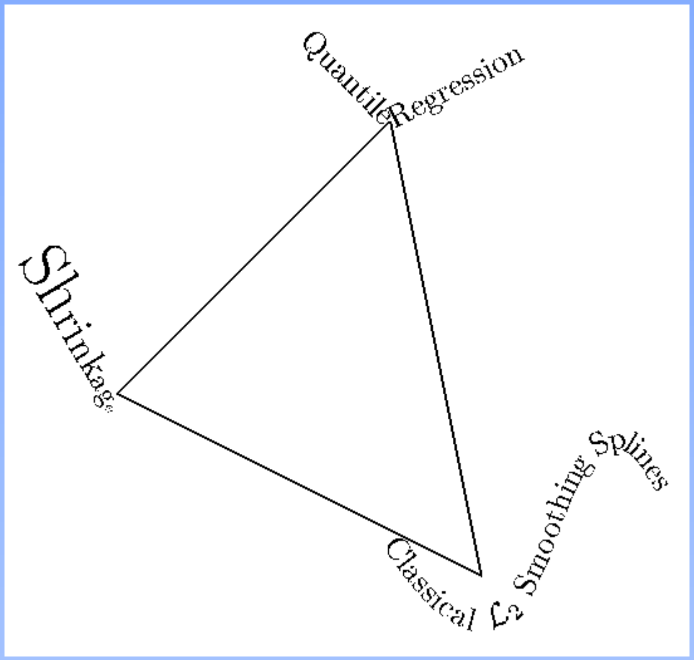
Penalty Methods for Bivariate Smoothing and Chicago Land Values

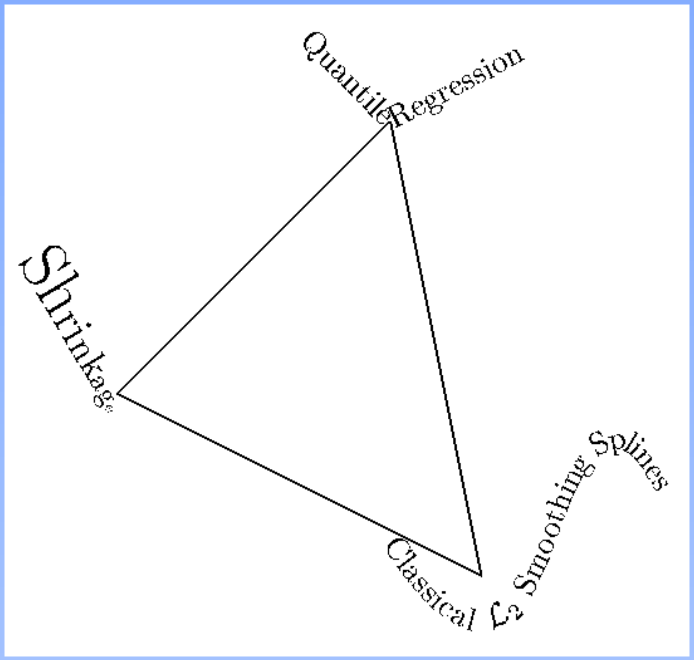
Roger Koenker
University of Illinois, Urbana-Champaign
Ivan Mizera
University of Alberta, Edmonton

Northwestern University: October 2001

Or ... Pragmatic Goniolatry



Or ... Pragmatic Goniolatry



"Goniolatry, or the worship of angles, ..."
Thomas Pynchon (*Mason and Dixon*, 1997).

Univariate $\mathcal{L}_2$ Smoothing Splines

The Problem:

$$\min_{g \in \mathcal{G}} \sum_{i=1}^{n} (y_i - g(x_i))^2 + \lambda \int_a^b (g''(x))^2 dx,$$

Gaussian Fidelity to the data:

$$\sum_{i=1}^{n} (y_i - g(x_i))^2$$

Roughness Penalty on $\hat{g}$:

$$\lambda \int_a^b (g''(x))^2 dx,$$

Quantile Smoothing Splines

The Problem:

$$\min_{g \in \mathcal{G}} \sum_{i=1}^{n} \rho_\tau(y_i - g(x_i)) + \lambda J(g),$$

Quantile Fidelity to the Data:

$$\rho_\tau(u) = u(\tau - I(u < 0))$$

Total Variation Roughness Penalty on $\hat{g}$:

$$J(g) = V(g') = \int |g''(x)| dx,$$

Ref: Koenker, Ng, Portnoy (*Biometrika*, 1994)

Thin Plate Smoothing Splines

Problem:

$$\min_g \sum_{i=1}^{n} (z_i - g(x_i, y_i))^2 + \lambda J(g)$$

Roughness Penalty:

$$J(g, \Omega) = \int\int_{\Omega} (g_{xx}^2 + 2g_{xy}^2 + g_{yy}^2) dx dy$$

Equivariant to translations and rotations.

Easy to compute provided $\Omega = \mathbb{R}^2$. But this creates boundary problems.

References: Wahba(1990), Green and Silverman(1998).

Thin Plate Smoothing Splines

Problem:

$$\min_g \sum_{i=1}^{n} (z_i - g(x_i, y_i))^2 + \lambda J(g)$$

Roughness Penalty:

$$J(g, \Omega) = \int\int_\Omega (g_{xx}^2 + 2g_{xy}^2 + g_{yy}^2) dxdy$$

Equivariant to translations and rotations.

Easy to compute provided $\Omega = \mathbb{R}^2$. But this creates boundary problems.

References: Wahba(1990), Green and Silverman(1998).

Question: How to extend total variation penalties to $g : \mathbb{R}^2 \to \mathbb{R}$?

Thin Plate Example

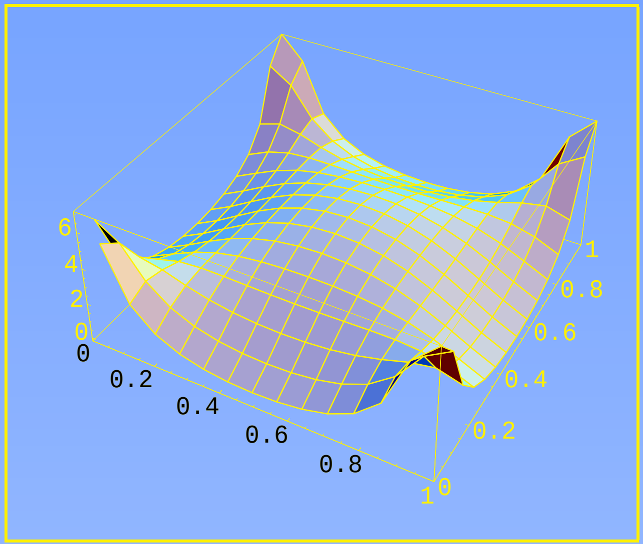


Figure 1: Integrand of the thin plate penalty for the He, Ng, and Portnoy tent function interpolant of the points $\{(0,0,0),(0,1,0),(1,0,0),(1,1,1)\}$. The boundary effects are created by extension of the optimization over all of $\mathbb{R}^2$. For the restricted domain $\Omega = [0,1]^2$ the optimal solution $g(x,y) = xy$ has considerably smaller penalty: 2 versus 2.77 for the unrestricted domain solution.

Three Variations on Total Variation for $f : [a, b] \to \mathbb{R}$

1. Jordan(1881)

$$V(f) = \sup_{\pi} \sum_{k=0}^{n-1} |f(x_{k+1}) - f(x_k)|$$

where π denotes partitions: $a = x_0 < x_1 < \ldots < x_n = b$.

Three Variations on Total Variation for $f : [a, b] \to \mathbb{R}$

1. Jordan(1881)

$$V(f) = \sup_{\pi} \sum_{k=0}^{n-1} |f(x_{k+1}) - f(x_k)|$$

where π denotes partitions: $a = x_0 < x_1 < \ldots < x_n = b$.

2. Banach (1925)

$$V(f) = \int N(y) dy$$

where $N(y) = \text{card}\{x : f(x) = y\}$ is the Banach indicatrix

Three Variations on Total Variation for $f : [a, b] \to \mathbb{R}$

1. Jordan(1881)

$$V(f) = \sup_{\pi} \sum_{k=0}^{n-1} |f(x_{k+1}) - f(x_k)|$$

   where π denotes partitions: $a = x_0 < x_1 < \ldots < x_n = b$.

2. Banach (1925)

$$V(f) = \int N(y) dy$$

   where $N(y) = \text{card}\{x : f(x) = y\}$ is the Banach indicatrix

3. Vitali (1905)

$$V(f) = \int |f'(x)| dx$$

   for absolutely continuous f.

Total Variation for $f : \mathbb{R}^k \longrightarrow \mathbb{R}^m$

A convoluted history ... de Giorgi (1954)

For smooth $f : \mathbb{R} \longrightarrow \mathbb{R}$

$$V(f, \Omega) = \int_\Omega |f'(x)| dx$$

Total Variation for $f: \mathbb{R}^k \to \mathbb{R}^m$

A convoluted history ... de Giorgi (1954)

For smooth $f: \mathbb{R} \to \mathbb{R}$

$$V(f, \Omega) = \int_{\Omega} |f'(x)| dx$$

For smooth $f: \mathbb{R}^k \to \mathbb{R}^m$

$$V(f, \Omega, \|\cdot\|) = \int_{\Omega} \| \nabla f(x) \| dx$$

Total Variation for $f : \mathbb{R}^k \to \mathbb{R}^m$

A convoluted history ... de Giorgi (1954)

For smooth $f : \mathbb{R} \to \mathbb{R}$

$$V(f, \Omega) = \int_\Omega |f'(x)| dx$$

For smooth $f : \mathbb{R}^k \to \mathbb{R}^m$

$$V(f, \Omega, \| \cdot \|) = \int_\Omega \| \nabla f(x) \| dx$$

Extension to nondifferentiable f via theory of distributions.

$$V(f, \Omega, \| \cdot \|) = \int_\Omega \| \nabla f(x) * \varphi_\epsilon \| dx$$

Roughness Penalties for $g : \mathbb{R}^k \to \mathbb{R}$

For smooth $g : \mathbb{R} \to \mathbb{R}$

$$J(g, \Omega) = V(g', \Omega) = \int_\Omega |g''(x)| dx$$

Roughness Penalties for $g: \mathbb{R}^k \to \mathbb{R}$

For smooth $g: \mathbb{R} \to \mathbb{R}$

$$J(g, \Omega) = V(g', \Omega) = \int_\Omega |g''(x)| dx$$

For smooth $g: \mathbb{R}^k \to \mathbb{R}$

$$J(g, \Omega, \|\cdot\|) = V(\nabla g, \Omega, \|\cdot\|) = \int_\Omega \|\nabla^2 g\| dx$$

Roughness Penalties for $g : \mathbb{R}^k \to \mathbb{R}$

For smooth $g : \mathbb{R} \to \mathbb{R}$

$$J(g, \Omega) = V(g', \Omega) = \int_\Omega |g''(x)| dx$$

For smooth $g : \mathbb{R}^k \to \mathbb{R}$

$$J(g, \Omega, \| \cdot \|) = V(\nabla g, \Omega, \| \cdot \|) = \int_\Omega \| \nabla^2 g \| dx$$

Again, extension to nondifferentiable g via theory of distributions.

Roughness Penalties for $g : \mathbb{R}^k \to \mathbb{R}$

For smooth $g : \mathbb{R} \to \mathbb{R}$

$$J(g, \Omega) = V(g', \Omega) = \int_\Omega |g''(x)| dx$$

For smooth $g : \mathbb{R}^k \to \mathbb{R}$

$$J(g, \Omega, \|\cdot\|) = V(\nabla g, \Omega, \|\cdot\|) = \int_\Omega \|\nabla^2 g\| dx$$

Again, extension to nondifferentiable g via theory of distributions.

Choice of norm is subject to dispute.

Invariance Considerations

Invariance helps to narrow the choice of norm.

For orthogonal U and symmetric matrix H, we would like:

$$\|U^\top H U\| = \|H\|$$

Invariance Considerations

Invariance helps to narrow the choice of norm.

For orthogonal U and symmetric matrix H, we would like:

$$\|U^\top HU\| = \|H\|$$

Examples:

$$\|\nabla^2 g\| = \sqrt{g_{xx}^2 + 2g_{xy}^2 + g_{yy}^2}$$

$$\|\nabla^2 g\| = |trace\,\nabla^2 g|$$

$$\|\nabla^2 g\| = max|eigenvalue(H)|$$

Invariance Considerations

Invariance helps to narrow the choice of norm.

For orthogonal U and symmetric matrix H, we would like:

$$\|U^\top HU\| = \|H\|$$

Examples:

$$\|\nabla^2 g\| = \sqrt{g_{xx}^2 + 2g_{xy}^2 + g_{yy}^2}$$
$$\|\nabla^2 g\| = |trace\, \nabla^2 g|$$
$$\|\nabla^2 g\| = max|eigenvalue(H)|$$

$$\|\nabla^2 g\| = |g_{xx}| + 2|g_{xy}| + |g_{yy}|$$
$$\|\nabla^2 g\| = |g_{xx}| + |g_{yy}|$$

Invariance Considerations

Invariance helps to narrow the choice of norm.

For orthogonal U and symmetric matrix H, we would like:

$$\|U^\top HU\| = \|H\|$$

Examples:

$$\|\nabla^2 g\| = \sqrt{g_{xx}^2 + 2g_{xy}^2 + g_{yy}^2}$$

$$\|\nabla^2 g\| = |trace\, \nabla^2 g|$$

$$\|\nabla^2 g\| = max|eigenvalue(H)|$$

$$\|\nabla^2 g\| = |g_{xx}| + 2|g_{xy}| + |g_{yy}|$$

$$\|\nabla^2 g\| = |g_{xx}| + |g_{yy}|$$

Solution of associated variational problems is difficult!

Triograms

Following Hansen, Kooperberg and Sardy (JASA, 1998):

Let $\mathcal{U}$ be a compact region of the plane, and let Δ denote a collection of sets $\delta_i : i = 1, \ldots, n$ with disjoint interiors such that $\mathcal{U} = \cup_{\delta \in \Delta} \delta$.

If $\delta \in \Delta$ are planar triangles, Δ is a triangulation of $\mathcal{U}$,

Definition: *A continuous, piecewise linear function on a triangulation, Δ, is called a triogram.*

Triograms

Following Hansen, Kooperberg and Sardy (JASA, 1998):

Let $\mathcal{U}$ be a compact region of the plane, and let Δ denote a collection of sets $\delta_i : i = 1, \ldots, n$ with disjoint interiors such that $\mathcal{U} = \cup_{\delta \in \Delta} \delta$.

If $\delta \in \Delta$ are planar triangles, Δ is a triangulation of $\mathcal{U}$,

Definition: *A continuous, piecewise linear function on a triangulation, Δ, is called a triogram.*

For triograms roughness is less ambiguous.

A Roughness Penalty for Triograms

For triograms the "ambiguity of the norm" problem for total variation roughness penalties is resolved.

Theorem. Suppose that $g : \mathbb{R}^2 \to \mathbb{R}$, is a piecewise-linear function on the triangulation, Δ. For any coordinate-independent penalty, J, there is a constant c dependent only on the choice of the norm such that

$$J(g) = cJ_{\triangle}(g) = c\sum_{e} \|\nabla g_e^{+} - \nabla g_e^{-}\| \, \|e\| \tag{1}$$

where e runs over all the interior edges of the triangulation $\|e\|$ is the length of the edge e, and $\|\nabla g_e^{+} - \nabla g_e^{-}\|$ is the length of the difference between gradients of g on the triangles adjacent to e.

Computation of Median Triograms

The Problem:

$$\min_{g \in \mathcal{G}_\triangle} \sum |z_i - g(x_i, y_i)| + \lambda J_\triangle(g)$$

can be reformulated as an augmented ℓ_1 (median) regression problem,

$$\min_{\beta \in \mathrm{IR}^p} \sum_{i=1}^{n} |z_i - a_i^\top \beta| + \lambda \sum_{k=1}^{M} |h_k^\top \beta|.$$

where β denotes a vector of parameters representing the values taken by the function g at the vertices of the triangulation $\triangle$. The a_i are barycentric coordinates of the (x_i, y_i) points in terms of these vertices, and the h_k represent the penalty contribution in terms of these vertices.

Computation of Median Triograms

The Problem:

$$\min_{g \in \mathcal{G}_\triangle} \sum |z_i - g(x_i, y_i)| + \lambda J_\triangle(g)$$

can be reformulated as an augmented ℓ_1 (median) regression problem,

$$\min_{\beta \in \mathbb{R}^p} \sum_{i=1}^{n} |z_i - a_i^\top \beta| + \lambda \sum_{k=1}^{M} |h_k^\top \beta|.$$

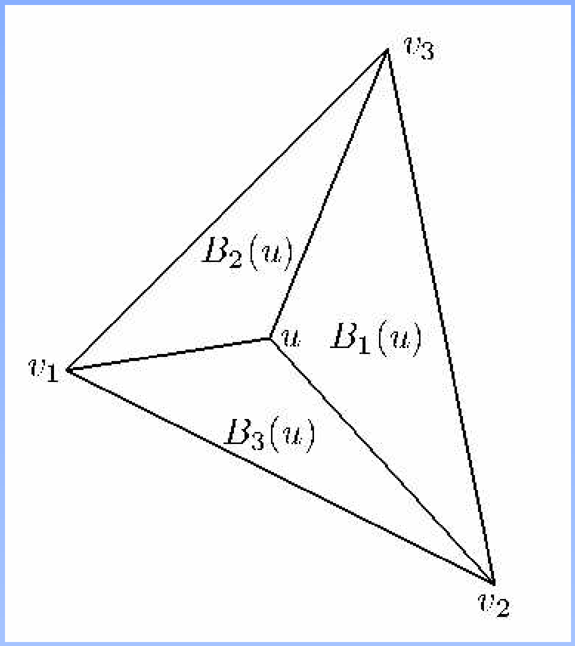
where β denotes a vector of parameters representing the values taken by the function g at the vertices of the triangulation $\triangle$. The a_i are barycentric coordinates of the (x_i, y_i) points in terms of these vertices, and the h_k represent the penalty contribution in terms of these vertices.

Extensions to quantile and mean triograms are straightforward.

Barycentric Coordinates

Triograms, $\mathcal{G}$, on Δ constitute a linear space with elements

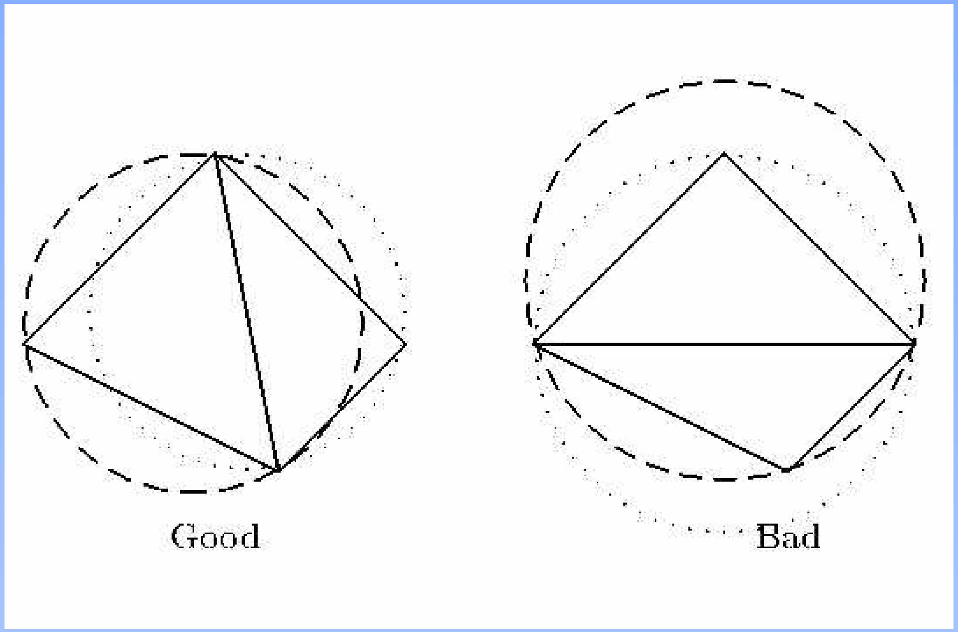
$$g(u) = \sum_{i=1}^{3} \alpha_i B_i(u) \quad u \in \delta \subset \Delta \qquad B_1(u) = \frac{\text{Area }(u, v_2, v_3)}{\text{Area }(v_1, v_2, v_3)} \text{ etc.}$$



Delaunay Triangulation

Properties of Delaunay triangles:

- Circumscribing circles of Delaunay triangles exclude other vertices,
- Maximize the minimum angle of the triangulation.



Robert Delaunay

B.N. Delone (1890-1973)

Four Median Triograms Fits

Consider estimating the noisy cone:

$$z_i = \max\{0, 1/3 - 1/2\sqrt{x_i^2 + y_i^2}\} + u_i,$$

with the (x_i, y_i)'s generated as independent uniforms on $[-1, 1]^2$, and with the u_i are iid Gaussian with standard deviation $\sigma = .02$. With sample size $n = 400$, the triogram problems are roughly 1600 by 400, but very sparse.

Four Median Triograms Fits

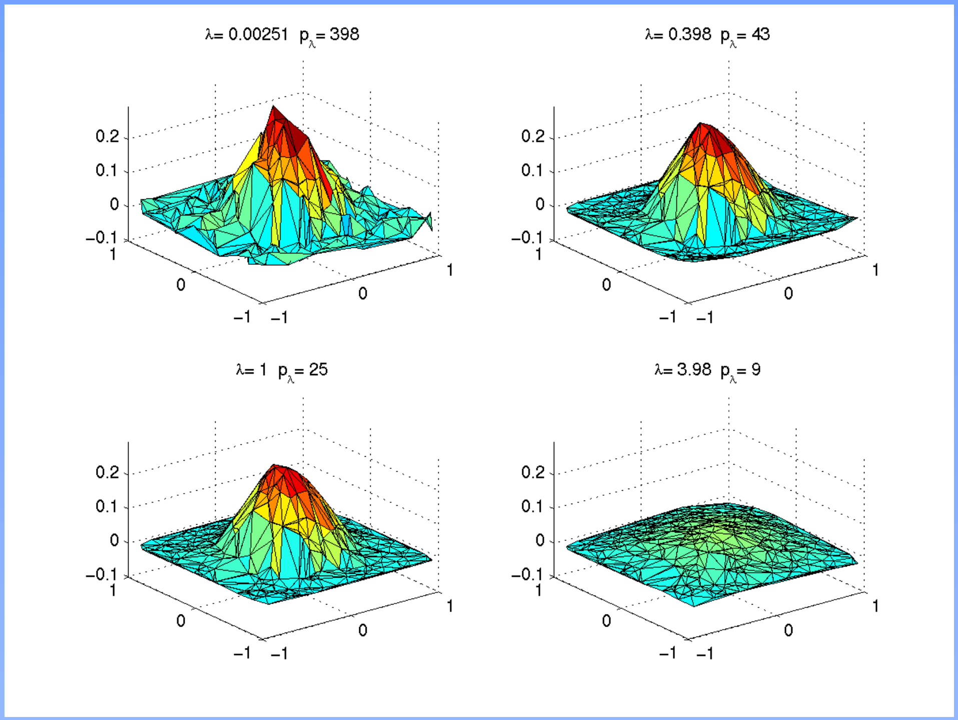
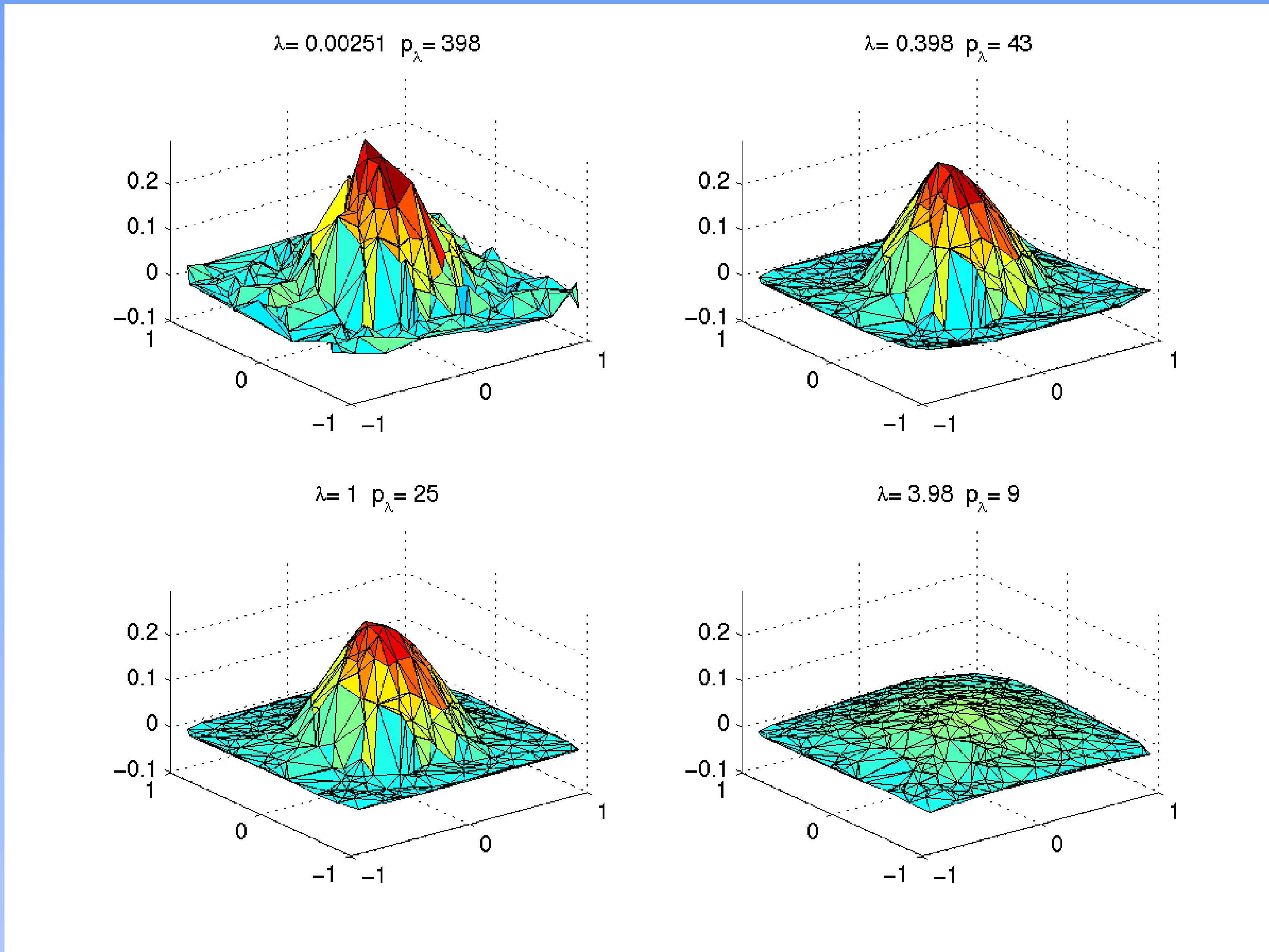


Figure 2: Four median triogram fits for the inverted cone example. The values of the smoothing parameter λ and the number of interpolated points in the fidelity component of the objective function, p_λ are indicated above each of the four plots.


λ= 0.00251 p_λ= 398
λ= 0.398 p_λ= 43
λ= 1 p_λ= 25
λ= 3.98 p_λ= 9
0.2
0.1
0
−0.1
1
0
−1
−1
0
1

Four Mean Triograms Fits

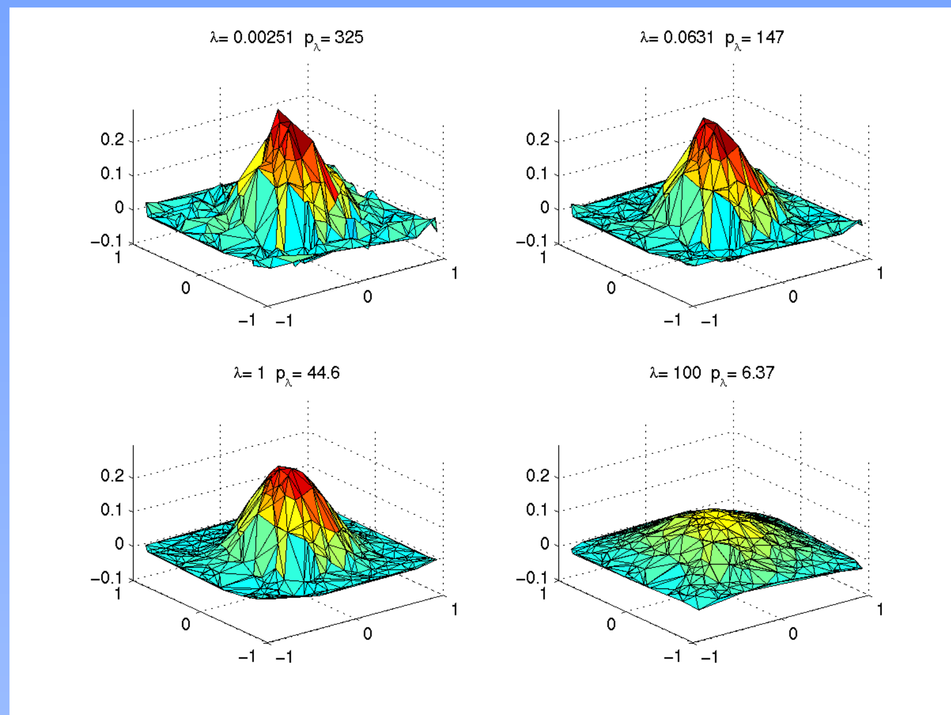


Figure 3: Four mean triogram fits for the noisy cone example. The values of the smoothing parameter λ and the trace of the linear operator defining the estimator, p_λ are indicated above each of the four plots.

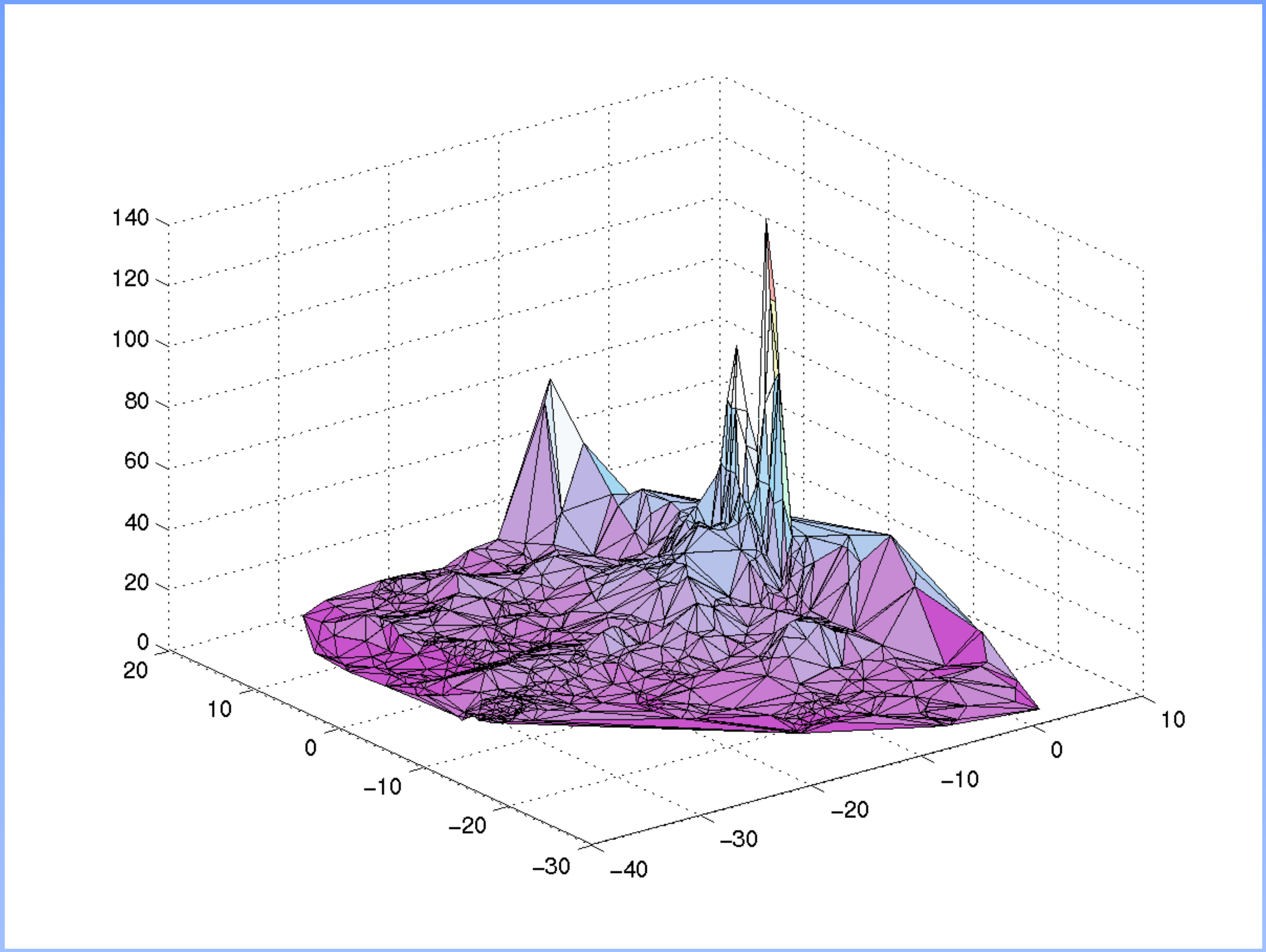


Figure 4: Perspective Plot of Median Model for Chicago Land Values. Based on 1194 land sales in Chicago Metropolitan Area in 1995-97, prices in dollars per square foot.

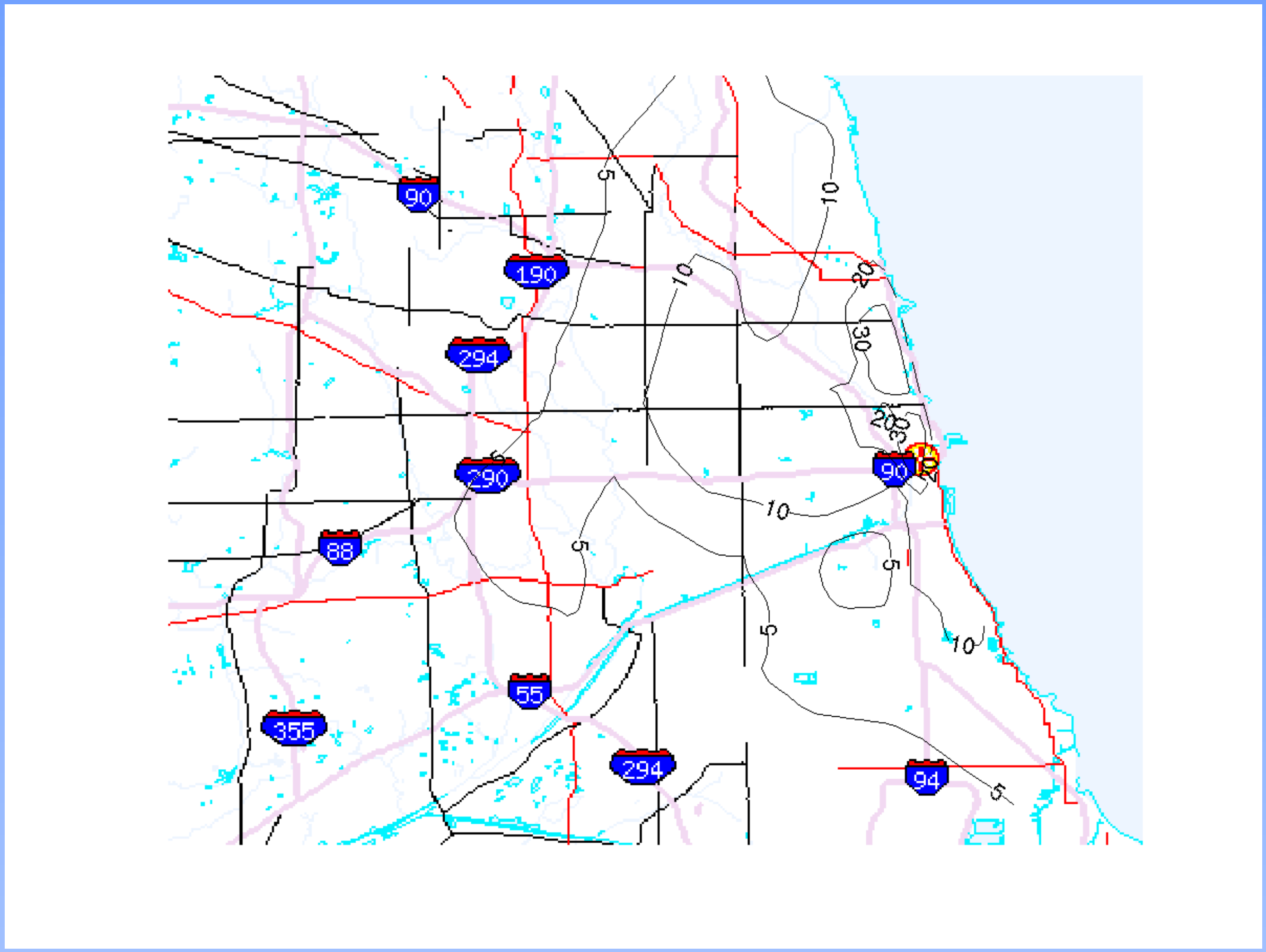


Figure 5: Contour Plot of First Quartile Model for Chicago Land Values.

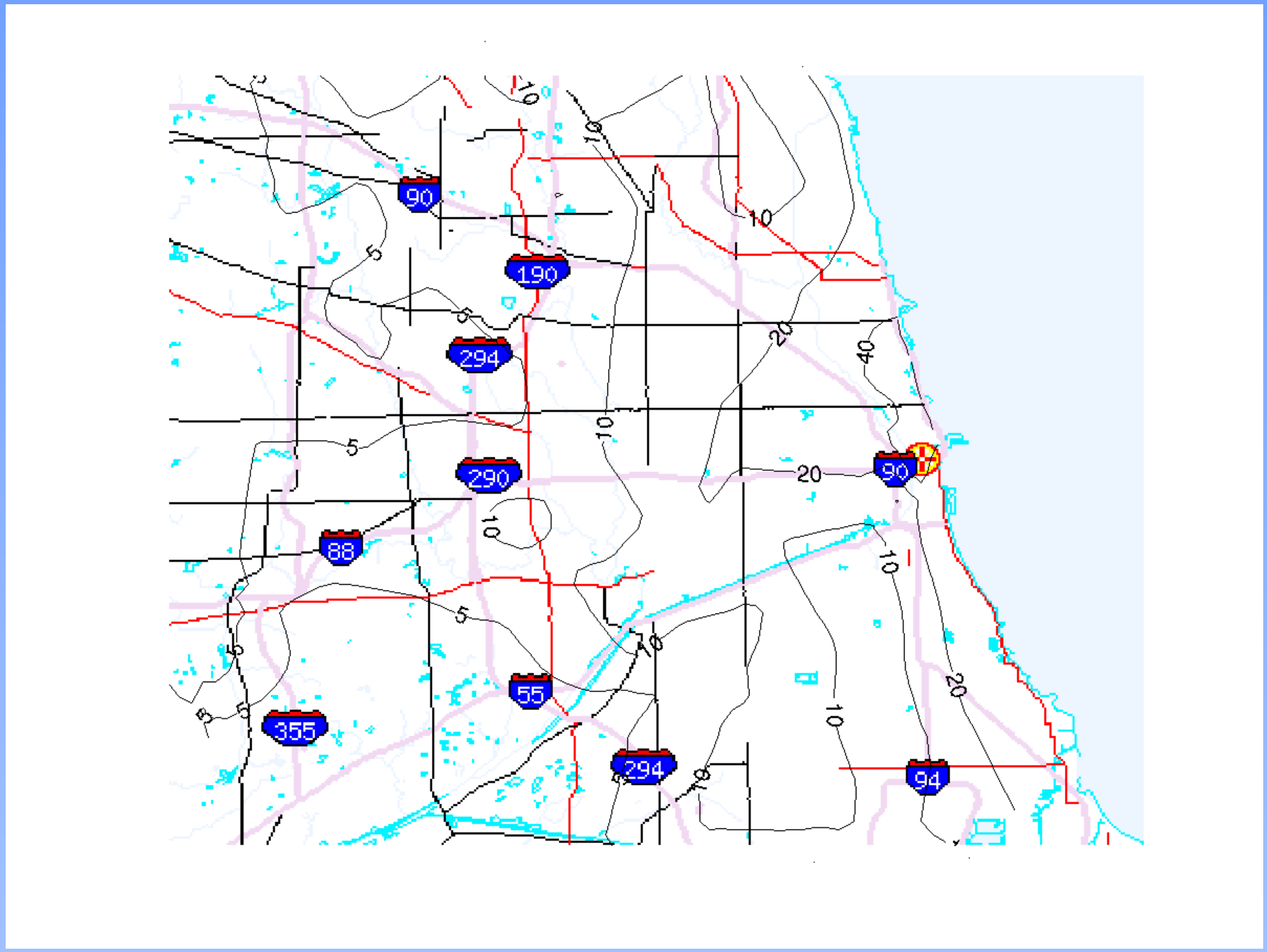


Figure 6: Contour Plot of Median Model for Chicago Land Values.

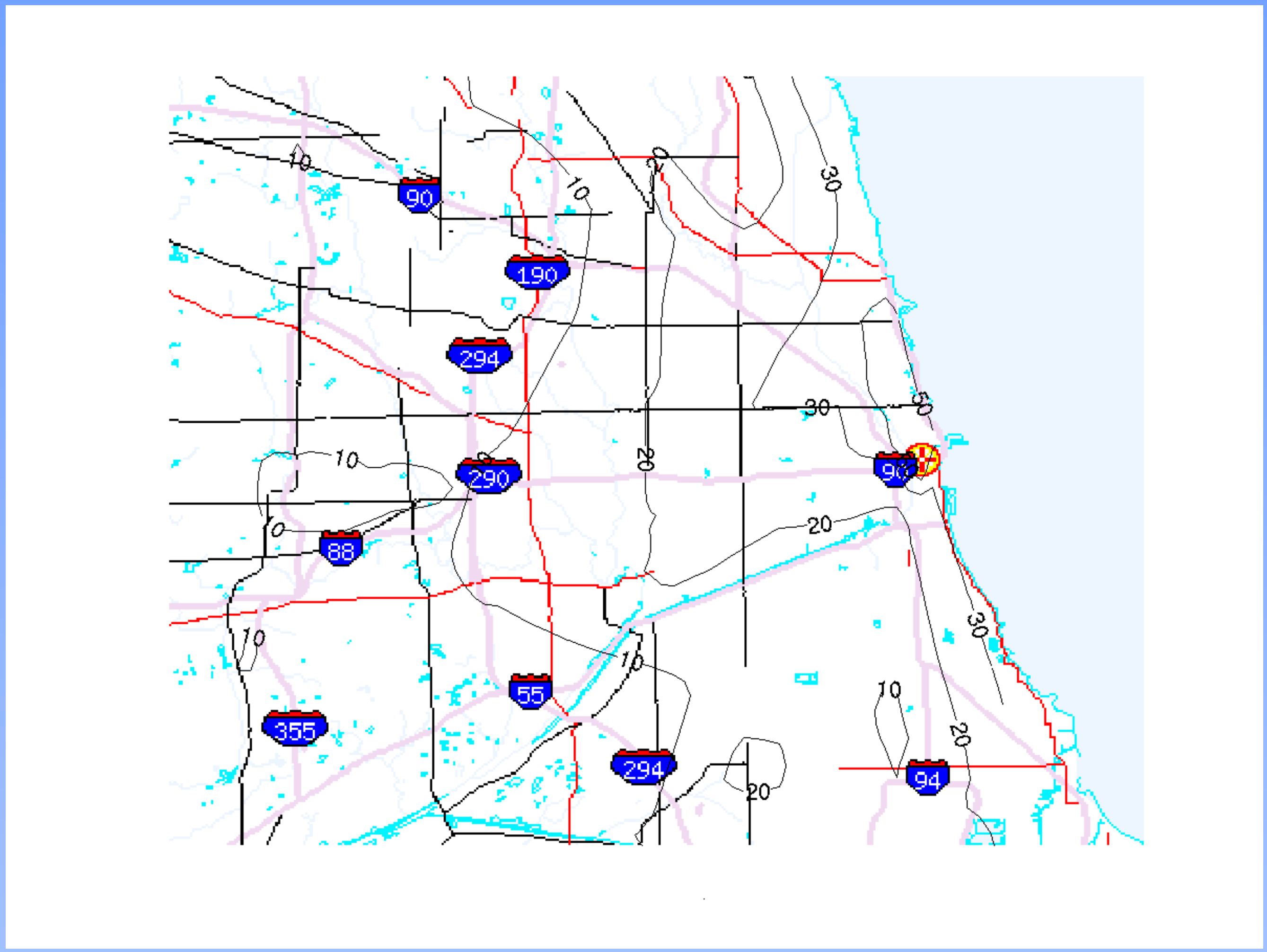


Figure 7: Contour Plot of Third Quartile Model for Chicago Land Values.

Automatic λ Selection

Schwarz Criterion:

$$\log(n^{-1}\sum \rho_\tau(z_i - \hat{g}_\lambda(x_i, y_i))) + (2n)^{-1} p_\lambda \log n.$$

where the dimension of the fitted function, p_λ, is defined as the number of points interpolated by the fitted function $\hat{g}_\lambda$. Other approaches: Stein's unbiased risk estimator, Donoho and Johnstone (1995), and e.g. Antoniadis and Fan (2001).

Extensions

Triograms can be constrained to be convex (or concave) by imposing m additional linear inequality constraints, one for each interior edge of the triangulation. This might be interesting for estimating bivariate densities since we could impose, or test (?) for log-concavity. Now computation is somewhat harder since the fidelity is more complicated.

Partial linear model applications are quite straightforward.

Extensions to penalties involving $V(g)$ may also prove interesting.

Monte-Carlo Performance

Design: He and Shi (1996)

$$z_i = g_0(x_i, y_i) + u_i \;\; i = 1, ..., 100.$$

$$g_0(x, y) = \frac{40 \exp(8((x - .5)^2 + (y - .5)^2))}{(\exp(8((x - .2)^2 + (y - .7)^2)) + \exp(8((x - .7)^2 + (y - .2)^2)))}$$

with (x, y) iid uniform on $[0, 1]^2$ and u_i distributed as normal, normal scale mixture, or slash.

Monte-Carlo Performance

Design: He and Shi (1996)

$$z_i = g_0(x_i, y_i) + u_i \;\; i = 1, ..., 100.$$

$$g_0(x, y) = \frac{40 \exp(8((x - .5)^2 + (y - .5)^2))}{(\exp(8((x - .2)^2 + (y - .7)^2)) + \exp(8((x - .7)^2 + (y - .2)^2)))}$$

with (x, y) iid uniform on $[0, 1]^2$ and u_i distributed as normal, normal scale mixture, or slash.

Comparison of both L_1 and L_2 triogram and tensor product splines.

Monte-Carlo MISE (1000 Replications)

Distribution	L_1 tensor	L_1 triogram	L_2 tensor	L_2 triogram
Normal	0.609	0.442	0.544	0.3102
	(0.095)	(0.161)	(0.072)	(0.093)
Normal Mixture	0.691	0.515	0.747	0.602
	(0.233)	(0.245)	(0.327)	(0.187)
Slash	0.689	4.79	31.1	171.1
	(6.52)	(125.22)	(18135)	(4723)

Table 1: Comparative mean integrated squared error

Monte-Carlo MISE (1000 Replications)

Distribution	L_1 tensor	L_1 triogram	L_2 tensor	L_2 triogram
Normal	0.609	0.442	0.544	0.3102
	(0.095)	(0.161)	(0.072)	(0.093)
Normal Mixture	0.691	0.515	0.747	0.602
	(0.233)	(0.245)	(0.327)	(0.187)
Slash	0.689	4.79	31.1	171.1
	(6.52)	(125.22)	(18135)	(4723)

Table 2: Comparative mean integrated squared error

Monte-Carlo MISE (998 Replications)

Distribution	L_1 tensor	L_1 triogram	L_2 tensor	L_2 triogram
Normal	0.609	0.442	0.544	0.3102
	(0.095)	(0.161)	(0.072)	(0.093)
Normal Mixture	0.691	0.515	0.747	0.602
	(0.233)	(0.245)	(0.327)	(0.187)
Slash	0.689	0.486	31.1	171.1
	(6.52)	(3.25)	(18135)	(4723)

Table 3: Comparative mean integrated squared error

Dogma of Goniolatry

Dogma of Goniolatry

- Triograms are nice elementary surfaces

Dogma of Goniolatry

- Triograms are nice elementary surfaces
- Roughness penalties are preferable to knot selection

Dogma of Goniolatry

- Triograms are nice elementary surfaces
- Roughness penalties are preferable to knot selection
- Total variation provides a natural roughness penalty

Dogma of Goniolatry

- Triograms are nice elementary surfaces
- Roughness penalties are preferable to knot selection
- Total variation provides a natural roughness penalty
- Schwarz penalty for λ selection based on model dimension

Dogma of Goniolatry

- Triograms are nice elementary surfaces
- Roughness penalties are preferable to knot selection
- Total variation provides a natural roughness penalty
- Schwarz penalty for λ selection based on model dimension
- Sparsity of linear algebra facilitates computability

Dogma of Goniolatry

- Triograms are nice elementary surfaces
- Roughness penalties are preferable to knot selection
- Total variation provides a natural roughness penalty
- Schwarz penalty for λ selection based on model dimension
- Sparsity of linear algebra facilitates computability
- Quantile fidelity yields a family of fitted surfaces

Dogma of Goniolatry

- Triograms are nice elementary surfaces
- Roughness penalties are preferable to knot selection
- Total variation provides a natural roughness penalty
- Schwarz penalty for λ selection based on model dimension
- Sparsity of linear algebra facilitates computability
- Quantile fidelity yields a family of fitted surfaces