BAYESIAN ART VERSUS CONJUGATE GRADIENTf METHODS IN TOMOGRAPHIC SEISMIC IMAGING: AN APPLICATION AT MOUNT ST. HELENS, WASHINGTON

J. M. Lees and R. S. Crosson
Geophysics Program
University of Washington
Seattle, WA 98195

ABSTRACT

We compare two approaches for solving the large, sparse linear systems that arise in tomographic velocity inversion problems. When noise is present in the data, the system typically is inconsistent and quasi-overdetermined, and some form of regularization must be implemented to avoid the strong, undesired influence of small singular values dominating the solutions. First, a Bayesian ART (Algebraic Reconstruction Technique) algorithm is applied to the system where we solve for a set of model parameters and residuals simultaneously. Careful choice of relaxation parameters and smoothing filters insures convergence and acceptable results. This approach avoids the undesirable effects of implicit row weighting inherent in simple ART or SIRT (Simultaneous Iterative Reconstruction Technique) applications. Second, a conjugate gradient approach is implemented via algorithm LSQR where regularization is achieved by augmenting the system with additional constraint equations which minimize the roughness of the model. Specifically, applied to a 3-dimensional tomographic inversion we constrain the second derivative (the Laplacian) to be zero within horizontal layers. Comparison on synthetic data reveals that these techniques produce nearly equivalent results. By applying these methods to local earthquake data in the vicinity of Mount St. Helens, Washington, we have produced a 3-dimensional, laterally varying velocity structure in the top 40 km of the crust which correlates well with known geological and geophysical features and delineates possible accumulation of magma beneath the crater.

Introduction

Seismic tomography uses the travel time of elastic waves to probe the internal structure of the earth. It differs from traditional medical tomography in four major aspects: 1) acoustic signals travel in highly curved raypaths in media that vary in 3-dimensions, 2) the travel time is a non-linear function of the velocity field ("velocity" field in the seismic sense is the scalar wave speed), 3) when the sources are earthquakes, the distribution of rays covering the target cannot be controlled and is often highly inhomogeneous and 4) uncertainties in the travel time exist because the source location and origin time must be determined from the observations themselves. These differences indicate that special care must be taken when techniques borrowed from the medical field are applied to seismic data. Specifically, due to the non-uniform distribution of sources and receivers, the convolutional techniques of inversion, common in medical tomography, are inapplicable in the seismic case and iterative approaches are used instead. These are usually grouped in two camps: the algebraic techniques, typified by ART (Algebraic Reconstruction Technique) and its variants, and the projection methods, a name used for conjugate gradients and its variants. In this paper we introduce the usual approach to linearization used in seismic tomography and then we compare, with synthetic and real data, a specific application of these two approaches, namely a Bayesian form of ART [*Herman et al.*, 1979] and the LSQR algorithm of *Paige and Saunders* [1982], and discuss the relationship between the methods of regularization used for each.

Artzy et al. [1979] performed an early comparison of this sort on synthetic data and they found that conjugate gradient methods had superior convergence behavior over Richardson's method (also termed SIRT for Simultaneous Iterative Reconstruction techniques), which is a variant of the ART techniques [Herman, 1980]. *Scales* [1986], *Spakman and Nolet* [1987], and *Van der Sluis and Van der Vorst* [1988] each made comparisons between SIRT, which is commonly used in geophysical applications [*Dines and Lytle*, 1979; *Nakanishi*, 1985; *Humphreys and Clayton*, 1988], and LSQR. They found that LSQR out performed SIRT on synthetic data as well as on real data. In addition, *Van der Sluis and Van der Vorst* argued that LSQR is preferred over SIRT due to inherent row weighting implicit in the SIRT algorithm. In this study we have applied a simple variation of Bayesian ART which does not share the undesirable feature of implicit row weighting and also performs competitively with LSQR. Furthermore, in situations of non-standard regularization it may prove easier to implement than the LSQR routine.

To illustrate this comparison we have synthesized artificial data that simulates many characteristics of the distribution of real earthquake data in 2-dimensions. For a real data situation we use arrival time data from the western Washington network near Mount St. Helens, where the successful inversion of P-wave velocity fields has led to delineation of faults and magmatic accumulation in the region [*Lees and Crosson*, 1988].

Theory: Linearization and Discretization

Under the approximations of geometrical optics, we assume that the time a seismic signal takes to travel from point A to point B (Figure 1), in a given medium, is a function of the seismic velocity of the intervening material and the path the wave traverses (the raypath). Determining the travel time, given the velocity, source and receiver locations, is called the forward problem, written mathematically as

$$T = \int_{ray} \frac{1}{\nu(\mathbf{x})} dr \tag{1.1}$$

where T is the travel time, $\mathbf{x}$ is the spatial position vector, $\nu(\mathbf{x})$ is the value of the velocity field at position $\mathbf{x}$, and dr is a differential line element along the path from A to B. The travel time is thus the line integral of the inverse of the velocity along the ray-path. Since the travel time T does not depend linearly on the velocity, $\nu(\mathbf{x})$, it is convenient to introduce the "slowness", $s(\mathbf{x})$, where $s(\mathbf{x}) = 1/\nu(\mathbf{x})$. Then our functional relationship between the travel time and the model becomes,

$$T = \int_{ray} s(\mathbf{x}) dr \tag{1.2}$$

Since the raypath itself depends on the velocity (slowness) the problem is still non-linear.

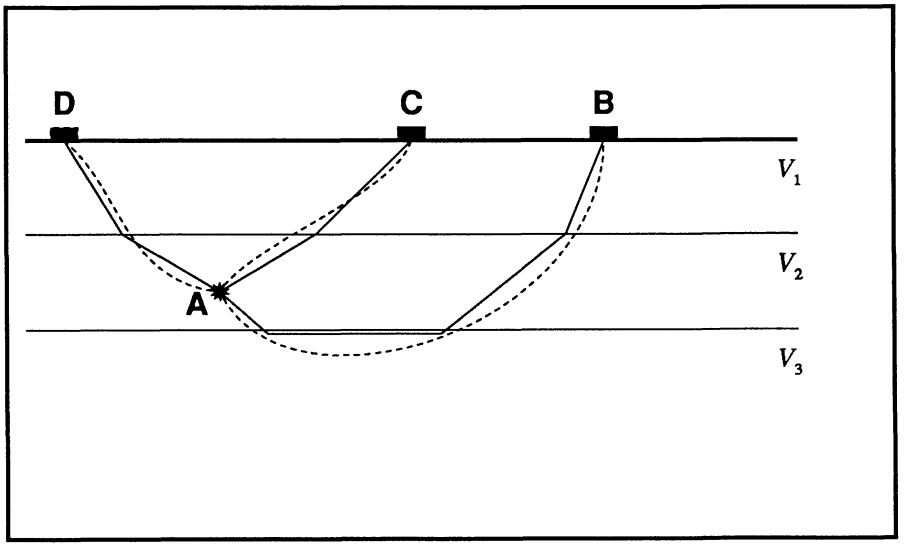

Figure 1. Hypothetical rays in a laterally varying media are displayed with dashed lines and their approximate raypaths in a horizontally stratified model in solid lines. The paths A-D and A-C are direct raypaths where as A-B is refracted along the lowest layer.

We consider the following perturbation approach. Suppose the model, $s(\mathbf{x})$, that we are seeking can be assumed to be a reference model, $s_0(\mathbf{x})$, plus a small perturbation $\delta s(\mathbf{x})$, i.e.

$$s(\mathbf{x}) = s_0(\mathbf{x}) + \delta s(\mathbf{x}) \tag{1.3}$$

Then,

$$T = \int_{ray} s_0(\mathbf{x}) + \delta s(\mathbf{x}) dr \tag{1.4}$$

Fermat's principle states that, to first order, the travel time is stationary with respect to small perturbations in the ray-path [*Aki et al.*, 1977]. For small $\delta s(\mathbf{x})$ we can therefore integrate over the raypath in the unperturbed model to determine an approximation to the right hand side of equation (1.4). If T_0 is the travel time for the unperturbed model then,

$$T \approx \int_{ray_0} s_0(\mathbf{x}) dr + \int_{ray_0} \delta s(\mathbf{x}) dr = T_0 + \int_{ray_0} \delta s(\mathbf{x}) dr \tag{1.5}$$

or, rearranging,

$$\delta T = T - T_0 \approx \int_{ray_0} \delta s(\mathbf{x}) dr \tag{1.6}$$

where δT is called the travel time residual. We now have a linear relationship between the travel time residual observations, δT, and the slowness perturbations, δs. For simplicity in notation we drop the spatial dependence $\mathbf{x}$ from our equations and refer to δs as simply s and δT as simply t, remembering that these are the slowness perturbations and residuals respectively.

Using techniques borrowed from medical tomography we parametrize the structure by partitioning it into small cells within which the slowness perturbation is considered constant. The real slowness perturbation field s_{TRUE} is thus approximated by the discrete version s, where it is assumed the blocks are chosen small enough such that $s \approx s_{\text{TRUE}}$, and there is no aliasing of structure. The inverse problem is then discretized by considering sums instead of integrals in equation (1.6). The travel time residual will be the sum of the slowness perturbation in each cell times the length of the ray within that cell. For many such observations we have,

$$t_n = \sum_{j=1}^{m} a_{nj} s_j \tag{1.7}$$

where t_n is the travel time residual associated with the *n-th* ray, and a_{nj} is the length of the *n-th* ray in the *j-th* cell. In matrix notation this can be expressed as,

$$\mathbf{t} = \mathbf{A}\mathbf{s} \tag{1.8}$$

where $\mathbf{t}$ is an n-row column vector of observations, $\mathbf{A}$ is an $n \times m$ matrix of coefficients describing the lengths of each ray in each cell, and $\mathbf{s}$ is the slowness perturbation vector of length m.

The "inverse problem" involves finding a solution $\tilde{s}$ which satisfies (1.8). If $\mathbf{A}^\dagger$ is a generalized inverse of $\mathbf{A}$, then $\tilde{s} = \mathbf{A}^\dagger \mathbf{t}$ is such a solution [*Lanczos*, 1961]. Typically $n > m$ and the system is said to be overdetermined, but the number of singular values of the matrix $\mathbf{A}$ is generally less than m, and the system will be underconstrained. The problem is thus ill-posed and the solution is not unique. We apply a common method of regularization for determining a unique solution by minimizing the functional

$$\phi = \|\mathbf{t} - \mathbf{A}\mathbf{s}\|^2 + \lambda^2 \|\mathbf{s}\|^2 \tag{1.9}$$

where λ is the trade-off parameter that regulates the relative importance we assign to models that predict the data versus models that have a characteristic, *a priori* variance.

The quality of the data depends on a variety of factors, one of which is the ability of an analyst or machine to pick the time of arrival on the seismogram. For signals of high frequency this can be done fairly consistently, but emergent signals have low frequency components and picking the onset of the first arrival can be very difficult, involving a high degree of uncertainty. Since we have more confidence in travel time picks that come from better quality data we wish to weight these data more heavily in the inversion process. The analyst who picks the data estimates a confidence interval that represents the standard error of the pick about the mean. Using these estimates we multiply each equation in (1.8) by $1/\sigma_i$, where σ_i is the estimated uncertainty in

the *i-th* datum. In addition to simple row weighting we may anticipate covariances between data points, as would be the case for several observations obtained from the same earthquake. If the covariance matrix of data is $\mathbf{C}_{data}$, then we form the weighting matrix

$$\mathbf{W} = \mathbf{C}_{data}^{-1/2}.$$

The diagonal elements of $\mathbf{W}$ are the weights for each equation, $1/\sigma_i$, and the off-diagonal elements are the covariances between the data values (which are typically not available). Equation (1.8) is transformed to:

$$\mathbf{Wt} = \mathbf{WAs}. \tag{1.10}$$

In a similar fashion we may have *a priori* knowledge of the covariance of the model parameters. First, if we know in advance what the ray coverage is, we may weight blocks differently depending on the configuration of rays in the vicinity of the block. For instance, blocks that have heavy coverage may be weighted in such a way as to reflect the fact that we have more confidence in them as opposed to blocks that are lightly sampled. Second, we may have *a priori* knowledge of the geological structures in the subsurface, like the wavelength of typical features, the location of faults or the presence of specific buried features which we obtain from external sources. In either case we can require that the solution have the desired covariance $\mathbf{C}_{model}$ by considering a change of variables,

$$\mathbf{x} = \mathbf{C}^{-1/2}\mathbf{s}$$

Then the system is replaced by

$$\mathbf{Wt} = \mathbf{WAC}^{1/2}\mathbf{x}. \tag{1.11}$$

Alternatively, *a priori* information can be introduced by forcing the system of equations in (1.8) to satisfy an additional, artificial set of equations, one for each model parameter:

$$\begin{bmatrix} \mathbf{A} \\ \lambda\mathbf{F} \end{bmatrix} \mathbf{s} = \begin{bmatrix} \mathbf{t} \\ 0 \end{bmatrix} \tag{1.12}$$

We then minimize the functional,

$$\phi = \|\mathbf{t} - \mathbf{As}\|^2 + \lambda^2\|\mathbf{Fs}\|^2 \tag{1.13}$$

where, as before, λ is a trade-off parameter that regulates the relative importance one puts in minimization of the prediction error versus minimization of the variance of the model. If $\mathbf{F} = \mathbf{I}$, this expression leads to the minimization of the functional ϕ from equation (1.9), which is the Levenberg-Marquardt method of regularization. In our case we wish to impose a degree of smoothness on the model, and $\mathbf{F}$ will be a roughening matrix.

Inversion Techniques: Bayesian ART

As a first approach to finding a solution of (1.8) where we minimize (1.9) we apply a Bayesian version of the simple ART (Kazcmarz' method) proposed by G.T. Herman [*Herman et al.*, 1979; *Herman*, 1980]. Since the system of equations in (1.8) is inconsistent we consider the system,

$$\mathbf{t} = \mathbf{As} + \mathbf{r} \tag{2.1}$$

where $\mathbf{r}$ is chosen such that given any $\mathbf{s}$, $\mathbf{r} = \mathbf{t} - \mathbf{As}$. This means that (2.1) is a well-posed problem and we will solve for $\mathbf{s}$ and $\mathbf{r}$ simultaneously. We form the following consistent system of equations,

$$[\lambda \mathbf{I}\ \mathbf{A}] \begin{bmatrix} \mathbf{r} \\ \mathbf{s} \end{bmatrix} = \mathbf{t}. \tag{2.2}$$

whose solution minimizes ϕ in (1.9), [*Herman*, 1980]. Application of the simple ART algorithm to (2.2) yields the following modified algorithm:

Bayesian ART Algorithm:

$$\textit{Initialize}: \begin{cases} \mathbf{r}^0 = 0 \\ \mathbf{s}^0 = 0 \end{cases}$$

For $k = K \bmod(n) + 1 = \{1, 2, \ldots, n, 1, 2, \ldots, n, \ldots\}$:

$$\gamma_k = \rho_k \frac{t_i - \lambda(\mathbf{r}_i) - (\mathbf{a}_i^T \mathbf{s}^{(k)})}{\lambda^2 + \|\mathbf{a}_i\|^2} \tag{2.3}$$

$$\mathbf{r}^{(k+1)} \leftarrow \mathbf{r}^{(k)} + \gamma_k \lambda \hat{\mathbf{e}}_i$$

$$\mathbf{s}^{(k+1)} \leftarrow \mathbf{s}^{(k)} + \gamma_k \mathbf{a}_i$$

where $\mathbf{a}_i$ is the ith row of $\mathbf{A}$ ($\mathbf{a}_i^T$ is its transpose), $\hat{\mathbf{e}}_i$ is a unit vector with the $i-th$ element set to one, ρ_k is a relaxation parameter used to stabilize convergence, λ is the regularization parameter of (1.13) when $\mathbf{F} = \mathbf{I}$, and K is the number of iterations. Notice that if there is no regularization, i.e. $\lambda = 0$, this algorithm becomes Kazcmarz' simple ART algorithm. An important aspect of the introduction of regularization is that with $\lambda \neq 0$ the explicit weighting of (1.10) is retained where as with Kazcmarz' method there is implicit weighting in (2.3) (in effect, normalizing the rows to unit length) which may not be desirable [*Van der Sluis and Van der Vorst*, 1987]. The value of ρ_k may remain constant or it may change between iterations. In situations where we have allowed ρ_k to change we have used the following scheme:

$$\rho_k = \frac{\kappa_1}{\kappa_2 + k} \tag{2.4}$$

where κ_1 and κ_2 are suitably chosen constants. This way of decreasing ρ_k is in accordance with the sufficient conditions required for convergence as outlined in *Trummer* [1981].

Smoothing is introduced by applying a low-pass filter after each iteration, i.e. after all the equations in the system have been used. In this study the updated version of the model is formed by taking a linear combination of the present version and the smoothed version,

$$\mathbf{S}^{(k+1)} \leftarrow (1 - \psi_k)\mathbf{S}^{(k)} + \psi_k \mathbf{S}_{smooth}$$

Such that the amount of smoothing can depend on the iteration (k) without changing the filter. Thus for the $i-th$ element of $\mathbf{s}$,

$$\mathbf{S}_i^{(k+1)} \leftarrow (1 - \psi_k)\mathbf{S}_i^{(k)} + \psi_k \sum_{j=1}^{m} a_j \mathbf{S}_j^{(k)} \tag{2.5}$$

The reduction in smoothing is tied into the reduction of relaxation by multiplicative constant:

$$\psi_k = \kappa_3 \rho_k \tag{2.6}$$

such that the smoothness constraint does not dominate over reduction of data misfit. Smoothing between iterations is often called a "trick" in the literature [*Herman*, 1980] for two basic reasons. First, the smoothing procedure destroys the least squares nature of the solution, meaning the solution after smoothing is no longer one that uniquely minimizes (1.9). Second, the smoothing operation is applied to all the model parameters simultaneously, much like a SIRT iterative step, and thus does not fit mathematically into the ART inversion framework. Experience indicates that in spite of these difficulties, solutions obtained in this fashion are nearly identical to more analytical, least squares approaches. A comparison is provided in the next section. The advantage of using ART with tricks lies in its flexibility. For example, one could apply a nonlinear smoothing scheme to the images where the amount of smoothing depends on the image itself. This might even be applied in an interactive mode where the analyst adjusts smoothing or other parameters between iterations as needed. The ART approach also lends itself easily to robust inversion where the magnitude of the residual that is backprojected depends on the size of the residual itself. Again, applying these "tricks" implies a departure from strict least squares inversions. The desirability of this departure depends on the nature of the problem and data set.

Conjugate Gradient - LSQR

A different approach to solving equation (1.8) has been suggested by *Spackman and Nolet* [1988] and *Scales* [1987]. Here we use an algorithm developed by *Paige and Saunders* [1982] called LSQR. It is based on Lanczos procedures used to bidiagonalize the system of equations followed by a QR decomposition to obtain the solution. At each step the procedure can be shown to produce iterative approximations which are equivalent to conjugate gradient solutions, and so we call it conjugate gradients. The details of the method are outlined in *Golub and Van Loan* [1983], *Spakman and Nolet* [1988], and *Van der Sluis and Van der Vorst* [1987].

In this case we will implement regularization by augmenting the system in (1.8) by m additional equations as in (1.12) where the matrix $\mathbf{F}$ is the second order differential operator applied in horizontal layers, i.e. the 2-dimensional Laplacian. Using a first order discrete approximation for the second derivatives, this constraint corresponds to the following equation for the $j-th$ block:

$$4^*(j\text{–}th \text{ block slowness}) - \sum(\text{adjacent slowness}) = 4x_j - (x_{j+1} + x_{j-1} + x_{j+n} + x_{j-n}) = 0$$

where n is the number of blocks on the side of the model. The result is a minimization of the roughness of the model as measured by the differential operator. A discussion of the use of differential operators as roughening constraints can be found in *Titterington* [1985] and *Constable et al.* [1987], and the discrete Laplacian operator is introduced in *Young* [1971].

The equivalence of maximizing smoothness with ART and minimizing roughness with LSQR can be illustrated best in the spatial frequency domain. A (k_x, k_y) plot of the frequency response of the Laplacian filter is displayed in Figure 2a, where the high pass nature of the roughening operator is clearly evident. The inverse filter can be obtained by taking a suitably normalized spike and subtracting it from the Laplacian in the space domain [*Crosson and Lees*, 1989]. The response of this filter is displayed

in Figure 2b, and the result is a low-pass filter as expected. This low-pass filter, or one similar to it, was used for smoothing in the ART scheme outlined above.

Resolution

The resolution of the inversion is primarily a function of the ray distribution. In the context of classical least squares, if **A** is the matrix that describes the way the rays sample the model blocks, the resolution matrix is $\mathbf{R} = (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{A}$ [*Crosson*, 1976]. For our modified version of least squares, $\mathbf{R} = (\mathbf{A}^T\mathbf{W}^T\mathbf{W}\mathbf{A} + \lambda^2\mathbf{F}^T\mathbf{F})^{-1}\mathbf{A}^T\mathbf{W}^T\mathbf{W}\mathbf{A}$. However, the sheer volume of data used in an inversion of this sort makes the calculation of the resolution matrix prohibitive. For this reason we resort to estimating the resolution by other means. To accomplish this we examined the response of the system, that is ray coverage and inversion procedure, to spikes located in areas where resolution information was required. If the ray coverage over a region is relatively homogeneous and isotropic, the resolution kernels will be nearly identical for all blocks in the region [*Humphreys*, 1984]. Ray coverage diagrams were thus used in conjunction with impulse responses to estimate the resolution at critical regions in the target volume.

Error Analysis

For the same reasons outlined above, the model covariance matrix is also unavailable from a practical standpoint. In this case we applied the statistical technique called "jackknife" [*Efron*, 1982] to estimate the standard errors of model parameters. This approach is similar to the bootstrap technique suggested by *Willmott et al.* [1985]. The jackknife involves partitioning the data set into subsets, performing inversions on the subsets, and calculating a standard error from the set of image vectors that result [*Mosteller and Tukey*, 1977; *Efron*, 1982]. The process is described as follows. Divide the data into k sets, with each set leaving out a random $1/k$th portion of the data without replacement. Perform an inversion for each of the k subsets and call the slowness image derived from such an inversion $\hat{\mathbf{s}}_j$. Create from these "mini-inversions" a "pseudo-inversion" by forming the following linear combination:

$$\tilde{\mathbf{s}}_j = k\hat{\mathbf{s}}_{all} - (k-1)\hat{\mathbf{s}}_j. \tag{3.1}$$

The jackknife estimate of the slowness is simply the average of the pseudo-inversions:

$$\tilde{\mathbf{s}} = \frac{\sum^k \tilde{\mathbf{s}}_j}{k} \tag{3.2}$$

which has variance,

$$\nu = \frac{\sum \tilde{\mathbf{s}}_j^2 - \frac{1}{k}(\sum \tilde{\mathbf{s}}_j)^2}{k(k-1)}. \tag{3.3}$$

From this the standard error is $\mathbf{E}_\sigma = \sqrt{\nu}$. This will be an estimate of the variability of the model and can be used to project how large the errors are in each block of the target. Normally k would be the number of rays in the data set, meaning, leaving out one ray for each pseudo-inversion [*Mosteller and Tukey*, 1977]. However, this would involve performing thousands of inversions, a very time consuming operation. A compromise can be struck if we assume that the variability in the pseudo-inversions will be represented in far fewer partitions of the data. In this study, for example, only 30 partitions were used. The advantage of this statistical approach is that we are using the

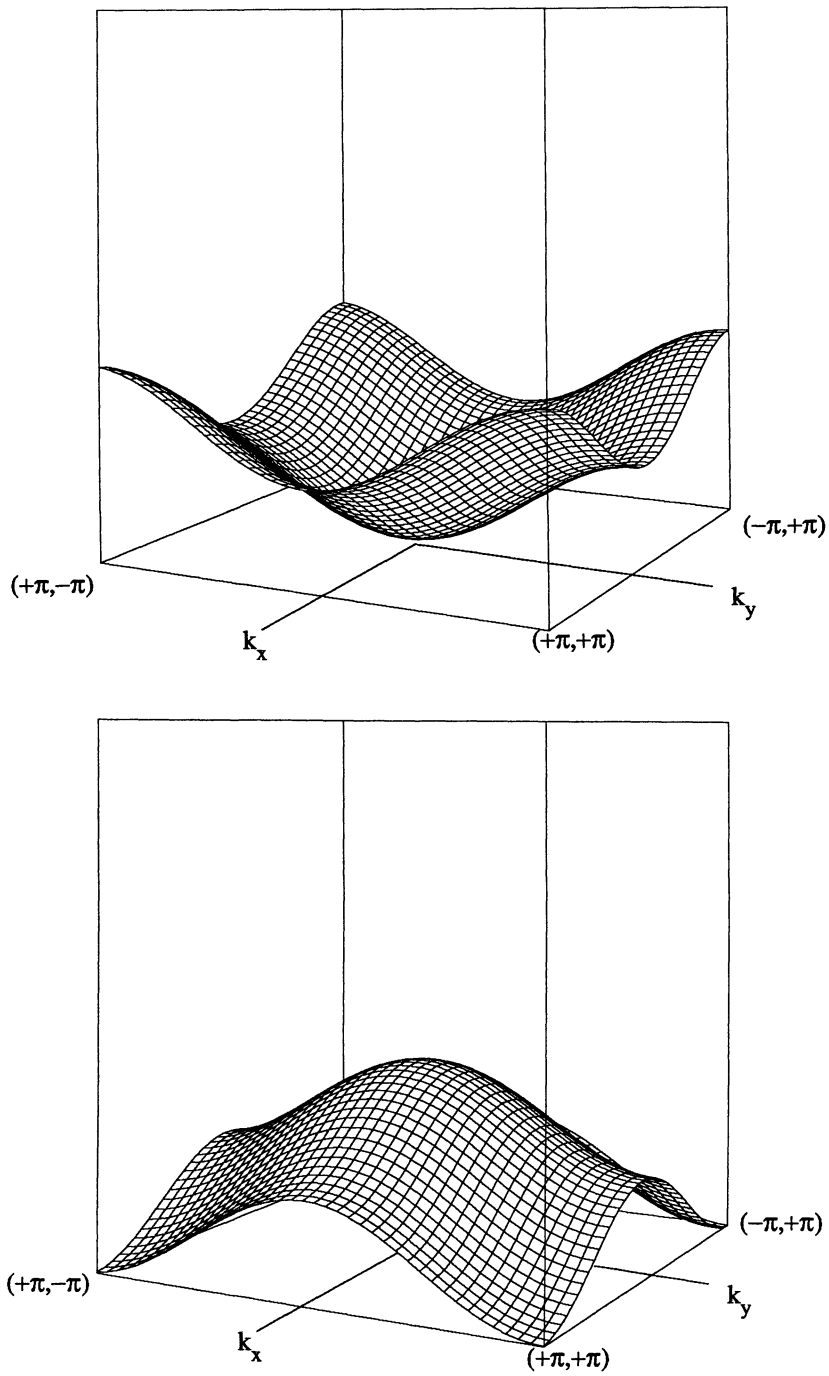

Figure 2. (a) The spatial frequency domain representation of the Laplacian filter as described in the text. Notice the high pass character of this roughening operator. (b) The low-pass complement of the filter in (a).

data themselves to estimate the covariance and we have made no assumptions about how the data are distributed.

Synthetic Phantom

The comparison of the two approaches is illustrated by simulating seismic data on a synthetic phantom (model) and applying the methods to the data set. For speed and simplicity the synthetic is defined in two dimensions, however extension to three dimensions is straightforward.

The phantom, consisting of a cross and a torus, is illustrated in Figure 3(a), where we use grayshades to represent levels of percent perturbation from the reference slowness. 20 stations (Figure 3(b)) are randomly distributed over the phantom and 236 sources are similarly distributed (Figure 3(c)) and connected to nearby stations in a fashion that simulates real earthquake data, resulting in 3000 earthquake-station rays (Figure 3(d)). Forward modeling is performed by summing the perturbation and background velocities in the blocks along straight raypaths. To simulate real data, perturbations are added to each observation such that 80% of the root mean square travel time residual is random Gaussian noise.

To monitor the convergence behavior we have plotted the sum of the squared weighted prediction error,

$$\chi^2 \equiv \sum_{i=1}^{n} [w_i(t_i - \sum_j a_{ij} s_j)]^2 \tag{4.1}$$

versus iteration number, k. Using $\mathbf{s}$, $\tilde{\mathbf{s}}$ and $\bar{\mathbf{s}}$ to represent the phantom, the reconstructed image and the mean value of $\tilde{\mathbf{s}}$ respectively, the following quantitative measures of distance from the phantom are provided for each synthetic inversion [*Herman*, 1980]:

$$d_1 = \left[\sum_{i=1}^{m} (\mathbf{s}_i - \tilde{\mathbf{s}}_i)^2 / \sum_{i=1}^{m} (\tilde{\mathbf{s}}_i - \bar{\mathbf{s}})^2 \right]^{1/2} \tag{4.2a}$$

$$d_2 = \sum_{i=1}^{m} |\mathbf{s}_i - \tilde{\mathbf{s}}_i| / \sum_{i=1}^{m} |\tilde{\mathbf{s}}_i| \tag{4.2b}$$

$$d_3 = \max |\mathbf{s}_i - \tilde{\mathbf{s}}_i| \tag{4.2c}$$

These represent the normalized root mean squared distance, the average absolute value distance and the worst case distance respectively and are listed in Table 1.

Table 1. Distance measures of inversion images to phantom model. Here d_1 is the root mean square normalized distance, d_2 is the average absolute value difference, and d_3 is a measure of worst case.

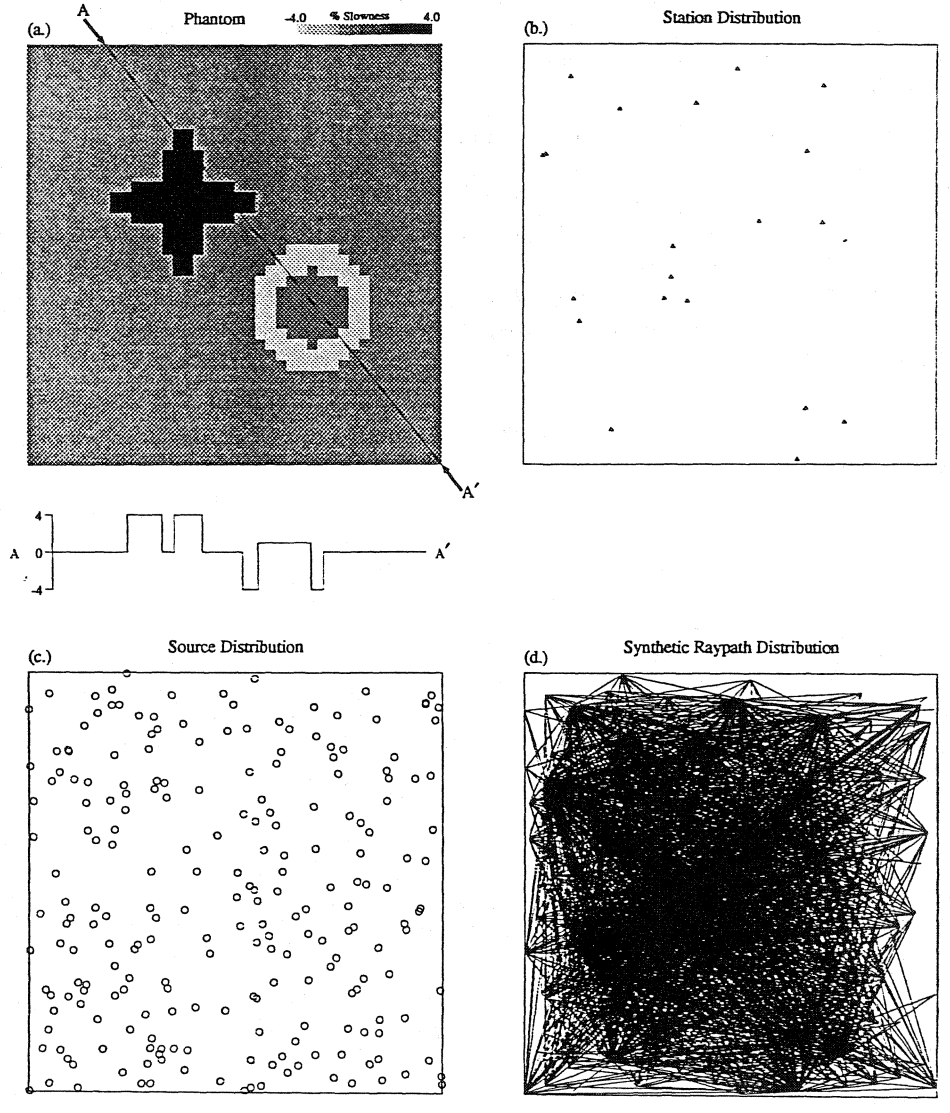

Figure 3. Description of the phantom and synthetic data set. (a) Phantom model represented in percent perturbation from a constant background slowness. The line from A-A' is a cross section displayed for each inversion. (b) Random synthetic station distribution. (c) Random synthetic source distribution. (d) Raypaths connecting stations in (b) to sources in (c).

Distance Measures to Phantom			
Model	d_1	d_2	d_3
ART1	0.816994	0.065758	0.001112
LSQR1	1.938234	0.144696	0.002443
ART2	9.232765	0.691058	0.007067
LSQR2	9.232765	0.691058	0.007067
ART3	8.686169	0.613092	0.006573
LSQR3	8.985689	0.615608	0.006460
ART4	8.234601	0.752227	0.006503
ART5	8.353343	0.608981	0.006614

Comparison

As a first comparison we show the results of inversion with equal number of iterations (labeled ART1 and LSQR1) when no noise is added to the data. In this case we need neither regularization nor relaxation. Note in Figure 4(a) and (b) that the reconstruction is nearly perfect for both ART1 and LSQR1, but ART1 has outperformed LSQR1 (for the same number of iterations) in the lower right corner where the data coverage is more sparse. This is because ART converges faster than LSQR in this case. All three distance measures of Table 1 also indicate that ART1 is a better reconstruction than LSQR1.

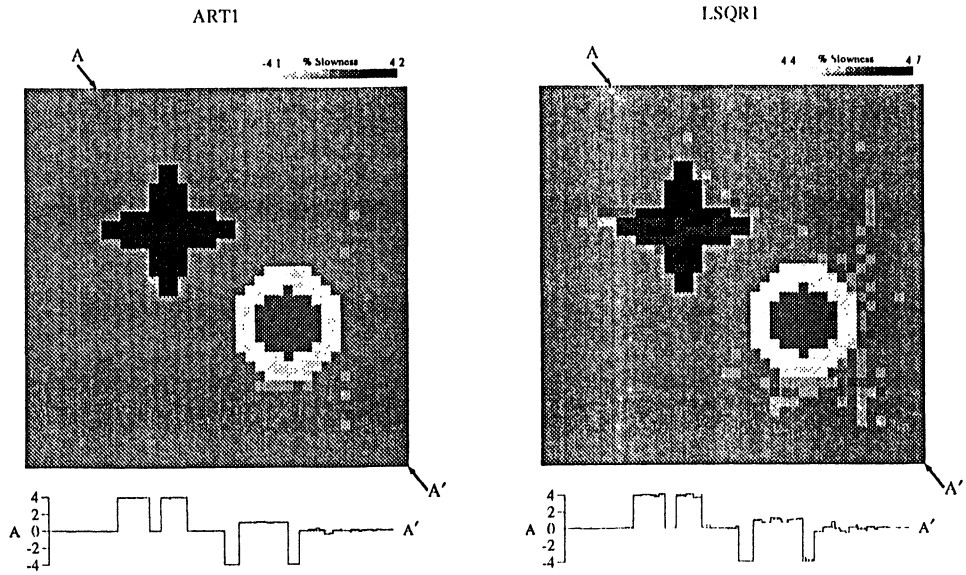

Figure 4. Comparison of ART1 (a) and LSQR1 (b) using data with no noise. Notice that the LSQR1 inversion is worse than the ART1 inversion in the lower right corner where the coverage is sparse.

In Figures 5(a) and (b) noise (80% of the travel time misfit) is introduced and damping is applied but no smoothness constraints are implemented (ART2 and LSQR2). Here, using the same damping, the resulting images are identical. The characteristic

"salt and pepper" degradation appears, as noise from the data is projected into the image but the phantom is still evident beneath the noise.

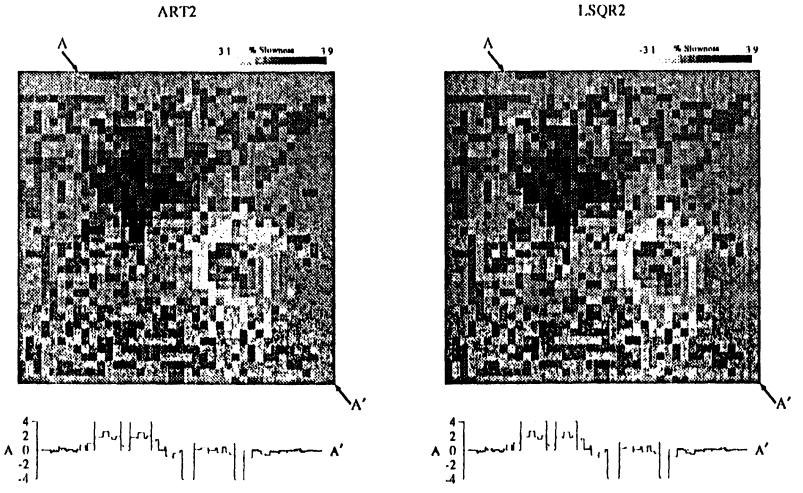

Figure 5. Comparison of ART2 (a) and LSQR2 (b) using data with noise and a damping of $\lambda = 400$. The results are identical after 30 iterations.

Since there is no widely accepted analytical method for determining the choice of λ, a range of values were tried and the best results (both visually and in terms of the distance measures) were obtained for $\lambda = 400$. The method of *Hoerl et al.* [1975], where λ is chosen to be the ratio of the mean square data variance to the mean square variance of the least squares model, produced unacceptable results (λ was too small). For the synthetic data cross validatory techniques [*Titterington*, 1985] indicated a choice of λ between 300 and 500, although for the real data inversion no optimal choice of λ could be determined.

Next, smoothing is implemented by constraining the high-pass Laplacian for the LSQR3 inversion and applying the complementary low-pass filter between iterations for ART3 (Figure 6(a) and (b)).

In the ART3 case a relaxation of $\rho^{(k)} = 0.02$ is used with regularization parameter adjusted to $\lambda = 65$. Here, even though the solutions predict the data to the same χ^2 accuracy, they appear slightly different, due to the different way the smoothing is performed in each method. Table 1 indicates that ART3 is over all slightly closer to the phantom than LSQR3, but has a higher worst case, d_3. If we apply this same filter but keep $\rho^{(k)} = 1.0$ with $\lambda = 400.0$ as in the previous example, the technique does not converge in 30 iterations. After the first few iterations, the smoothing dominates the inversion process, producing successively smoother models which converges to a different χ^2.

Figure 7(a) shows ART4 after 30 iterations. Notice in Table 1 that the d_1 measure is smaller for ART4 than ART3 but d_2 is larger. As a final example, ART5, the relaxation parameter is allowed to decrease between iterations as in (2.4) with $\kappa_1 = 1$, $\kappa_2 = 30$ and a suitable reduction in smoothing following (2.6) with $\kappa_3 = 30$. This results in an

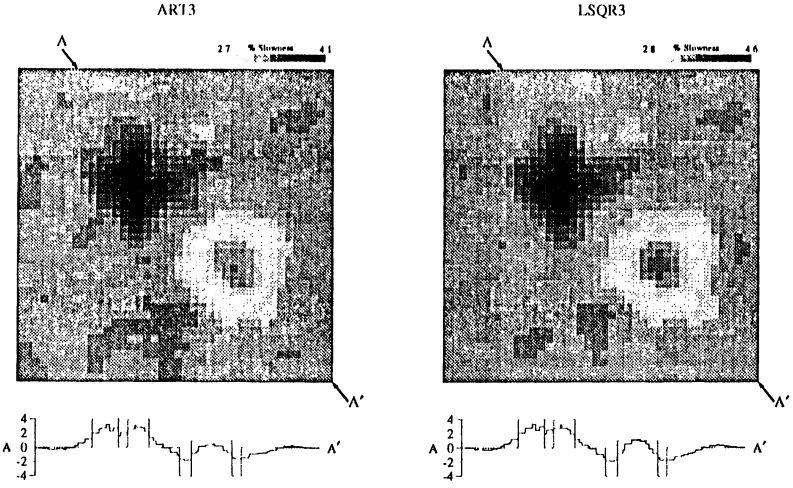

Figure 6. Comparison of ART3 (a) and LSQR3 (b) using data with noise. In this example smoothing has been applied as filtering for ART3 and roughening constraints for LSQR3.

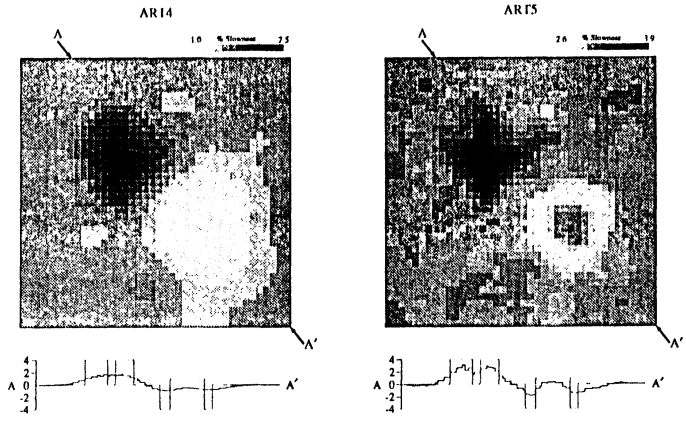

Figure 7. (a) Smoothed ART4 with $\rho_k = 1.0$, $\lambda = 400.0$ and a smoothing filter applied. This result is considered unacceptable because it has not converged to the appropriate χ^2. (b) Smoothed ART5 with ρ_k decreasing and ψ_k, the smoothing parameter, also decreasing.

image (Figure 7 (b)) very close to that of ART3 in Figure 6(a) and is indicated as such in Table 1. Since the selection of these various parameters $(\kappa_1, \kappa_2, \kappa_3)$ requires very careful analysis of many trial inversions, we prefer the approach taken in ART3 where the relaxation and regularization is fixed. ART5 is presented mainly to show that this approach can be used if desired.

In Figure 8 we present a plot of the $\%\chi^2$ reduction versus iteration number for each of the synthetic inversions. For the case of perfect data, ART1 reduces the χ^2 faster than LSQR1 until the 8-*th* iteration where the respective χ^2 converges to nearly 100% reduction. In the damped case ART2 out performs LSQR2 until 4 iterations when the

two methods converge.

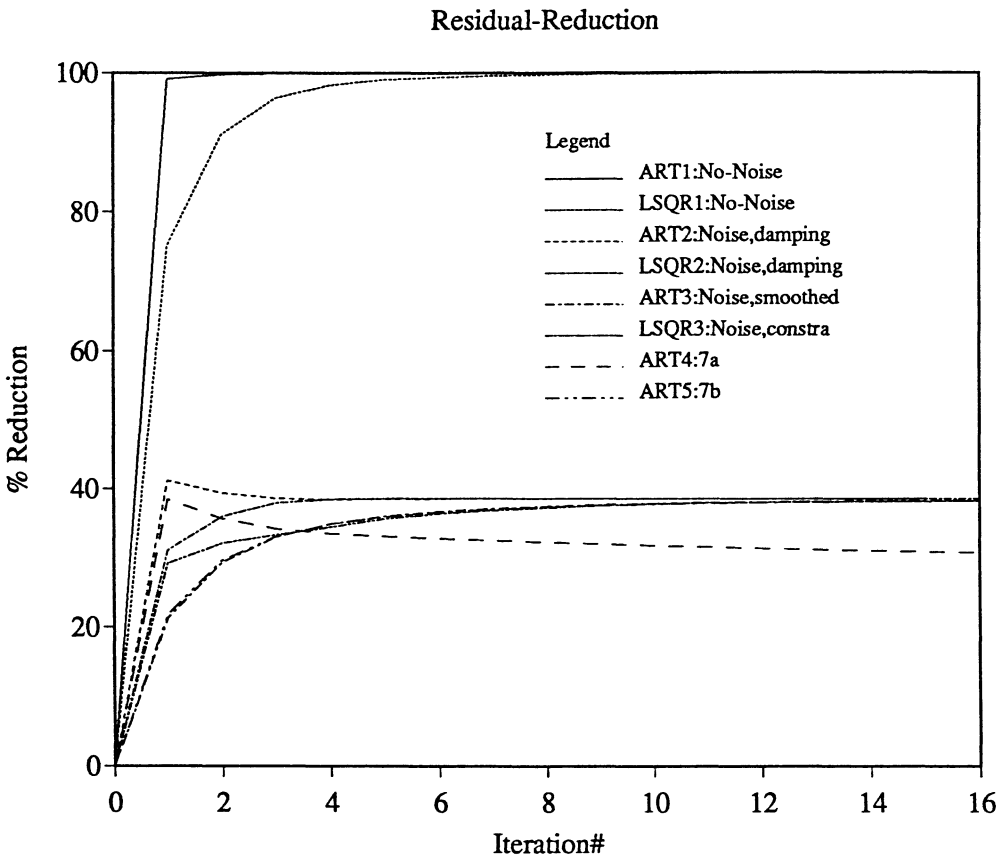

Figure 8. Plots of the χ^2 verses iteration for the synthetic inversions.

When smoothing is applied in ART3 and constraining in LSQR3 the situation is reversed. Here LSQR3 appears to have reduced the χ^2 faster for the first 3 iterations where the χ^2 for each merge. The χ^2 reduction for ART4, however, is better than that of LSQR3 for the first 3 iterations, after which the smoothing degrades the data predictive nature of the solution. If the relaxation is allowed to decay with appropriate decay of smoothing, the χ^2 of ART5 is made to follow that of ART3.

Figure 9(a) and (b) displays the standard error for the inversions in ART3 and LSQR3 of Figure 6 respectively. Since the methods of regularization are different in each case, the mini-inversions and pseudo values calculated will vary accordingly giving rise to different distributions of standard errors. Overall, though, the errors are of the same magnitude. In both cases the estimated errors are generally larger in the lower right corner where coverage is sparse.

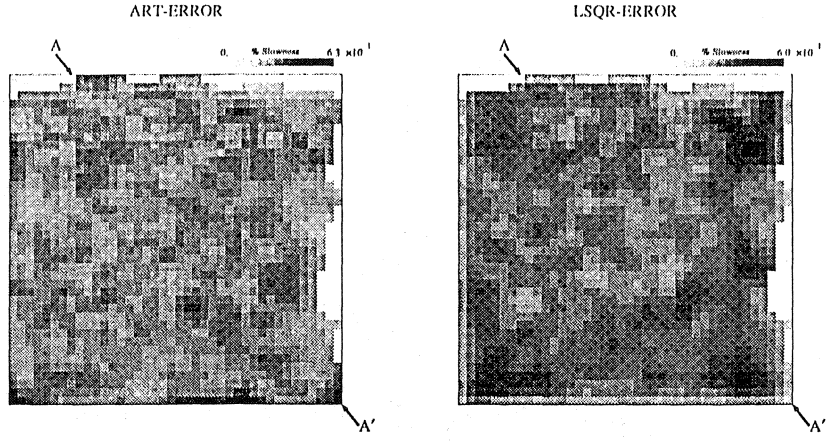

Figure 9. (a) Standard errors for the ART3 inversion in Figure 6(a) as estimated by the jackknife technique. (b) same as (a) but for the LSQR3 inversion in 6(b).

Real Data: Mount St. Helens

To illustrate the usage of these techniques on real, 3-dimensional data, results from an inversion in the vicinity of Mount St. Helens, Washington, are presented. A geographic plot of the target region is displayed in Figure 10 where the 21 stations and 2,023 sources (Figure 11) are plotted. Earthquake data were taken from the University of Washington database ranging in time from 1972 to 1988. There were 17,659 rays covering the $80 \times 80 \times 40$ km region which was partitioned into 2×2 km blocks over 10 layers of varying thickness.

Since data coverage is best between 2-12 km depth, only the four layers spanning this region are presented. For the ART case, Figure 12(a-d) a relaxation of $\rho = .02$ and $\lambda = 260$ have been used for 30 iterations where a 24% reduction of χ^2 was achieved. Figure 13(a-d) shows the results of applying LSQR, where Laplacian smoothing with a damping parameter $\lambda = 1600$ was used for 40 iterations when the same χ^2 was reached. Minor differences are evident for the two procedures but in general results are nearly the same. Small adjustments in parameters can refine this comparison. Figure 14 shows a closeup of layer 2 with prominent geologic features juxtaposed. The high velocity regions correlate well with plutonic centers where we would expect rocks to be consolidated. Low velocity structures along the St. Helens seismic zone and the Goat Mt complex correlates with zones of faulting or intruded magmas where we expect fractured rocks to have lower velocities. These results correlate very well with other geophysical measurements, particularly aeromagnetic data [Finn and Williams, 1987]. We thus have a high degree of confidence in our results. Layers 4 and 5 show a strong low velocity anomaly just south of the crater, perhaps indicating the presence of the modern magma source beneath the volcano. For more details see Lees and Crosson [1989]. Future tomographic inversions will attempt to delineate the small scale details of the magma chamber. Inversions with the ART algorithm produce similar results.

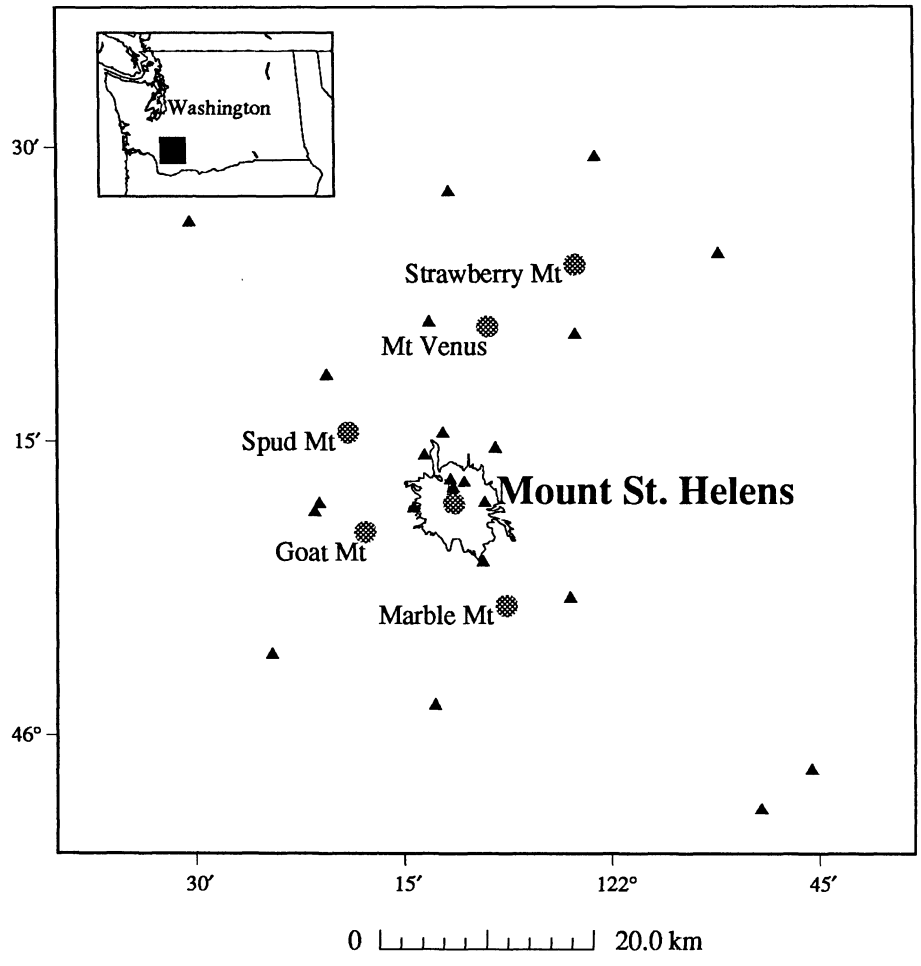

Figure 10. Map view of target area showing station distribution. The inset shows the location of target area in Washington. Triangles are station locations and labeled bullets identify major topographic highs in the area. The pre-1980 eruption tree line around Mount St. Helens is plotted for reference.

Conclusions

We conclude that with careful choice of parameters iterative techniques such as ART and projection methods like LSQR produce nearly equivalent results for tomographic inversions using seismic travel time data. Both methods can successfully incorporate regularization and various approaches to smoothing can be introduced. The ART approach has as its main draw-back that the application of smoothing destroys the least squares nature of the solution, although with adequate iteration and careful reduction of smoothing this effect may be diminished. The advantage of ART is its simplicity and ease of implementation when non-linear constraints are required, or when other, unusual *a priori* information is deemed important for constraining. Both methods have been found to work well and produce comparable results for synthetic as well as real data.

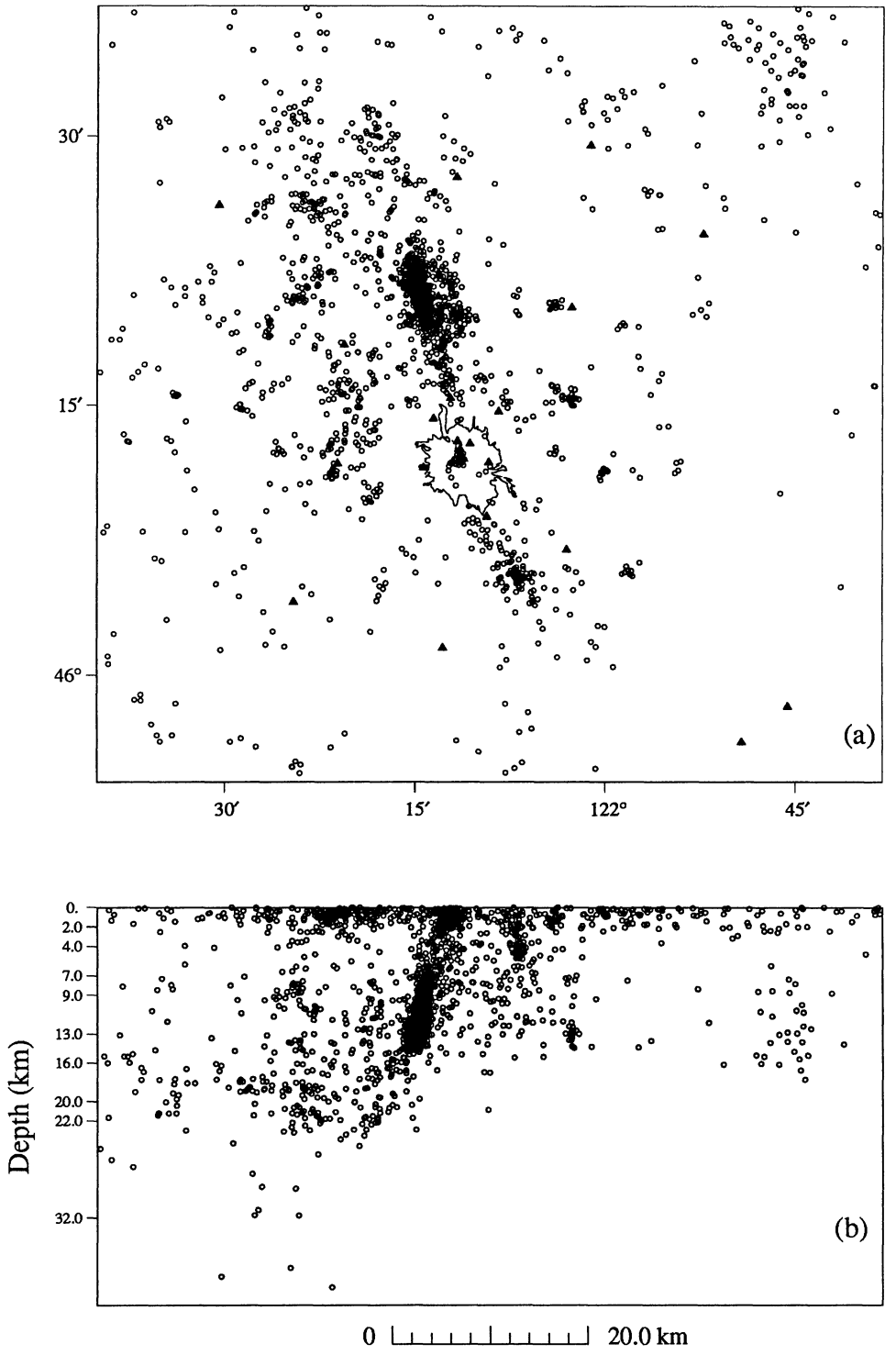

Figure 11. (a) Epicenter map and (b) east-west vertical projection of earthquake sources within target area at Mt. St. Helens.

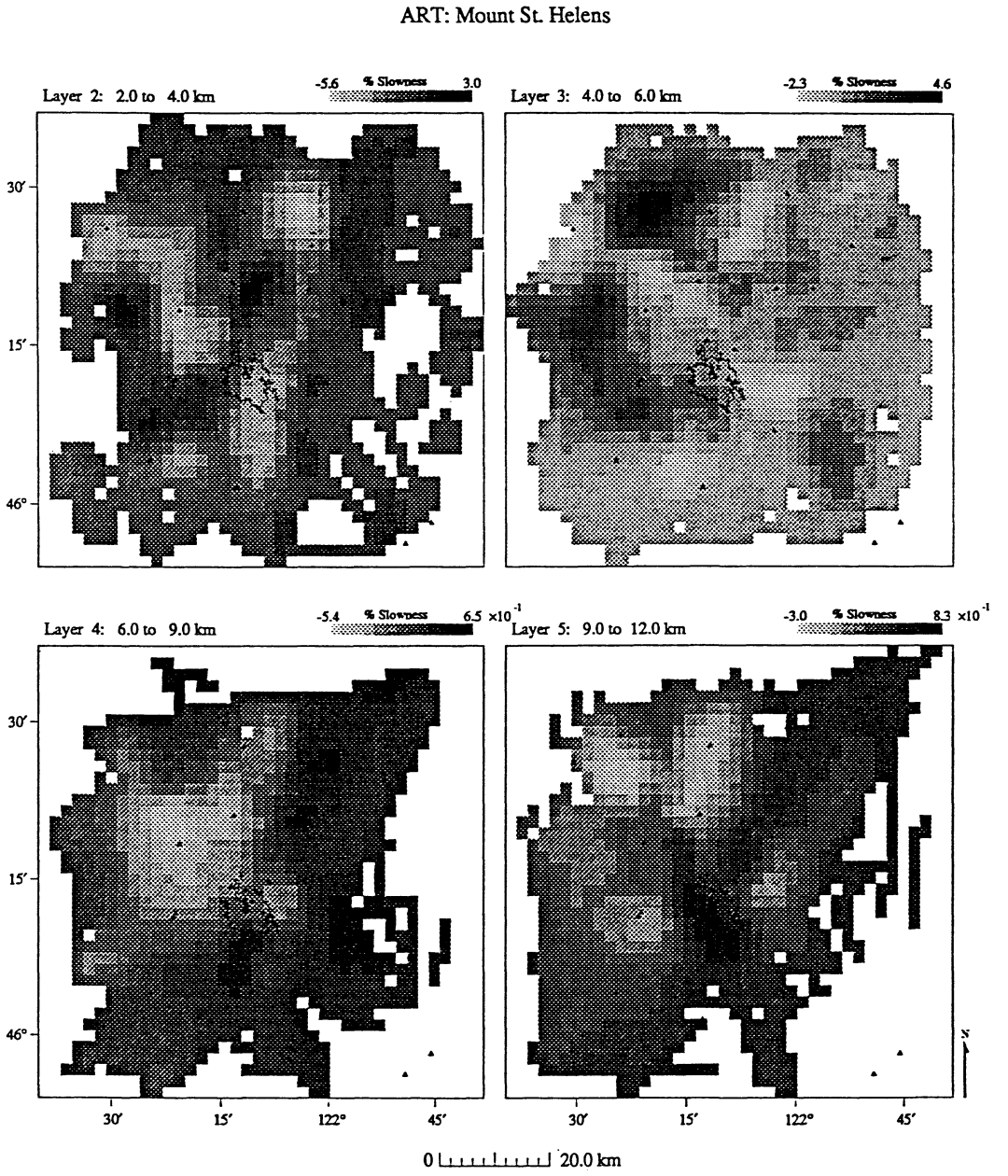

Figure 12. (a-b) 3-D tomographic inversion using ART for lateral variations in slowness perturbation beneath Mt. St. Helens. Dark triangles are station locations as illustrated in Figure 10.

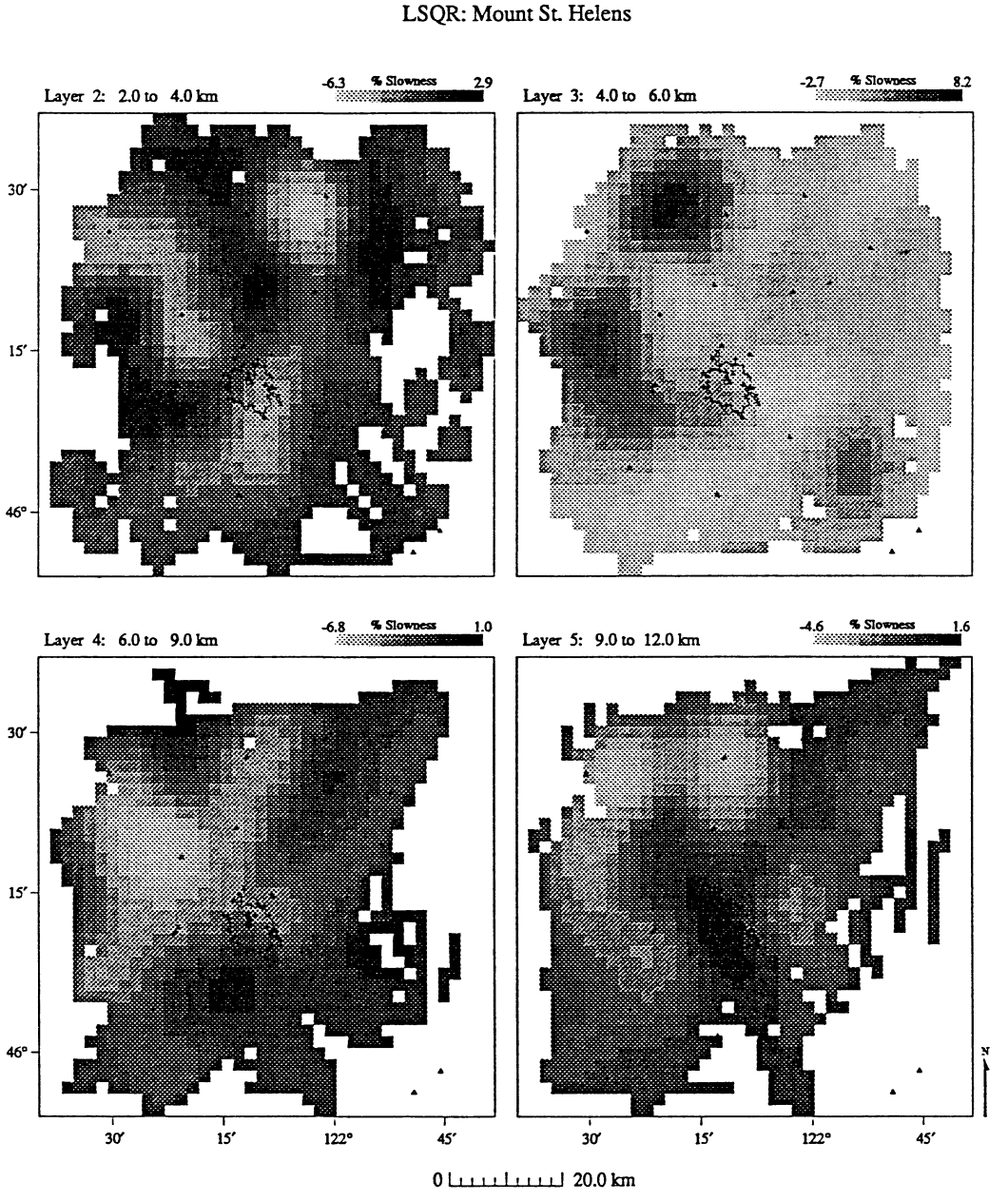

Figure 13. (a-b) 3-D tomographic inversion using LSQR for lateral variations in slowness perturbation beneath Mt. St. Helens.

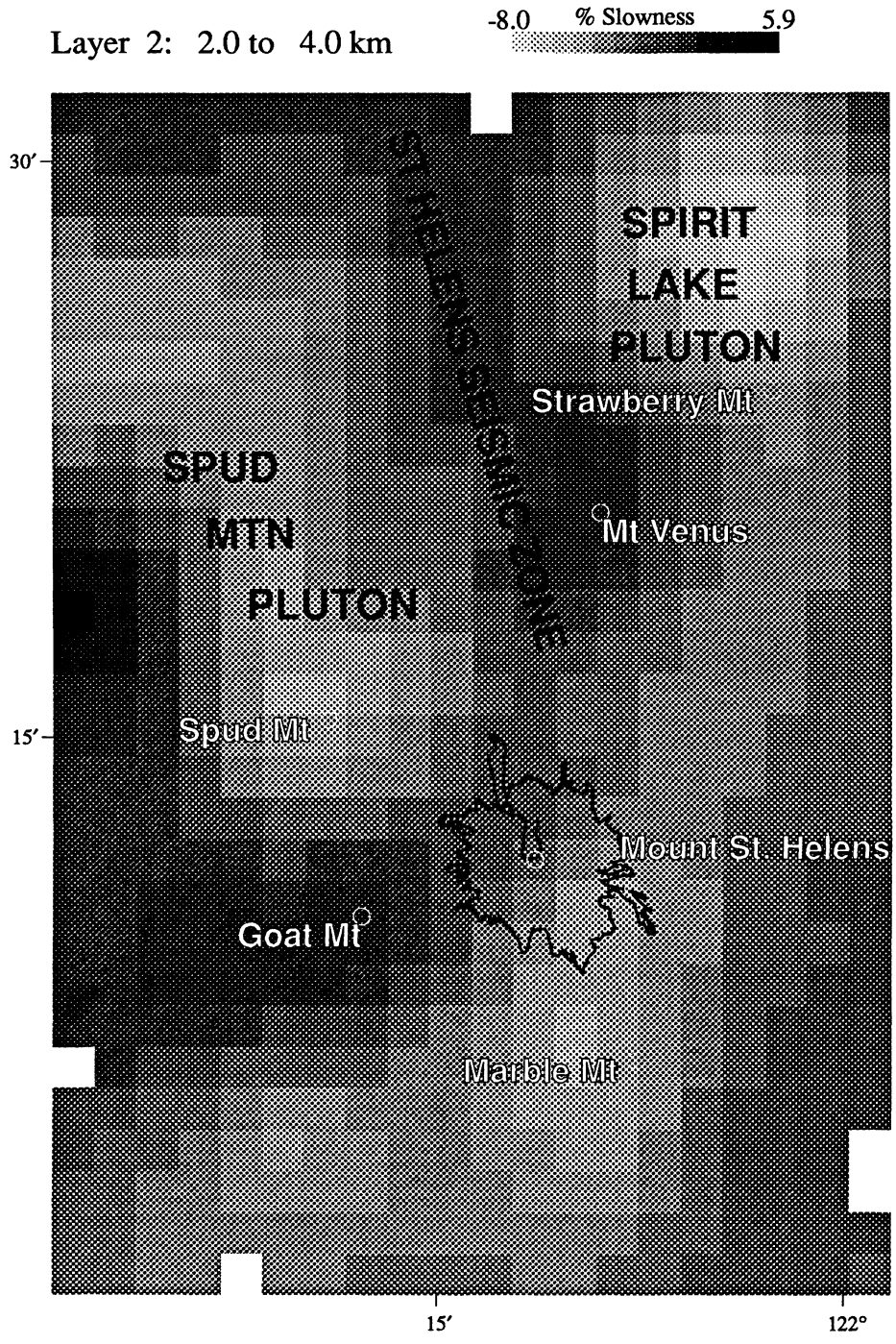

Figure 14. Blow-up of Mount St. Helens inversion Layer 2 with prominent geologic features marked.

Acknowledgements

The authors are grateful to Antonio Possolo and Torquil Smith for many helpful discussions. We thank John Vandecar and Larry Ruff for reviewing the manuscript. This work was supported by USGS Grants 14-08-0001-G1080 and 14-08-0001-G1390 and USGS Contract 14-08-0001-22169.

References

Aki, K., Christoffersson, A., & Husebye, E. S. (1977). Determination of the three-dimensional seismic structure of the lithosphere. *Journal of Geophysical Research* **82**, 277-296.

Artzy, E., Elfing, T., & Herman, G. T. (1983). Quadratic optimization for image reconstruction, II. *Computer Graphics and Image Processing* **11**, 242-261.

Constable, A. C., Parker, R. L., & Constable, C. G. (1987). Occam's inversion: A practical algorithm for generating smooth models from electromagnetic sounding data. *Geophysics* **52** (3), 289-300.

Crosson, R. S. (1976). Crustal structure modeling of earthquake data 1. Simultaneous least squares estimation of hypocenter and velocity parameters. *Journal of Geophysical Research* **81** (17), 3036-3046.

Crosson, R. S., & Lees, J. M. (1989). Regularization or smoothing in inversion and seismic tomography as a linear filtering operation. *Eos* **70** (20), 601.

Dines, K. A., & Lytle, R. J. (1979). Computerized geophysical tomography. *Proceedings of the IEEE* **67**, 1065-1073.

Efron, B. (1982). *The Jackknife, the Bootstrap and Other Resampling Plans.* Society for Industrial and Applied Mathematics.

Finn, C., & Williams, D. L. (1987). An aeromagnetic study of Mount St. Helens. *Journal of Geophysical Research* **92**, 10194-10206.

Golub, G. H., & Van Loan, C. F. (1983). *Matrix Computations.* The Johns Hopkins University Press.

Herman, G. T. (1980). *Image Reconstructions from Projections.* Academic Press.

Herman, G. T., Hurwitz, H., Lent, A., & Lung, H. P. (1979). On the Bayesian approach to image reconstruction. *Information and Control* **42**, 60-71.

Hoerl, A. E., Kennard, R. W., & Baldwin, K. F. (1975). Ridge regression: some simulations. *Communication Statistics* **4**, 105-123.

Humphreys, E., & Clayton, R. W. (1988). Adaptation of back projection tomography to seismic travel time problems. *Journal of Geophysical Research* **93**, 1073-1085.

Humphreys, E., Clayton, R. W., & Hager, B. H. (1984). A tomographic image of mantle structure beneath southern California. *Geophysical Research Letters* **11**, 625-627.

Lanczos, C. (1961). *Linear Differential Operators.* D. Van Nostrand.

Lees, J. M., & Crosson, R. S. (1989). Tomographic inversion for 3-D velocity structure at Mount St. Helens using earthquake data. *Journal of Geophysical Research* **94** (B5), 5716-5728.

Menke, W. (1984). *Geophysical Data Analysis: Discrete Inverse Theory.* Academic Press.

Mosteller, F., & Tukey, J. W. (1977). *Data Analysis and Regression.* Addison-Wesley.

Nakanishi, I. (1985). Three-dimensional structure beneath the Hokkaido-Tohoku region as derived from a tomographic inversion of P-arrival times. *Journal of Physics of the Earth* **33**, 241-256.

Paige, C. C., & Saunders, M. A. (1982). LSQR: An algorithm for sparse linear equations and sparse least squares. *Transactions on Mathematical Software* **8**, 43-71.

Scales, J. A. (1987). Tomographic inversion via the conjugate gradient method. *Geophysics* **52**, 179-185.

Spakman, W., & Nolet, G. (1988). *Imaging algorithms, accuracy and resolution in delay time tomography.* Mathematical Geophysics. 155-187, D. Reidel Publishing Co.

Titterington, D. M. (1985). General structure of regularization procedures in image reconstruction. *Astronomy and Astrophysics* **144**, 381-387.

Trummer, M. R. (1981). Reconstructing pictures from projections: on the convergence of the ART algorithm with relaxation. *Computing* **26**, 189-195.

VanderSluis, A., & VanderVorst, H. A. (1987). *Numerical solution of large, sparse linear algebraic systems arising from tomographic problems.* Seismic Tomography. D. Reidel Publishing Co.

Willmott, C. J., Ackleson, S. G., Davis, R. E., Feddema, J. J., Klink, K. M., Legates, D. R., O'Donnell, J., & Rowe, C. M. (1985). Statistics for the evaluation and comparison of models. *Journal of Geophysical Research* **90**, 8995-9005.

Young, D. M. (1971). *Iterative Solution of Large Linear Systems.* Academic Press.