Closed-Form Maximum Likelihood Estimates for Spatial Problems

R. Kelley Pace[1]
LREC Endowed Chair of Real Estate
Department of Finance
E.J. Ourso College of Business Administration
Louisiana State University
Baton Rouge, LA 70803-6308
OFF: (225)-388-6256, FAX: (225)-334-1227
kelley@pace.am, www.spatial-statistics.com

James P. LeSage
University of Toledo
Department of Economics
Toledo, OH 43606
jlesage@spatial-econometrics.com

September 3, 2000

[1]The first author gratefully acknowledges research support from Louisiana State University and the University of Connecticut. In addition, the authors would like to thank Ming-Long Lee, Carlos Slawson, and Dongya Zou for their comments. Please Do Not Quote Without Permission

Abstract

This manuscript introduces the matrix exponential as a way of specifying spatial transformations of the data. The matrix exponential spatial specification (MESS) simplifies the log-likelihood, leading to a closed form maximum likelihood solution. The computational advantages of this model make it ideal for applications involving large data sets such as census and real estate data. The manuscript demonstrates the utility of the techniques by estimating a model for housing prices across 57,647 census tracts. Amazingly, the MESS autoregression can take under a second to compute, despite the large sample size.

JEL: C29, R15

KEYWORDS: spatial statistics, spatial autoregression, nearest neighbor, maximum likelihood, sparse matrices, log-determinants, matrix exponentials

1 Introduction

Recent technology has increased the ability to analyze data, but has simultaneously increased the amount of data available for analysis. Spatial data technologies such as global positioning systems (GPS), geographic information systems (GIS), and address geocoding have created an explosion in the size of these data sets. For example, commercial vendors sell data on millions of housing sales across the US with address information that can be easily geocoded using GIS software to produce very large spatial data sets. Analysis of real estate transactions for even a single county may yield more than one hundred thousand annual observations. Not surprisingly, such data or functions of such data (e.g., regression residuals) exhibit a high degree of spatial dependence (e.g., Bell and Bockstael (2000) and Pace and Gilley (1997)). Although spatial location is important when analyzing these data, direct estimation via maximum likelihood of spatial models requires computation of a determinant involving an nxn matrix. Brute force implementations of maximum likelihood methods become prohibitively expensive for these large data sets.

One approach to overcoming these problems was proposed by Kelejian and Prucha (1998,1999) who set forth a generalized-moments (GM) estimation technique. Bell and Bockstael (2000) compare this GM estimation methodology to maximum likelihood methods concluding that GM estimation may represent a low-cost means of obtaining estimates that are comparable to those from maximum likelihood.

However, spatial maximum likelihood may not prove as difficult as initially thought. For the particular case of nearest neighbor spatial dependence Pace and Zou (2000) provide a closed-form solution that produces maximum likelihood estimates and illustrate their technique on samples sizes of up to 500,000 observations. As an alternative approach also leading to closed-form maximum likelihood estimates, this paper adapts the matrix exponential covariance specification introduced by Chiu, Leonard, and Tsui (1996). Specifically, this paper investigates the use of matrix exponentials for spatially transforming the dependent variable. Amazingly, common ways of specifying the spatial transformation ensure the determinant of the matrix exponential transformation identically equals 1, eliminating the log-determinant term from the log-likelihood. Elimination of the log-determinant term reduces maximum likelihood estimation to minimizing a quadratic form subject to a polynomial constraint. Further, this minimization problem has a unique, closed-form interior solution. Thus, maximum likelihood for this specification reduces to a particularly tractable form of

non-linearly constrained least squares.

This approach to spatial estimation which we label the matrix exponential spatial specification (MESS) possesses several outstanding advantages. First, the matrix exponential spatial specification can exhibit an operation count as low as O(n), the same as OLS. Second, the usual diagnostics and other useful tools associated with least squares easily transfer to spatial maximum likelihood estimation. Finally, the availability of the likelihood greatly facilitates both classical and Bayesian inference (see LeSage 1997, 2000 for Bayesian variants of spatial models). Hence, users do not need to adopt another inferential paradigm to overcome computational difficulties arising during analysis of problems involving large samples.

To illustrate the efficacy of these techniques, the MESS is estimated using nationwide housing data from 57,647 census tracts. Any individual MESS autoregression takes under a second to compute. The ensemble of finding the neighbors from the locational coordinates, calculating 203 spatial autoregressions (to estimate hyperparameters), and computing the likelihood ratio tests associated with variable deletions takes under four minutes on a 600 Mhz PC compatible machine.

Section 2 provides the theory underlying spatial estimation with matrix exponentials, section 3 applies the MESS model to US census tract data, and section 4 summarizes the key results.

2 Closed form estimation of spatial dependence using matrix exponentials

This section sets forth a unique interior optimal spatial transformation of the dependent variable. Section 2.1 presents the MESS model based on this spatial transformation, section 2.2 discusses the matrix exponential and its computation in a statistical context and section 2.3 provides the closed form solution that is based on the eigenvalues of a small matrix. Section 2.4 proves that the solution is both interior and unique, while section 2.5 illustrates hypothesis testing for the MESS model and section 2.6 details the construction of diagnostic statistics.

2.1 Model

Consider estimation of models where the dependent variable y undergoes a linear transformation Sy as in (1).

$$Sy = X\beta + \varepsilon \tag{1}$$

The vector y contains the n observations on the dependent variable, X represents the nxk matrix of observations on the independent variables, S is a positive definite nxn matrix, and the n-element vector ε is distributed $N(0, \sigma^2 I_n)$. The log-likelihood for the MESS model in (1) is,

$$L = C + \ln|S| - (n/2)\ln(y'S'MSy) \tag{2}$$

where C represents a scalar constant and both $M = I - H$ and $H = X(X'X)^{-1}X'$ are idempotent matrices. The term $|S|$ is the Jacobian of the transformation from y to Sy. Without the Jacobian term, S containing all zeros would lead to a perfect, albeit pathological, fit. The Jacobian term penalizes attempts to use singular or near singular transformations to artificially increase the regression fit.

We explore the use of the matrix exponential as defined by (3) in modeling S,

$$S = e^{\alpha D} = \sum_{i=0}^{\infty} \frac{\alpha^i D^i}{i!} \tag{3}$$

where D represents an nxn non-negative matrix with zeros on the diagonal and α represents a scalar real parameter. While a number of ways exist to specify D, a common specification sets $D_{ij} > 0$ for observations $j = 1 \ldots n$ sufficiently close (as measured by some metric) to observation i. By construction, $D_{ii} = 0$ to preclude an observation from directly predicting itself. If $D_{ij} > 0$ for the nearest neighbors of observation i, $D^2_{ij} > 0$ contains neighbors to these nearest neighbors for observation i. Similar relations hold for higher powers of D which identify higher-order neighbors. Thus the matrix exponential S, associated with matrix D, can be interpreted as assigning rapidly declining weights for observations involving higher-order neighboring relationships. That is, observations reflecting higher-order neighbors (neighbors of neighbors) receive less weight than lower-order neighbors.

If D is row-stochastic, S will be proportional to a row-stochastic matrix, since products of row-stochastic matrices are row-stochastic (i.e., by definition $D\iota = \iota$ and therefore $D(D\iota) = \iota$, and so on, where ι denotes a vector of ones). The same holds true for any power of S, since the powers are simply linear combinations of the powers of D, all of which are proportional to a row-stochastic matrix. Row-stochastic spatial weight matrices, or multidimensional linear filters, have a long history of application in spatial statistics (e.g., Ord (1975)). The row-stochastic weight matrix has very

favorable numeric as well as statistical properties. For example, the product of a row-stochastic weight matrix D and a random variable vector v produces a vector of spatially local averages, Dv.

Chiu, Leonard, and Tsui (1996) proposed the use of the matrix exponential and discussed several of its salient properties, some of which are enumerated below:

1. S is positive definite,
2. any positive definite matrix is the matrix exponential of some matrix,
3. $S^{-1} = e^{-\alpha D}$,
4. $|e^{\alpha D}| = e^{\text{trace}(\alpha D)}$.

The last property greatly simplifies the MESS log-likelihood. Since $\text{trace}(D) = 0$ and by extension $|e^{\alpha D}| = e^{\text{trace}(\alpha D)} = e^0 = 1$, the log-likelihood takes the form: $L = C - (n/2)\ln(y'S'MSy)$. Therefore, maximizing the log-likelihood is equivalent to minimizing $(y'S'MSy)$, the overall sum-of-squared errors. Thus, one can interpret the search for an optimal S as a search for a coordinate system (possibly oblique) which has the same multidimensional volume as the orthogonal Cartesian coordinate system, but yields a better goodness-of-fit among the variables (smaller $y'S'MSy$).

The MESS model in (1) is a bit more general than it appears. Let U represent a matrix of observations on p non-constant independent variables and let q be an integer large enough so that X approximately spans SU, but small enough so that X cannot span y. The design matrix X (assuming full rank) could have the form (4).

$$X = [\iota \; U \; DU \ldots D^{q-1}U] \tag{4}$$

In this case, X approximately spans SU and thus the MESS model based on (4) nests a spatial autoregression in the errors. Hence, a set of linear restrictions on the parameters associated with the columns of X could yield the error autoregression. The MESS model with X specified as in (4) results in an estimate for α that does not depend upon the variance of the errors, only the direction of these errors (Pace and Barry (1998)). This allows the MESS model to effectively accommodate different structures for the spatial lags of Y and U (Anselin (1988), p. 225-230). Hendry et al. (1984) advocate estimation of this type of general distributed lag model with subsequent imposition of restrictions that has been labeled the general to specific approach to model specification.

2.2 Computational considerations and comparisons

If the magnitude of the elements of the powers of D do not rise with the power, the power series converges rapidly. Note, row-stochastic, non-negative D can have a maximum of 1 in any row. Hence, the magnitude of the elements in the powers of D does not grow with the power. Given the rapid decline in the coefficients in the power series, achieving a satisfactory progression with six or seven terms seems feasible.

Straightforward implementation of the definition of S by a power series expansion truncated after q terms with D^{q-1} as the highest degree term would require $O((q-2)n^3)$ operations for dense D, as matrix multiplication requires $O(n^3)$ operations for dense matrices. This would not be practical and computing the matrix exponential using eigenvalues and eigenvectors would also be impractical for sufficiently large dense matrices, since this requires $O(n^3)$ operations.

If the graph of D is strongly connected, meaning that a path exists between every pair of observations, then $\sum_{r=1}^{n} \omega_i D^r$ will be dense (all non-zeros) for positive ω_i (Horn and Johnson, p. 361-362). Hence, S will be dense. Computing S separately would thus require prohibitive amounts of memory for large n. Fortunately, one does not need to compute S separately, as S always appears in conjunction with y.[1] This allows computation of Sy in $O((q-1)n^2)$ operations for dense D by sequential left-multiplication of y by D to form $n-$element vectors, (i.e., Dy, $D(Dy) = D^2y$, and so on).

For sparse D, that would result from a spatial weight matrix based on nearest neighbors computed using Delaunay triangles, or a sparse covariance structure based on the spherical variogram, the number of operations required to compute Sy declines dramatically.[2] The number of operations required drops to $O((q-1)n_{\neq 0})$, where $n_{\neq 0}$ denotes the number of non-zeros. For the nearest neighbor spatial weight matrix approach with m non-zero entries in each row, the operation count associated with computing Sy would decline to $O((q-1)mn)$. This results in an operation count for computing Sy in nearest neighbor specifications of D that is linear in n.

To illustrate this computation approach in detail, we define the nxq matrix Y comprised of powers of D times y in (5).

[1] Sidje (1998) has also used this as a point of departure in the computation of matrix exponentials. In addition, Sidje provides other algorithms for computing the matrix exponential right multiplied by a vector. Finally, Sidje provides software implementing these algorithms in both Matlab and Fortran 77.

[2] Myers (1997, p. 276) indicates the spherical model is the one most often used in geostatistical practice. See Barry and Pace (1997) for more about the use of sparsity with the spherical model.

$$Y = [y \; Dy \; D^2y \ldots D^{q-1}y] \tag{5}$$

Note that, this simple form of sparse matrix-vector multiplication can be implemented without explicit use of sparse matrix multiplication algorithms. Given a list of m neighbors for observation i (i.e., $(i, j_1), (i, j_2), \ldots, (i, j_m)$) and given equal weights for the neighbors, the product Dy for observation i is merely $(y_{j_1} + y_{j_2} + \ldots + y_{j_m})/m$. That is, the sparse matrix-vector products needed to compute Y in (5) require only indexing and addition, two of the fastest operations possible on digital computers. This means that the MESS model can be implemented in a variety of computing languages such as FORTRAN or C, and various statistical software environments. In fact, to demonstrate the feasibility of estimating MESS using standard software we coded the estimator in Fortran 90 as well as in Matlab. Employing a compiled language (Fortran 90) plus using indexing as opposed to sparse matrices increased the speed by more than 6 times.

We provide a comparison of maximum likelihood estimates and the time required to solve the MESS model and a traditional spatial autoregressive model (SAR) taking the form: $y = \rho Wy + X\beta + \varepsilon$, with a spatial weight matrix based on 30 nearest neighbors making it an asymmetric row-stochastic matrix. These comparative results were based on a dependent variable representing housing prices across 57,647 census tracts along with five explanatory variables described in section 3. This comparison makes the point that the parameter estimates for β in traditional spatial econometric models as well as inferences regarding the magnitude and significance of these parameters can be replicated using the MESS model. In this application the MESS model took the form: $Sy = X\beta + \varepsilon$, where the explanatory variables matrix X was not transformed for comparability with the SAR model.

The estimates from both models are presented in Table 1 along with computed deviances that would be used to draw inferences about the statistical significance of each variable. These deviances reflect the change in the log likelihood arising from sequential deletion of each variable from the model.

From the table we see that the parameter magnitudes are very similar from both models. Since the model is in log form, the coefficients can be interpreted as the percentage change in housing prices that would result from a one percent change in the explanatory variables. The deviances would produce the same inferences regarding the significance or lack thereof for all variables. (In this case, all of the explanatory variables are significant at the 0.01 level.)

Table 1: A Comparison of estimates from SAR and MESS models

Variables	SAR Model†	MESS Model
ln(Land Area)	-0.0152	-0.0188
Deviance	1142.64	1394.59
ln(Population)	0.0189	0.0233
Deviance	132.87	177.70
ln(Per Capita Income)	0.4613	0.4786
Deviance	31229.72	23987.24
ln(Age)	-0.0685	-0.0689
Deviance	1213.15	1097.93
Intercept	-1.6861	-1.3492
Deviance	3464.33	1945.65
Time (in seconds) Matlab	2414.92	3.36
Time (in seconds) FORTRAN 90	—	0.536
n	57,647	57,647
k	5	5

The table also shows a timing comparison for computing estimates using the two methods, where we see a dramatic improvement in speed associated with the MESS model which is over 1,000 times faster than the more traditional SAR model. This speed gain was obtained in Matlab. The Fortran coding accelerated this already fast implementation by over a factor of 6.[3] In section 2.5 we motivate that the computational speed of the MESS model allows quick hypothesis testing of alternative specifications for the functional form taken by the spatial weight matrix, as well as more traditional specification tests used in regression models. These testing methods are illustrated in an application presented in section 3.

2.3 Closed form solution of the estimated parameters

We define the diagonal matrix W containing part of the coefficients of the power series as shown in (6).

[3]The use of an asymmetric 30 neighbor spatial weight matrix poses a substantial computational challenge to computing the log-determinant term used in maximum likelihood. The times reported in Table 1 could be improved to around 400 seconds by using symmetric spatial weight matrix. The times would improve further if fewer neighbors were used or if the Monte Carlo Log-determinant estimator proposed by Barry and Pace (1999) were employed to compute the SAR estimates.

$$W = \begin{pmatrix} 1/0! & & & \\ & 1/1! & & \\ & & \ddots & \\ & & & 1/(q-1)! \end{pmatrix} \tag{6}$$

In addition, we define the q-element column vector v shown in (7) that contains powers of the scalar real parameter α, $|\alpha| < \infty$.

$$v = [1 \ \alpha \ \alpha^2 \dots \alpha^{q-1}]' \tag{7}$$

Using (5), (6) and (7), we can rewrite Sy as shown in (8).

$$Sy = YWv \tag{8}$$

Premultiplying Sy by the least-squares idempotent matrix M yields the residuals e, allowing us to express the overall sum-of-squared errors as in (9),

$$e'e = v'W(Y'M'MY)Wv = v'W(Y'MY)Wv = v'Qv \tag{9}$$

where $Q = W(Y'MY)W$. The matrix MY represents residuals from regressing the dependent variable and the spatial lags of the dependent variable on the independent variables X. Multiplying MY by a vector ϕ results in a linear combination of these residuals, $MY\phi$. Hence, the sum-of-squared errors associated with this vector equals $(MY\phi)'(MY\phi) = \phi'(Y'MY)\phi$. If a linear combination of the residuals, $MY\phi$ produces a zero vector (columns of MY are not linearly independent), then $\phi'(Y'MY)\phi = 0$ and $Y'MY$ is positive semidefinite in this case, since the product cannot be negative. This seems unlikely to arise in practice, so we assume the regression residuals MY are linearly independent so that MY is nonsingular. In this case, $\phi'(Y'MY)\phi > 0$ and $(Y'MY)$ is positive definite. Given this, both W and $(Y'MY)$ are symmetric positive definite matrices, so Q must be congruent to $(Y'MY)$ and have the same number of positive eigenvalues as $(Y'MY)$ by Sylvester's law of inertia (Strang(1976, p. 246)). Since $(Y'MY)$ is a symmetric positive definite matrix, Q will have all positive eigenvalues and must be a symmetric positive definite matrix (Horn and Johnson (1993), p. 402).

The overall sum-of-squared errors $v'Qv$ is a $2q-2$ degree polynomial in the variable α. The coefficients in the polynomial are the sum of all terms appearing in Q associated with each power of α. The number of coefficients of a $2q-2$ degree polynomial equals $2q-1$ due to the constant

term (coefficient associated with the degree 0). Specifically, the coefficients c, a $2q-1$ element column vector are shown in (10),

$$c_{t-1} = \sum_{i=1}^{q} \sum_{j=1}^{q} Q_{ij} Ind((i+j)=t) \tag{10}$$

where $Ind()$ is an indicator function taking on values of 1 when the condition is true. The terms associated with the same power of α have subscripts i, j that sum to the same value. For example, $\alpha^i \alpha^j = \alpha^t$ when $i+j=t$, which means that each coefficient c_i is the sum of the elements along the antidiagonals of Q. This allows us to rewrite $v'Qv$ as the $2q-2$ degree polynomial $P(\alpha)$, shown in (11).

$$P(\alpha) = \sum_{i=1}^{2q-1} c_i \alpha^{i-1} = v'Qv \tag{11}$$

To find the minimum of the sum-of-squared errors, we differentiate the polynomial $P(\alpha)$ in (11) with respect to α, equate to zero, and solve for α as shown in (12).

$$\frac{dP(\alpha)}{d\alpha} = \sum_{i=2}^{2q-1} c_i (i-1) \alpha^{i-2} = 2v'Q\left(\frac{dv}{d\alpha}\right) = 0 \tag{12}$$

The derivative $dP(\alpha)/d\alpha$ is a degree $2q-3$ polynomial and thus has $2q-3$ possible roots. The problem of finding all the roots of a polynomial has a well-defined solution. Specifically, the roots equal the eigenvalues of the companion matrix associated with the polynomial (Horn and Johnson (1993, p. 146-147)).[4] Computation of the eigenvalues requires $O(8q^3)$ operations in this case and does not depend upon n. Thus, the maximum likelihood estimates have a closed-form solution in terms of the eigenvalues of a small matrix.

2.4 Uniqueness of the solution

To narrow the possible number of solutions, we turn to the second order conditions. Positive definite Q would usually prove sufficient for an interior solution, but the vector v embodies a polynomial constraint. Therefore, we elaborate on the second order conditions taking into account the constraints

[4]Other methods also exist for finding the roots of polynomials. See Press et al. (1996, p. 362-372) for a review of these.

imposed by the structure of v. Consider the second derivative of the sum-of-squared errors with respect to α shown in (13).

$$\frac{d^2v'Qv}{d\alpha^2} = \sum_{i=3}^{2q-1} c_i(i-1)(i-2)\alpha^{i-3} = 2\left[(\frac{dv}{d\alpha})'Q(\frac{dv}{d\alpha}) + v'Q(\frac{d^2v}{d\alpha^2})\right] \quad (13)$$

The first term inside the brackets is positive because it represents a positive definite quadratic form. We can rewrite the second term in brackets as shown in (14),

$$v'Q\left(\frac{d^2v}{d\alpha^2}\right) = v'(QA)v \quad (14)$$

where A equals,

$$A = \left(\frac{1}{\alpha^2}\right)\begin{pmatrix} 0 & & & & \\ & 0 & & & \\ & & (i-1)(i-2) & & \\ & & & \ddots & \\ & & & & (q-1)(q-2) \end{pmatrix}. \quad (15)$$

The minimum value of α depends on the eigenvalues of $v'(QA)v$. Note that Q is positive definite and the real diagonal (and thus Hermitian) matrix A in (15) has two zero and $(q-2)$ positive eigenvalues. Horn and Johnson (1993, p. 465) state in Theorem 7.6.3 that the product of a positive definite matrix Q and a Hermitian matrix A has the same number of zero, positive, and negative eigenvalues as A. Hence, QA must have two zero and $(q-2)$ positive eigenvalues. Therefore, $v'(QA)v$ is positive semidefinite implying that $v'QAv$ has a minimum value of 0. Since the first term in brackets in (13) always has a positive value, the entire expression in (13) has a positive value and thus $v'Qv$ is positive definite and strictly convex in α. Hence, if an interior solution exists to the first order conditions, it must be unique.

There exists an interior solution to the first order conditions. To see this, examine the highest degree term in $P(\alpha)$, from (11) shown in (16).

$$T_{\max} = \frac{\alpha^{2(q-1)}[y'D'^{(q-1)}MD^{(q-1)}y]}{(q!)^2} \quad (16)$$

The term in brackets is the contribution to the overall sum-of-squared errors from the last term in the truncated Taylor's series and must be positive. Since $\alpha^{2(q-1)}$ is even in α, only the magnitude and not the sign of α matter

for this result. Since $\lim_{|\alpha|\to\infty} T_{\max} \to \infty$, implies $\lim_{|\alpha|\to\infty} v'Qv \to \infty$, there exists an interior solution to the first order conditions.

In conclusion, there exists a unique interior real α, say $\alpha^\star$, that minimizes the sum-of-squared errors and maximizes the MESS likelihood. Such unique optima are rare in spatial statistics. See Warnes and Ripley (1987) and Mardia and Watkins (1989)) for a discussion of the potential multimodality of the likelihood. The MESS unique closed-form optimum solution not only reduces computational time, but also increases the confidence users have in the numerical quality of the estimates.

2.5 Quick hypothesis testing

Efficient computation of likelihood ratio tests requires updating the sum-of-squared errors matrix Q without recomputing the actual regressions. Let $\hat{B} = (X'X)^{-1}X'Y$ denote the k by q matrix of estimates from the regression of Y on X and let $\hat{E} = Y - X\hat{B}$ denote the n by q matrix of errors from the regression. Expression (17) shows the restricted least squares estimate for $\tilde{B}_j, (j = 1 \dots q)$,

$$\tilde{B}_j = \hat{B}_j + (X'X)^{-1}R'(R(X'X)^{-1}R')^{-1}(r - R\hat{B}_j) \tag{17}$$

where r is a h by 1 vector, R denotes a h by k matrix, and h is the number of hypotheses imposed.[5]

Let ΔB_j in (18) denote the change in the restricted least squares estimates versus the unrestricted estimates for the jth regression.

$$\Delta B_j = \tilde{B}_j - \hat{B}_j = (X'X)^{-1}R'(R(X'X)^{-1}R')^{-1}(r - R\hat{B}_j) \tag{18}$$

The inner product of any two vectors of restricted regression errors appears in (19).

$$\tilde{E}'_{j1}\tilde{E}_{j2} = (Y - X\tilde{B}_{j1})'(Y - X\tilde{B}_{j2}) = (\hat{E}_{j1} - X\Delta B_{j1})'(\hat{E}_{j2} - X\Delta B_{j2}) \tag{19}$$

where $\tilde{E}_{j1}, \tilde{E}_{j2}$ represent the vectors of restricted regression errors and $j1, j2 = 1, \dots, q$. Expanding (19) yields (20).

$$\tilde{E}'_{j1}\tilde{E}_{j2} = \hat{E}'_{j1}\hat{E}_{j2} + (\Delta B_{j1})'(X'X)(\Delta B_{j2}) \tag{20}$$

[5]See Gentle (1998, p. 166) for the standard restricted least squares estimator as well as some other techniques for computing these estimates.

where two of the possible terms vanish due to the enforced orthogonality between the residuals and the data in least-squares. One can further expand the second term $(\Delta B_{j1})'(X'X)(\Delta B_{j2})$ from (20).

$$\begin{aligned}(\Delta B_{j1})'(X'X)(\Delta B_{j2}) &= [(r - R\hat{B}_{j1})'(R(X'X)^{-1}R')^{-1}R(X'X)^{-1}] \\ &\cdot \quad (X'X)[(X'X)^{-1}R'(R(X'X)^{-1}R')^{-1}(r - R\hat{B}_{j2})]\end{aligned}$$

Fortunately, many terms in the above expression cancel which leaves a simple expression in (21) for the increase in error arising from restrictions.

$$(\Delta B_{j1})'(X'X)(\Delta B_{j2}) = (r - R\hat{B}_{j1})'[R(X'X)^{-1}R']^{-1}(r - R\hat{B}_{j2}) \qquad (21)$$

Finally, define the q by q matrix of cross-products of restricted least-squares regressions as $\tilde{E}'\tilde{E}$ with $j1, j2$th element $\tilde{E}'_{j1}\tilde{E}_{j2}$ and therefore the restricted sum of squared errors, $Q_R = W(\tilde{E}'\tilde{E})W$.

Computation of the unrestricted regressions means the quantities $\hat{B}_j$ and the Cholesky factors (even if computed from the QR algorithm) of $X'X$ are already known. However, $[R(X'X)^{-1}R']^{-1}$ requires $O(h^3)$ operations for its decomposition. Typically, h will be small. Testing for the effect of the deletion of a single variable means h equals 1 and for a variable and its associated lags h equals 1 plus the number of independent variable lag terms. Since computing the increase in errors from the restrictions requires $O(h^3)$ operations and resolving the first order conditions requires $O(8q^3)$ operations, deviance (i.e., likelihood ratio) tests do not depend upon n and thus require very little time.

One advantage of the likelihood-based MESS methodology noted in the introduction is the ability to accommodate Bayesian extensions. An interesting point here is that Bayesian logic emphasizes the fact that the significance level should be a decreasing function of sample size. As the sample size grows, the Bayes factor region of rejection is a function of sample size, in contrast to the usual classical region which is held constant and independent of sample size. This distinction may be important for very large models of the type discussed here. The Bayes factor in favor of hypothesis i relative to hypothesis j is simply the ratio of the marginal likelihoods $f_i(Y)/f_j(Y)$, where computation of the marginal likelihood requires $f_i(Y) = \int f_i(Y|B_i)f(B_i)dB_i$, where $f(B_i)$ denotes the prior distribution. Leamer (1978,1983) discusses these issues.

This suggests that efficient computation of the likelihood made possible by the matrix exponential specification may enable Bayesian approaches to

hypothesis testing that allow the level of significance to vary with the sample size. Since the rejection region shrinks as the sample size grows, Bayesian testing procedures should lead to more parsimonious specifications in cases involving large sample sizes.

2.6 Spatial diagnostics

Christensen, Johnson, and Pearson (1992), Haining (1994), and Martin (1992) have investigated various aspects of diagnostics for spatial models. For the models which employ some form of estimated variance-covariance matrix $\Omega(\theta)$ parameterized by a vector of parameters θ, estimating the model via maximum likelihood equates to estimating a model where both sides have been transformed (i.e., $\Omega^{-1/2}(\theta)y = \Omega^{-1/2}(\theta)X\beta + \varepsilon$). If the desire is to find leverage points with the transformed independent variables, this becomes difficult since the transformation depends upon y. Fortunately, the right hand side of the MESS model in (1) does not involve parameters in the formulation of X, so one can employ standard leverage statistics (avoiding one of the problems Martin (1992) examined). Since minimizing the overall sum-of-squared errors maximizes the likelihood, one can easily modify some of the useful diagnostics commonly employed in regression. Let $Z = MYW$ (hence $Q = Z'Z$), let $z_i = (Z'_{ij}, j = 1, \ldots, q)$, a q by 1 vector, and let $Q_{(i)}$ represent Q when deleting the ith observation. Applying the standard regression results for one-out sum-of-squared errors (e.g., Christensen (1996, p. 345)), produces (22).

$$Q_{(i)} = Q - \frac{z_i z'_i}{(1 - H_{ii})} \tag{22}$$

Using the same quick mechanism for finding the roots of polynomials leads to a sequence of one-out autoregressive parameters $\alpha_{(i)}$ and one-out deviances. As well-known, each of these would correspond to the deviance associated with including a variable with 1 in the ith row and zeros in all other rows (Christensen (1996, p. 348)).

Note, the case deletion diagnostics just consider the direct effect of a particular observation and do not account for their role in the transformation of the other observations. Naturally, one could delete blocks of observations and remove the observation itself, the observations it neighbors, and so forth (e.g. Christensen (1996, p. 348)).

3 An application to US census tracts

This section applies the MESS model proposed in section 2 to census tract housing prices for the continental US. Section 3.1 discusses the construction of the dataset from the US Census data, section 3.2 describes the MESS model specification and details of the spatial transformation, section 3.3 discusses the MESS estimates and inference, and section 3.4 presents diagnostic information.

3.1 Data

The deciennial census provides a comprehensive set of data on demographic and economic conditions across a wide array of geographical units. One of the most disaggregated and useful geographical units is the census tract, of which 60,804 exist within the Continental US in the 1990 Census. Unfortunately, not every tract provides every data field due to privacy and other constraints. Restricting the final data set to observations with complete data resulted in 57,647 observations on the median price of housing (Price), median per capita income (Income), median year built (Age), population (Pop), the tract's land area (Area), as well as the latitude and longitude of the centroid of the tract. The constructed variable Age equals 1990 less the median year constructed and was strictly positive.

3.2 Model

The overall transformed MESS model of housing prices appears in (23).

$$\begin{aligned} S(\alpha)\ln(\text{Price}) &= \beta_1 + \beta_2\ln(\text{Area}) + \beta_3\ln(\text{Pop}) + \beta_4\ln(\text{Income}) + \beta_5\ln(\text{Age}) \\ &+ \beta_6 D\ln(\text{Area}) + \beta_7 D\ln(\text{Pop}) + \beta_8 D\ln(\text{Income}) \\ &+ \beta_9 D\ln(\text{Age}) + \varepsilon \end{aligned} \tag{23}$$

where ln denotes logarithm and D is a row-stochastic spatial weight matrix.

Construction of D first requires finding the nearest neighbors for each observation. One can use several algorithms for this, but all require at least $O(n \cdot \ln(n))$ operations for points on a plane (Eppstein, Paterson, and Yao (1997)). A Delaunay triangle based method was used here to compute the m nearest neighbors.

A set of individual neighbor matrices $D_1, D_2, \ldots, D_m$, was formed where D_1 represents the closest previously sold neighbor (shortest distance), D_2 represents the second previously sold neighbor (second shortest distance)

and so on. These very sparse matrices have a 1 in each row and contain zeros elsewhere.

The overall spatial matrix D was constructed based on the individual neighbor matrices D_i using (24).

$$D = \frac{\sum_{i=1}^{m} \rho^i D_i}{\sum_{i=1}^{m} \rho^i} \tag{24}$$

In (24), ρ^i weights the relative effect of the ith individual neighbor matrix, so that S depends on the parameters ρ as well as m in both its construction and the metric used. Thus (24) imposes an autoregressive distributed lag structure on the spatial variables. By construction, each row in D sums to 1 and has zeros on the diagonal.

The use of the individual neighbor matrices greatly speeds up investigation of the sensitivity of the results to different forms of D. Constructing the individual neighbor matrices requires some computational expense, with the set of 30 used here taking 96.7 seconds computational time. However, reweighting the individual matrices using (24) requires very little time.

3.3 Estimated parameters and deviances

Relative to a simple aspatial model of housing prices (i.e., $\alpha = \beta_{6-9} = 0$), the unrestricted log-likelihood rises from -266,505.2 to -228,850.4, a deviance of 75,309.6. Relative to a model with spatial independent variables, but no spatial transformation of the dependent variable (i.e., $\alpha = 0$), the deviance is 64,450.6. Controlling for some of the spatial dependence reduces the deviances associated with deletion of the independent variables and their spatial lags for three out of the four basic variables (i.e., $\beta_2 = \beta_6 = 0, \beta_3 = \beta_7 = 0, \beta_4 = \beta_8 = 0$) relative to the corresponding deletions of the independent variables in the aspatial model (i.e., $\beta_2 = \beta_3 = 0, \beta_4 = 0$). For the age variable, the deviance associated with deletion (i.e., $\beta_5 = \beta_9 = 0$) actually rises relative to the aspatial model (i.e., $\beta_5 = 0$).

The interpretation and inferences based on estimated parameters from the aspatial and MESS models are quite different. For example, estimates from the aspatial model suggest that increasing income of a tract by 1% would lead to a 1.08% increase in median housing prices in the tract. Implicitly, the aspatial model allows for individuals with higher incomes to both purchase a larger house and to locate in a different neighborhood. For the MESS model where the characteristics of neighboring tracts is held constant, increasing the income of a tract by 1% would lead to only a 0.68% increase in median housing prices in the tract relative to median prices in

surrounding tracts. This result is consistent with the economic construct of externalities (e.g., Bogart (1998, p. 216-218)).

The optimal number of neighbors was 30 ($m = 30$) and the optimal rate of geometric decline with order was 0.9 ($\rho = 0.9$), and α was -1.67. All of the deviances associated with perturbations in these choices were highly significant.

It took slightly under 16 minutes on a 600 megahertz Pentium III computer running Matlab 5.3 to estimate the MESS model for each of the 203 cases defined by the grid over values of $m = 2 \ldots 30$ and $\rho = 0.25$, 0.5, 0.75, 0.85, 0.90, 0.95, 1.0, implying that each regression took under 5 seconds. The Fortran 90 routines took only 132 seconds to perform the same computations. The speed gain from using Fortran 90 makes searches over these parameters rather easy. Figure 1 shows the profile log-likelihoods across these combinations of m and ρ (scaled by subtracting the maximum log-likelihood). Quick computation of the estimator allowed for the optimization of the transformation with respect to three parameters (m, ρ, α) despite the large size of the problem. Hopefully, such flexibility will approximate the true transformation.

3.4 Spatial diagnostics

As discussed earlier, the MESS model can employ standard leverage statistics, unlike many spatial models. Figure 2 identifies tracts that have the lowest and highest (one-half percentile) degrees of leverage on a map of the US. Note the clustering of high leverage points along the southeast coast of Florida, the coast of California, and in scattered interior points in the western US. Conversely, low leverage points seem concentrated in Pennsylvania, North Carolina, and in some of the major cities.

Truly random outliers should reduce the degree of estimated spatial dependence, so case deletion estimates of the autoregressive parameter may provide interesting results. Examination of the one-out autoregressive estimates, $\alpha_{(i)}$, shows some striking features. For example, the range of is extremely small ($\min(\alpha_{(i)})$=-1.6744, $\max(\alpha_{(i)})$=-1.6726).

If the estimated spatial dependencies showed a spatial pattern, this might suggest possible ways of improving the spatial dependencies component of the model. Figure 3 illustrates the lowest and highest percentile of the one-out autoregressive parameters plotted against latitude and longitude. The smallest $\alpha_{(i)}$ indicate where the degree of spatial dependence increased ($\alpha_{(i)}$ became more negative) upon deletion of the ith observation. These observations exhibit weaker spatial dependence with their neighbors than

the typical observation. From the figure, we see some regular patterns. For example, the southeast coast of Florida has a number of tracts whose deletion increases the degree of estimated spatial dependence. This may indicate the need for a variable measuring contiguity with the ocean.

In contrast, the largest $\alpha_{(i)}$ indicate that the degree of spatial dependence decreased ($\alpha_{(i)}$ became more positive) upon deletion of the ith observation. For these observations we see stronger spatial dependence with neighbors than that associated with the typical observation. Almost all of these large $\alpha_{(i)}$ values occur in urban areas. It required less than 5 minutes to compute the MESS spatial diagnostics presented here, despite the large number of observations used in our example.

4 Conclusion

Maximum likelihood estimation based on the matrix exponential spatial specification (MESS) introduced here was shown to be computationally superior to most spatial estimators, requiring O(n) operations (the same as OLS) conditional upon formation of the spatial weight matrix and as low as O($n$log(n)) for the formation of the spatial weight matrix. When used in conjunction with common approaches to specifying spatial influences, the MESS results in a situation where the log-determinant term in the spatial likelihood function vanishes. The matrix exponential spatial specification provides an unusual situation (in spatial problems) where non-linear least-squares and maximum likelihood methods yield the same estimates. A simplification of the log-likelihood stemming from use of the matrix exponential spatial specification produces a situation where a unique closed-form solution for the estimates exists. This unique closed form solution greatly accelerates computation.

As an illustration of the speed gained through the use of these techniques, it took only 3.36 seconds using Matlab to compute a spatial autoregression involving 57,647 observations. We also demonstrated that the estimates for the parameters β from the MESS model were almost identical to those from a spatial autoregressive (SAR) model and the inferences were identical, while the MESS model's computational speed was over 1000 times faster than the more traditional SAR model. *Afortiori*, the Fortran code ran 6 to 7 times faster than the Matlab code for the 57,647 observation census dataset. In other experiments (not reported here) we found that the MESS model when specified with spatially lagged explanatory variables produces estimates and inferences similar to those from the spatial Durbin model introduced in

Anselin (1988).

This speed, along with the simpler MESS log-likelihood, facilitates maximization over a host of spatial parameter settings that can be used to vary the nature and extent of spatial influences in the model. The application by Bell and Bockstael (2000) provides a compelling motivation for this type of exploration. It may also enable Bayesian model selection criterion to be used in place of traditional likelihood ratio tests which would allow the rejection regions to vary with the sample size. This may have the potential to produce more parsimonious global model specifications because the rejection regions would narrow with larger sample sizes. Recent literature in the area of 'Bayesian model averaging' suggests that another potential role for Bayesian methods may be to produce a single posterior model that averages over alternative specifications associated with alternative spatial parameter settings. This would greatly facilitate reporting of results that are not conditional on a particular setting for decay in spatial weights, or number of neighbors employed. A final point is that the simpler MESS log-likelihood may make it an easier model to use in theoretical derivations needed to produce Bayesian and other spatial econometric extensions.

The computational advantages of the MESS should prove useful in solving a number of problems that arise in application of spatial econometric analysis. First, the MESS should provide an easily calculated benchmark against which to gauge the performance of other spatial estimators. In other words, MESS can serve as a more sophisticated null hypothesis than the typical assumption of spatial independence. In addition, since MESS provides a unique optimal estimate, it could help identify when another more complex model has become trapped in a local optima.

Second, the computational efficiency for large problems means that the MESS model can serve as a global description for very large data sets. Such global descriptions can help identify smaller regions where it may be of interest to apply more computationally costly techniques for analysis. Policy decisions often require global descriptions, so a collection of regional descriptions based on smaller subsets of the data set may not serve the desired purpose. This could become particularly important with the pending release of the year 2000 Census that will contain nearly 250,000 observations at the block-group microlevel.

A third area of applications opened up by this approach is computation of diagnostic statistics that have traditionally been problematical in the maximum likelihood spatial estimation setting. We provided a brief demonstration of the application of these statistics, but the potential of these diagnostics in large sample problems represents a relatively unexplored area

for future research.

Another area where the MESS method could be useful is Monte Carlo experiments. Research examining the performance characteristics of alternative spatial estimation methodologies has been limited to relatively small data sets because of the computational burdens. Since both simulation and estimation proceed rapidly in the case of the MESS, this should facilitate Monte Carlo experiments based on larger, more realistic data sets.

References

Anselin, L. 1988. *Spatial Econometrics: Methods and Models*, (Dorddrecht: Kluwer Academic Publishers).

Barry, Ronald, and R. Kelley Pace, "Kriging with Large Data Sets Using Sparse Matrix Techniques," *Communications in Statistics: Computation and Simulation*, Volume 26, Number 2, 1997, p. 619-629.

Barry, Ronald, and R. Kelley Pace, "A Monte Carlo Estimator of the Log Determinant of Large Sparse Matrices," *Linear Algebra and its Applications*, Volume 289, Number 1-3, 1999, p. 41-54.

Bell, Kathleen P., and Nancy E. Bockstael, "Applying the Generalized-Moments Estimation Approach to Spatial Problems Involving Microlevel Data", *Review of Economics and Statistics*, Volume 87, Number 1, 2000, p. 72-82.

Bogart, William T., *The Economics of Cities and Suburbs*, Upper Saddle River: Prentice Hall, 1998.

Chiu, Tom Y.M., Tom Leonard, and Kam-Wah Tsui, "The Matrix-Logarithmic Covariance Model," *Journal of the American Statistical Association*, 91, 1996, p. 198-210.

Christensen, Ronald, *Plane Answers to Complex Questions*, Second Edition, New York: Springer-Verlag, 1996.

Christensen, Ronald, Wesley Johnson, and Larry Pearson, "Prediction Diagnostics for Spatial Linear Models," *Biometrika*, Volume 79, 1992, p. 583-591.

Eppstein, D., M.S. Paterson, and F.F. Yao, "On Nearest-Neighbor Graphs," *Discrete and Computational Geometry*, Volume 17, 1997, p. 263-282.

Gentle, James, *Numerical Linear Algebra for Applications in Statistics*, New York: Springer-Verlag, 1998.

Haining, Robert, "Diagnostics for Regression Modeling in Spatial Econometrics," *Journal of Regional Science*, Volume 34, 1994, p. 325-341.

Hendry, David, Adrian Pagan, and Denis Sargan, "Dynamic Specification," in: Griliches, Z., and M. Intrilligator, eds. *Handbook of*

Econometrics, Volume 2, Amsterdam: North-Holland, 1984, p. 1023-1100.

Horn, Roger, and Charles Johnson, *Matrix Analysis*, New York: Cambridge University Press, 1993.

Kelejian, H., and I.R. Prucha, "A Generalized Spatial Two-Stage Least Squares Procedure for Estimating a Spatial Autoregressive Model with Autoregressive Disturbances", *Journal of Real Estate and Finance Economics*, Volume 17, Number 1 1998, p. 99-121.

Kelejian, H., and I.R. Prucha, "A Generalized Moments Estimator for the Autoregressive Parameter in a Spatial Model", *International Economic Review*, Volume 40, 1999, p. 509-533.

Lay, David, *Linear Algebra and its Applications*, second edition, New York: Addison Wesley Longman, 1997.

Leamer, Edward E., "Regression-Selection Strategies and Revealed Priors," *Journal of the American Statistical Association,* Volume 73, 1978, p. 507-510.

Leamer, Edward E., "Model Choice", in Z. Grilliches and M.D. Intrilligator (eds.) *Handbook of Econometrics, Volume 1* Amsterdam: North-Holland, 1983, p. 286-330.

LeSage, James P., "Bayesian Estimation of Spatial Autoregressive Models", *International Regional Science Review*, Volume 20, number 1&2, 1997, p. 113-129.

LeSage, James P., "Bayesian Estimation of Limited Dependent variable Spatial Autoregressive Models", *Geographical Analysis*, Volume 32, number 1, 2000, p. 19-35.

Mardia, K.V., and A.J. Watkins, "On Multimodality of the Likelihood in the Spatial Linear Model," *Biometrika*, Volume 76, 1989, p. 289-295.

Martin, Richard J., "Leverage, Influence, and Residuals in Regression Models when Observations are Correlated," *Communications in Statistics: Theory and Methods*, Volume 21, 1992, p. 1183-1212.

Myers, Jeffrey, *Geostatistical Error Management*, New York: Van Nostrand Reinhold, 1997.

Ord, J.K., "Estimation Methods for Models of Spatial Interaction," *Journal of the American Statistical Association*, Volume 70, 1975, p. 120-126.

Pace, R. Kelley, and Ronald Barry, "Simulating Mixed Regressive Spatially Autoregressive Estimators," *Computational Statistics*, Volume 13, Number 3, 1998, p. 397-418.

Pace, R. Kelley, and Dongya Zou, "Closed-Form Maximum Likelihood Estimates of Nearest Neighbor Spatial Dependence," *Geographical Analysis*, Volume 32, Number 2, April 2000.

Press, William, Saul Teukolsky, William Vetterling, and Brian Flannery, *Numerical Recipes in Fortran 77*, second edition, New York: Cambridge University Press, 1996.

Ripley, Brian D., *Statistical Inference for Spatial Processes*, Cambridge: Cambridge University Press, 1988.

Sidje, Roger B., "Expokit: a Software Package for Computing Matrix Exponentials," *ACM Transactions on Mathematical Software*, Volume 24, 1998, p. 130-156.

Strang, Gilbert, *Linear Algebra and its Applications*, New York: Academic Press, 1976.

Warnes, J.J., and Brian Ripley, "Problems with Likelihood Estimation of Covariance Functions of Spatial Gaussian Processes," *Biometrika*, Volume 74, 1987, p. 640-642.

Table 2: Spatial and Aspatial Regression Models

Variables	Aspatial Model	Spatial Model
Intercept	1.224	-0.151
ln(Land Area)	-0.085	-0.003
Dln(Land Area)		-0.017
Deviance	9,379.1	1,218.8
ln(Population)	0.115	0.022
Dln(Population)		0.030
Deviance	1,358.6	366.6
ln(Per Capita Income)	1.084	0.677
Dln(Per Capita Income)		-0.463
Deviance	43,355.2	29,764.3
ln(Age)	-0.127	-0.138
Dln(Age)		0.127
Deviance	1,175.2	2,289.5
m (# of neighbors)		30
Deviance (m=29)		2.72
ρ (geometric decay)		0.90
Deviance (ρ=0.95)		612.82
Deviance (ρ=0.85)		1240.52
α (autoregressive parameter)		-1.673
Deviance ($\alpha = 0$)		64,450.6
n	57,647	57,647
k	5	12
Maximum Log-likelihood	-266,505.2	-228,850.4

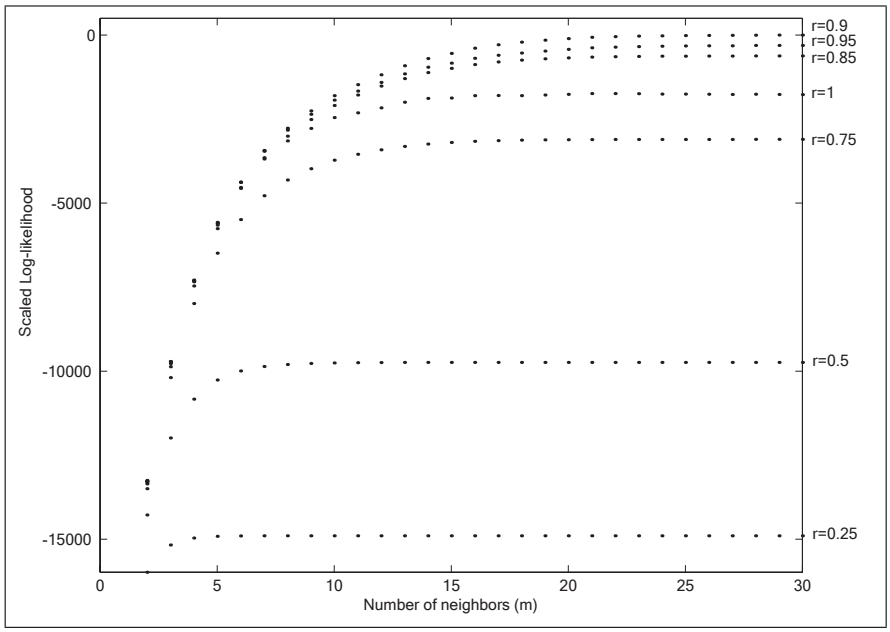

Figure 1: Scaled Log-likelihood vs. Number of Neighbors across Differing ρ

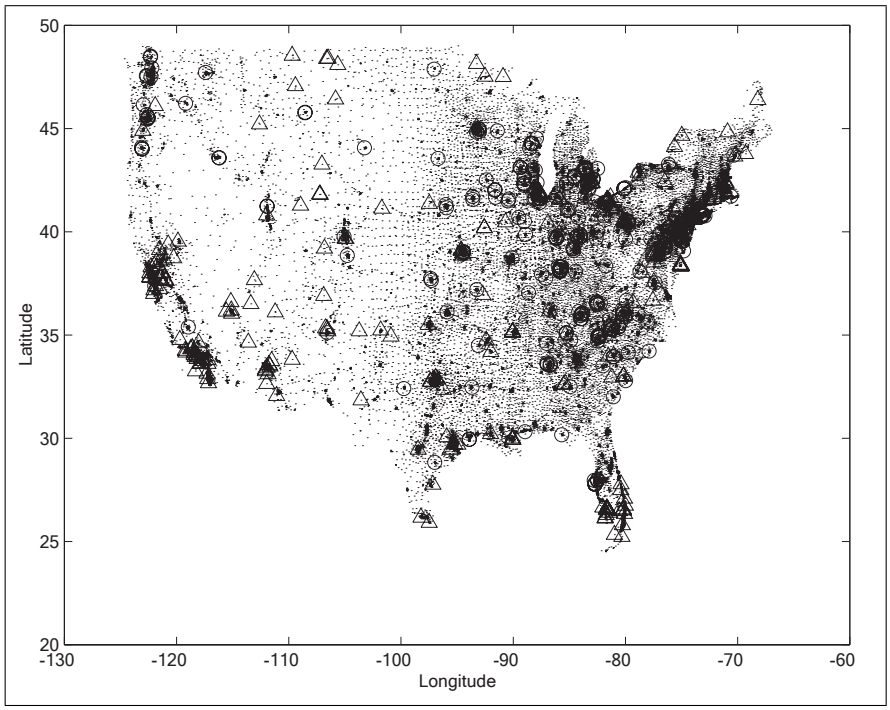

Figure 2: US Census Tract Locations with Smallest (O) and Largest (Δ) Leverage Observations Identified

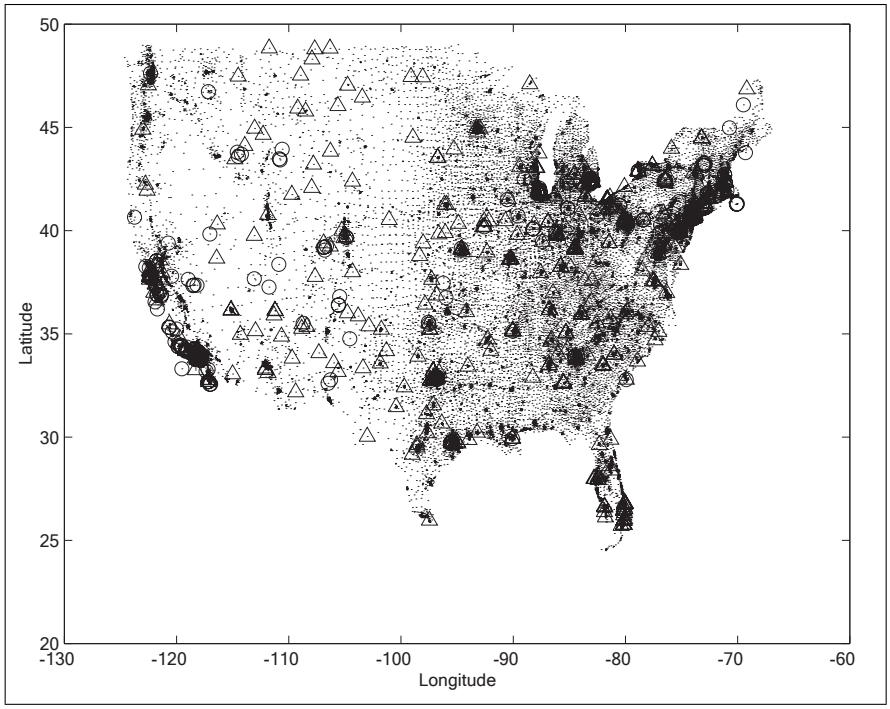

Figure 3: US Census Tract Locations with Largest (O) and Smallest (Δ) Delete-1 α Identified