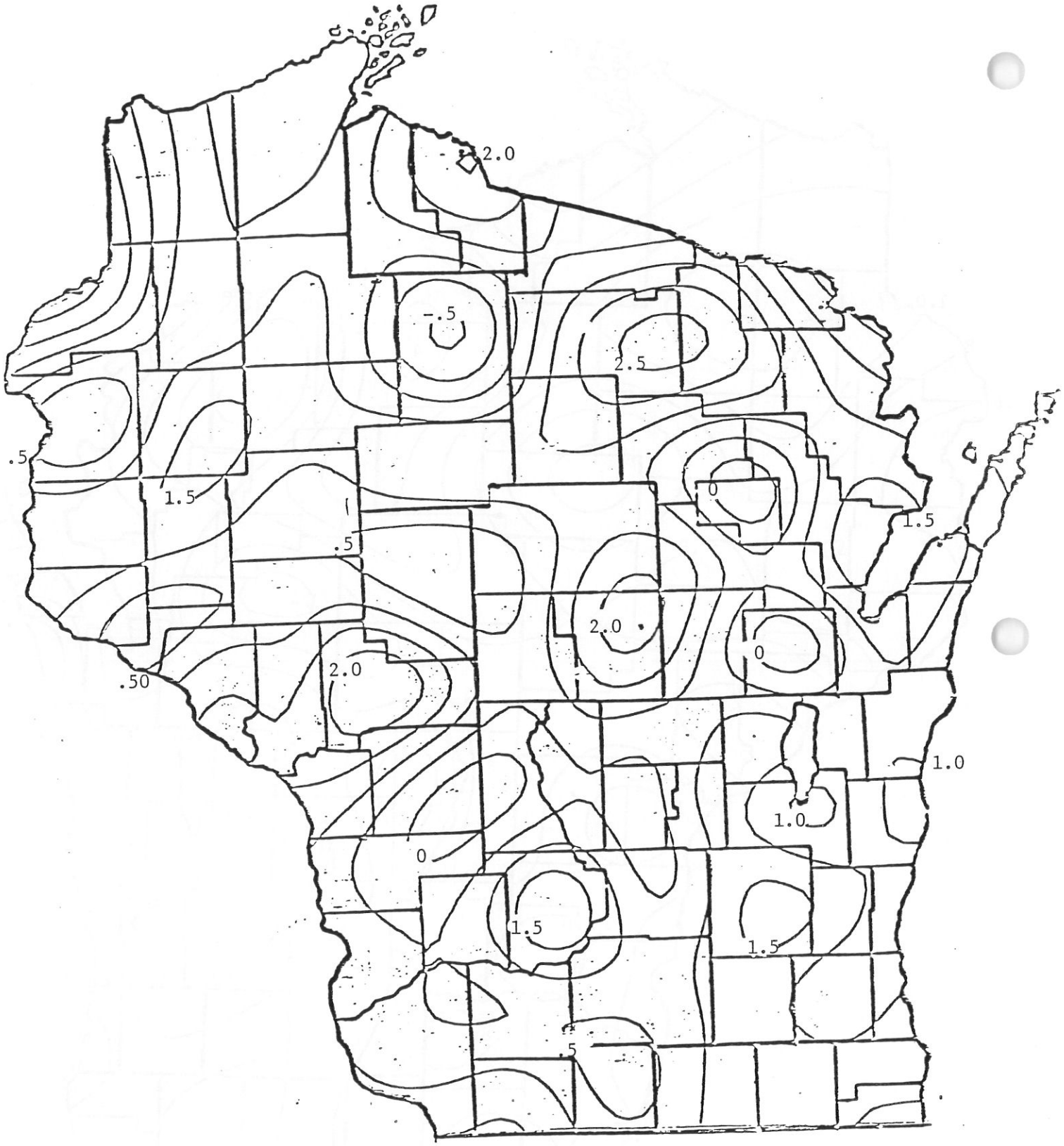

Figure 3.19. 1970-1975 Male Pancreatic Cancer, Volume Matching Histospline, $\lambda = \hat{\lambda} = 0$. Contour Interval: 0.5.

for use in "fishing expeditions" where the analyst may be seeking similar geographic patterns for variables that may, or may not be correlated. For example, contour maps made from point values of, for instance rainfall acidity, atmospheric or drinking water pollutants or soil components can be visually compared with revised SMR maps. For contour maps from point data, see Wahba and Wendelberger (1980). Variables that are aggregated differently can also be reduced to contour maps which can be compared. Maps for many variables could be screened visually and if common patterns are suspected, then possible relations between variables could be studied further, by more rigorous methods.

It is important to note that, while the calculation of one map can be very expensive, the calculations that depend on the data are the solution of equations (2.9), for $\tilde{c}$ and $\tilde{d}$, the calculation of ω via (2.14) and the minimization of V of (2.15). Other ingredients as noted at the end of section 2 are computed only once and stored. We originally evaluated $\tilde{f}_{n,m,\lambda}(t)$ from (2.7) on a 41×41 grid for input to the contour map subroutine used to make the plots. This size grid turned out to be sometimes esthetically unsatisfactory and was replaced by an 81×81 grid. Repeated evaluation of the $\tilde{\xi}_i$ on this grid was expensive. For repeated use, the values of the $\tilde{\xi}_i$ on an adequate grid should be done once and stored. Repeated use is then quite cheap.

For data from a non-negative source which may vary over several orders of magnitude (i.e. Wisconsin population data) the approach given here can be unsatisfactory since large negative estimates can occur. Nonnegatively constrained estimates are probably more appropriate. (See Dyn and Wong (1980), Wong (1980)). For data similar to the revised SMR's, however, the results can be quite reasonable. This remark is not intended to support methods which aggregate data before the analysis,

because as can be seen from the synthetic data, much resolution can be lost in the aggregation process. However, much medical data is not readily available in disaggregated form.

We have not discussed any theoretical properties of this estimate and in fact none are known to this author, other than those which can be inferred from the theoretical results known for the general class of nonparametric methods of which this is a member as discussed in, for example Wahba (1977a,b,1978). Also, the error structure of revised SMR's does not fit into the error structure commonly assumed to justify the volume smoothing used here. Significance tests for spatial differences and geographic correlations between different variables (some of which might be aggregated in different ways!) remains to be developed in the present context and we feel that it is necessary to do this before considering the above methods as more than a graphical tool.

ACKNOWLEDGEMENTS

The computer program was written by A. Kirsch. The work of A. Kirsch and the author was supported by the USARO under Grants DAAG29-77-G-0207 and DAAG29-80-K0042. We thank S. Dahlberg for assistance with the data and J. Crowley and S. Leurgans for additional computer time, supported by the NIH under Grant 144-P-301. We thank D. Bates, G. Golub and J. Wendelberger for advice concerning computational problems.

BIBLIOGRAPHY

Adams, R.A. (1975). Sobolev Spaces. Academic Press, New York.

Agarwal, G. and Studden, W. (1980). Asymptotic integrated mean square error using least squares and bias-minimizing splines. Ann. Statist. 8, 6, 1307-1325.

Boneva, L., Kendall, D. and Stefanov, J. (1971). Spline transformations. J. Royal Statistical Society, Ser. B, 33, 1-70.

Breslow, N.E. and Day, N.E. (1975). Indirect standardization and multiplicative models for rates, with reference to the age adjustment of cancer incidence and relative frequency data. J. Chron. Dis. 28, 289-303.

Chambless, D. (1980). Radiological data analysis in the time and frequency domain, II, Department of Mathematics, Auburn University, Montgomery, Alabama.

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

Dahlberg, S. (1980). Detailed description of Wisconsin cancer mortality data. Wisconsin Clinical Cancer Center-Biostatistics, University of Wisconsin-Madison, TR #5.

Dongarra, J.J., Bunch, J.R., Moler, C.B., Stewart, G.W. (1979). LINPACK Users' Guide. Society for Industrial and Applied Mathematics, Philadelphia, PA.

Duchon, J. (1975). Fonctions-spline du type plaque mince en dimension 2. #231, Seminaire d'Analyse Numerique, Universite Scientifique et Medicale de Grenoble.

Duchon, J. (1977). 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 471, Springer 1977.

Duchon, J. (1978). Sur l'erreur d'interpolation des fonctions de plusieurs variables par les D^m-splines. R.A.I.R.O. analyse numerique, 12, 4, 325-334.

Dyn, N. and Wahba, G. (1979). On the estimation of functions of several variables from aggregated data. Mathematics Research Center, University of Wisconsin-Madison, TSR # 1974, to appear, SIAM J. Math. Anal.

Dyn, N., Wahba, G. and Wong, W.H. (1979). Comments to "Smooth pycnophylactic interpolation for geographical regions, by W. Tobler. J. Am. Stat. Assoc., 74, 367, 530-535.

Dyn, N. and Wong, W.H. (1981). On the solution of a constrained minimization problem in $H'(\Omega)$ related to density estimation, Mathematics Research Center, University of Wisconsin-Madison TSR, to appear.

Kuhn, R. (1975). Reproducing kernel Hilbert spaces, with applications to control theory, univariate and bivariate spline density estimation. Ph.D. Thesis, Department of Statistics, University of Wisconsin-Madison.

Laurent, P.J. (1979). An algorithm for the computation of spline functions with inequality constraints, manuscript, #335, Seminaire d'Analyse Numerique, Universite Scientifique et Medicale de Grenoble.

Meinquet, J. (1978). Multivariate interpolation at arbitrary points made simple. To appear, Z. Angewandte Mathematik und Physik.

Paihua Montes, L. (1978). Quelques methodes numeriques pour le calcul de fonctions splines a une et plusiers variables, Thesis, Universite Scientifique et Medicale de Grenoble, Analyse Numerique.

Schoenberg, I.J. (1973). Splines and histograms, with an appendix by C. deBoor, in Spline Functions and Approximation Theory, Proceedings of the Symposium at the University of Alberta, A. Meir and A. Sharma, eds. Berkhaeuser Verlag, Basel, 277-358.

Smith, B.T., Boyle, J.M., Garbow, B.S., Ikebe, Y., Klema, V.C. and Moler, C.B. (1976). Matrix Eigensystem Routines, EISPACK Guide, Springer-Verlag.

Stone, C.J. (1980). Optimal rates of convergence for nonparametric estimators, Ann. Statist. 8, 6, 1348-1360.

Tobler, W. (1979). Smooth pycnophylactic interpolation for geographical regions. J. Am. Stat. Assoc., 74, 367, 519-530.

Wahba, G. (1975). Interpolating spline methods for density estimation I. Equi-spaced knots. Ann. Statist. 3, 1, 30-48.

Wahba, G. (1977a). Comments to "Consistent nonparametric regression by C.J. Stone, Ann. Statist. 5, 4, 637-645.

Wahba, G. (1977b). Practical approximate solutions to linear operator equations when the data are noisy, SIAM J. Num. Anal, 14, 4.

Wahba, G. (1978). Improper priors, spline smoothing and the problem of guarding against model errors in regression. J. Roy. Stat. Soc. B, 40, 3, 364-372.

Wahba, G. (1979). How to smooth curves and surfaces with splines and cross-validation, in Proceedings of the 24th Conference on the Design of Experiments, U.S.A.R.O. Report 79-2.

Wahba, G., and Wendelberger, J. (1980). Some new mathematical methods for variational objective analysis using splines and cross validation. Monthly Weather Review, 108, 8, 1122-1143.

Wong, W.J. (1980). An analysis of the volume matching problem and related topics in smooth density estimation. Ph.D. Thesis, Department of Statistics, University of Wisconsin-Madison.

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER: Technical Report No. 638

2. GOVT ACCESSION NO.:

3. RECIPIENT'S CATALOG NUMBER:

4. TITLE (and Subtitle): NUMERICAL EXPERIMENTS WITH THE THIN PLATE HISTOSPLINE

5. TYPE OF REPORT & PERIOD COVERED: Scientific Interim

6. PERFORMING ORG. REPORT NUMBER:

7. AUTHOR(s): Grace Wahba

8. CONTRACT OR GRANT NUMBER(s): DAAG-29-80-K0042

9. PERFORMING ORGANIZATION NAME AND ADDRESS: 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, Durham, N.C.

12. REPORT DATE: March 1981

13. NUMBER OF PAGES: 39

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)

Distribution of this document 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 histosplines, volume-matching splines, contour plots for standardized mortality ratios

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

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

ABSTRACT

The thin plate volume matching and volume smoothing histosplines are described. These histosplines are suitable for estimating densities or incidence rates as a function of position on the plane when data is aggregated by area, for example by county. We give a numerical algorithm for the volume matching histospline and for the volume smoothing histospline using generalized cross validation to estimate the smoothing parameter. Some numerical experiments were performed using synthetic data, population data and SMR's (standardized mortality ratios) aggregated by county over the state of Wisconsin. The method turns out to be not particularly suited for obtaining population density maps where the population density can vary by two orders of magnitude, because the histospline can be negative in unpleasant ways. However the fitting of SMR's, which are all about the same order of magnitude, results in some esthetically pleasing pictures which may be used to search visually for geographic patterns. A number of open questions remain.