PAPER J

STATISTICS FROM STRINGS

Nicholas Smalley and Jerry M. Harris
Seismic Tomography Project

ABSTRACT

The computation and interpretation of higher order statistics for the imaging region in tomographic traveltime inversion is presented in this paper. Tomographic String Inversion allows for easy computation of several orders of statistics in the imaging process. These statistics can be used to make interpretations about the values which go into calculating the mean or first order statistic, and can be used to improve the convergence of the string inversion as well as the convergence slowness value.

INTRODUCTION

In traveltime tomography, the objective is to obtain as much information from the traveltime data as possible. Usually the slowness value that is assigned to individual square regions or pixels is a mean or optimum value that results from inversion of the matrix equation

$$\mathbf{A}\,\mathbf{s} = \mathbf{t}, \tag{1}$$

where **s** is the slowness vector consisting of all the square pixels, A is the projection matrix, and **t** is the traveltime vector consisting of all the raypath traveltimes. The raypaths that intersect each pixel affect the mean or optimum value of that pixel by varying degrees. Knowing the degree to which each raypath or traveltime affects the result might yield additional information about the region.

Statistically, many different distributions of numbers can yield the same mean or optimum value. Higher order statistics of the distribution, such as standard deviation, skewness, and kurtosis constrain the possible interpretation of the distribution. Knowing how each of the traveltimes contributes to the mean value can give us a distribution of

possible slownesses for each pixel. It is hoped that higher order statistics can give us some additional information about the region.

STRING TOMOGRAPHY

A recent paper (Harris 1990) in traveltime tomography allows for a direct method of statistical analysis of slowness values. Tomographic string inversion separates traveltime tomography into separate steps of inversion and imaging.

Inversion

The inversion process uses the raypaths as basis functions for the slowness field reconstruction. It is the assignment of slowness residuals to each individual raypath that allows for direct statistical analysis. The slowness residual for each raypath is determined by

$$\delta S = \frac{\Delta t}{L}, \tag{2}$$

where $\Delta t = t_{calculated} - t_{measured}$, L is the length of the raypath through the input (background) model, $t_{calculated}$ is the calculated traveltime through the input model, and $t_{measured}$ is the measured traveltime between the source and the receiver. By writing the definition of Δt as

$$\Delta t = \int_L \Delta S(r) dl , \tag{3}$$

where ΔS is the error of the slowness estimate in the input model, we see the residual slowness of the raypath is the average of the slowness error of the input model along the raypath

$$\delta S = \frac{1}{L} \int_L \Delta S(r) dl . \tag{4}$$

Imaging Statistics

The imaging process is now carried out separately from the inversion. Post inversion imaging areas will be refered to as bins. Each bin will have a number of raypaths intersecting it (Figure 1). Each of the residual slowness values represents a possible residual slowness value for that bin. The residual slownesses for each bin form the set of numbers to be used for statistical analysis. The statistics assigned to each bin come from raypaths which sample many other bins. Therefore the statistics contain information about the region of coverage, not just the imaging bin. Nevertheless, the set of statistics will have a localized weighting; regions close to and including the bin of measurement will have a stronger effect on the statistics than other bins. This is due to the greater sampling of bins near and including the bin of measurement (Figure 1).

FOUR ORDERS OF STATISTICS

The distribution of numbers or residual slownesses can be quantified in terms of four orders of statistics. These statistics are the mean (first order) , standard deviation (second order), skewness (third order), and kurtosis (fourth order). The mathematical definitions of each of these statistics are given in table 1. The mean and the standard deviation describe a set of numbers or residual slownesses that have a normal distribution (Figure 2). The skewness and kurtosis are quantities that describe deviations from a normal distribution.

The Mean

The mean value of a bin is calculated by the average of the slowness residuals corresponding to the raypaths which intersect the bin

$$\Delta S_c = \frac{1}{N} \sum_{i=1}^{N} \delta S_i , \tag{5a}$$

where

$$\delta S_i = \frac{1}{L_i} \int_{L_i} \Delta S(r) dl , \tag{5b}$$

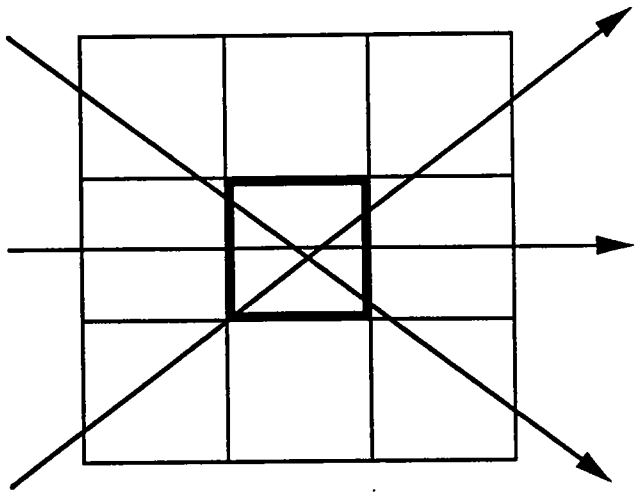

Figure 1. Raypaths intersecting a post inversion imaging area; a bin (solid box). The residual slownesses of the intersecting raypaths constitute the distribution of numbers for statisical analysis. The intersecting raypaths also contain information about other bins (lighter boxes).

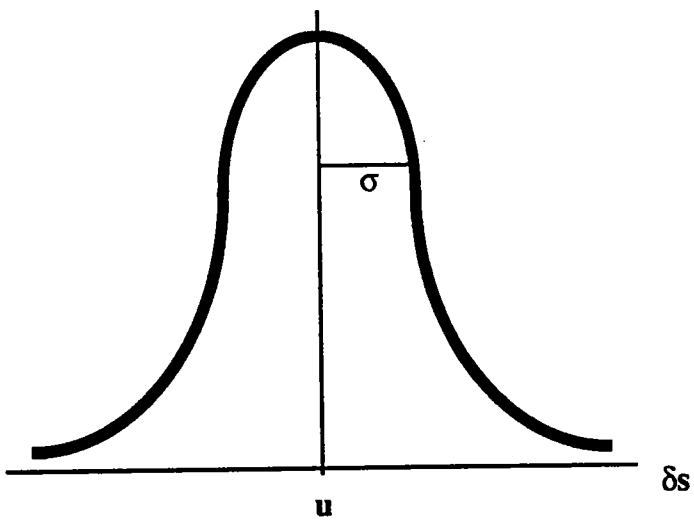

Figure 2. A normal distribution of data points. This type of distribution can be fully described by its mean, **u** , and its standard deviation, σ.

ORDER OF STATISTIC	MATHEMATICAL DEFINITION
FIRST ORDER (MEAN)	$S1 = \frac{1}{N}\sum_{I=1}^{N} S_I$
SECOND ORDER (STD. DEVIATION)	$S2 = \left[\frac{\sum_{I=1}^{N}(S_I - S1)^2}{N}\right]^{\frac{1}{2}}$
THIRD ORDER (SKEWNESS)	$S3 = \frac{\sum_{I=1}^{N}(S_I - S1)^3}{N \cdot S2^3}$
FOURTH ORDER (KURTOSIS)	$S4 = \frac{\sum_{I=1}^{N}(S_I - S1)^4}{N \cdot S2^4}$

Table 1. Four orders of statistics for a distribution of numbers.

N is the number of raypaths that intersect the bin, and L_i is the length of the i th raypath. The calculated mean serves as the approximation to the true residual slowness

$$\Delta S_e \cong \Delta S . \tag{6}$$

This estimate is added to the input slowness, S_o, to obtain an estimate for the slowness S,

$$S = S_o + \Delta S_e . \tag{7}$$

By adding S_o to each raypath residual slowness, statistics can be calculated for absolute slowness.

The Standard Deviation

The calculated standard deviation of a bin is a measure of the average deviation of the raypath slowness residuals that intersect the bin from the average raypath slowness residual. Since each slowness residual represents an average of the input slowness estimate error, ΔS , the standard deviation is an estimate of how ΔS varies throughout the region of coverage. Therefore the standard deviation is an estimate of the heterogeneity of the region of coverage relative to the input model; heterogeneity of ΔS.

The heterogeneity of ΔS in the region can give an indication of the error in the approximation of ΔS; $| \Delta S_e - \Delta S |$. A large source of error in traveltime inversion comes from the nature of the cross-well traveltime measurement. The slowness estimates for local regions (bins) are made from non-localized measurements. How greatly the non-localized measurements differ from each other can influence the accuracy of the estimate of ΔS.

Therefore the standard deviation might give an indication of the slowness estimate error; $| \Delta S_e - \Delta S |$. The logic sequence is

STANDARD DEVIATION => HETEROGENEITY ABOUT ΔS,

HETEROGENEITY ABOUT ΔS => ERROR IN ESTIMATE OF ΔS,

STANDARD DEVIATION => ERROR IN ESTIMATE OF ΔS,

where => symbolizes "implies".

Skewness

The skewness is a measure of the asymmetry of the distribution of slownesses for a bin (Figure 3). There are many geologic situations where a large number of raypaths intersecting a bin have many slownesses concentrated near one value, and a few slownesses with considerably smaller or larger values (Figure 4). These extreme values often will not carry useful information about the bin we are trying to image, yet these values can greatly influence the mean slowness value. Certain magnitudes of skewness might indicate the median instead of the mean should be used to estimate the slowness for a bin, since the median will not be affected by these extreme values.

Kurtosis

The kurtosis is a measure of the modality or bimodality of a distribution of slownesses for a bin (Figure 5) (Chissom 1970). It can tell us if the distribution is concentrated more closely around one or two different values of residual slowness. Geologically these two situations are distinct (Figure 6). This quantity can tell us about the homogeneity of the region of raypath coverage for a bin (Figure 1). An estimated slowness for a bin is more likely to be accurate if the slowness values of the intersecting raypaths are concentrated around one slowness value as opposed to two slowness values.

INVERTING FOR THE FOUR ORDERS OF STATISTICS

Synthetic measured traveltimes were computed from a model with a background slowness of 117.65 usec/ft (velocity = 8500 ft/sec), with four higher velocity layers in between (Figure 7). The first iteration of the string inversion used a constant slowness model of 117.65 usec/ft as the input model. The result of the first iteration is the first order statistic added to the input model, and is shown with the second, third, and fourth order statistics in figures 8a,b,c, and d respectively.

The inversion result shows resolution of the layers, with slowness errors within the layers. This is due to the limitations of the transmission measurement. Areas near the middle of the layers, particularly the thicker layers, are more accurate. The error between the inversion and the true model is shown in figure 9. When compared to the standard deviation in figure 10, we see a very small error in the slowness estimate for bins with standard deviations between 0 and 1. For higher standard deviations we see a gradual increase in error with increasing standard deviation. Overall the second order statistic

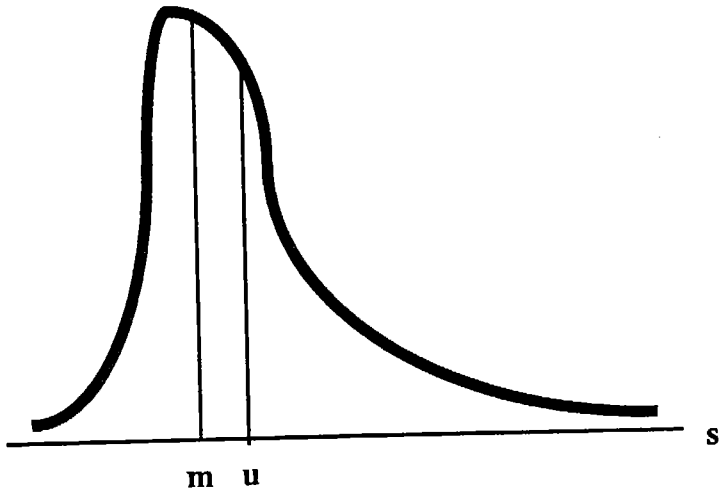

Figure 3. A skewed distribution of data points. This type of distribution is often better described by its median, **m**, rather than the mean. The mean , **u**, is strongly influenced by the few extreme values, while the median remains the same.

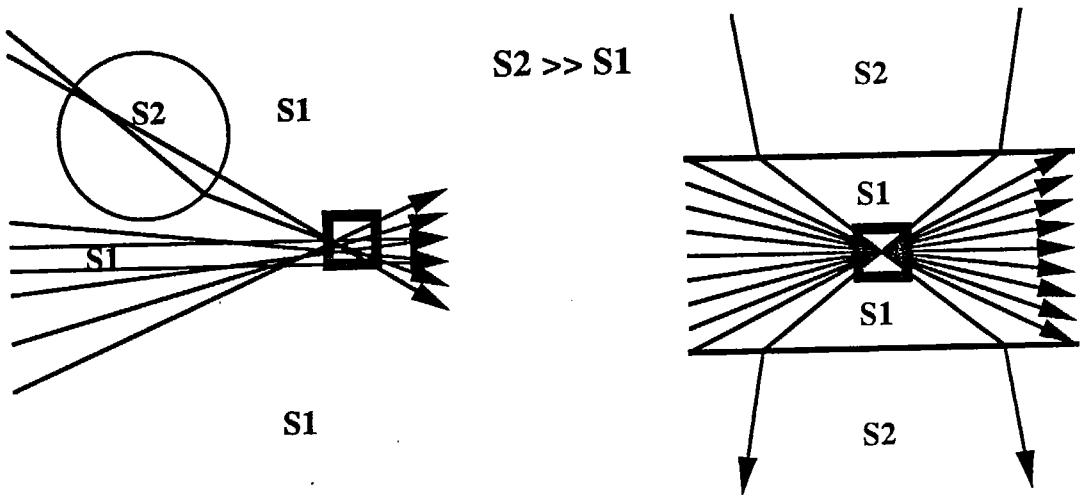

Figure 4. Two geologic situations that can yield high values of skewness for the imaging bin (dark box). Both will have values clustered around one value, with a few extreme values. The imaging bin on the left will have a negative skewness value, due to a few smaller values in the distribution. The imaging bin on the right will have a positive skewness value, due to a few larger values in the distribution. In both cases the median of the distribution will give a more accurate estimate of the slowness than the mean.

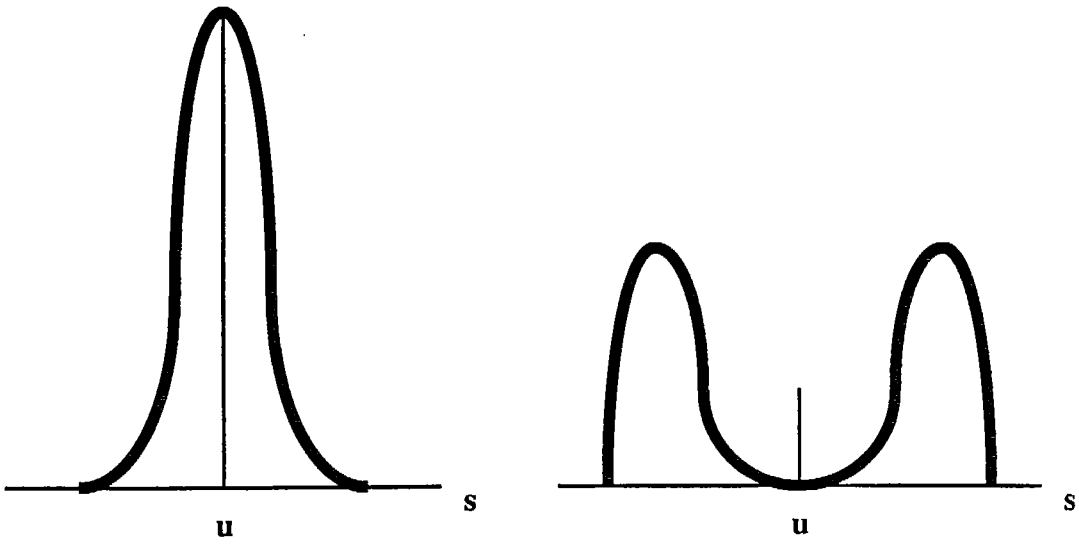

Figure 5. Two extreme values of kurtosis. The right graph is a unimodal distribution; representing a high value of kurtosis. The second graph is a bimodal distribution; representing a low value of kurtosis.

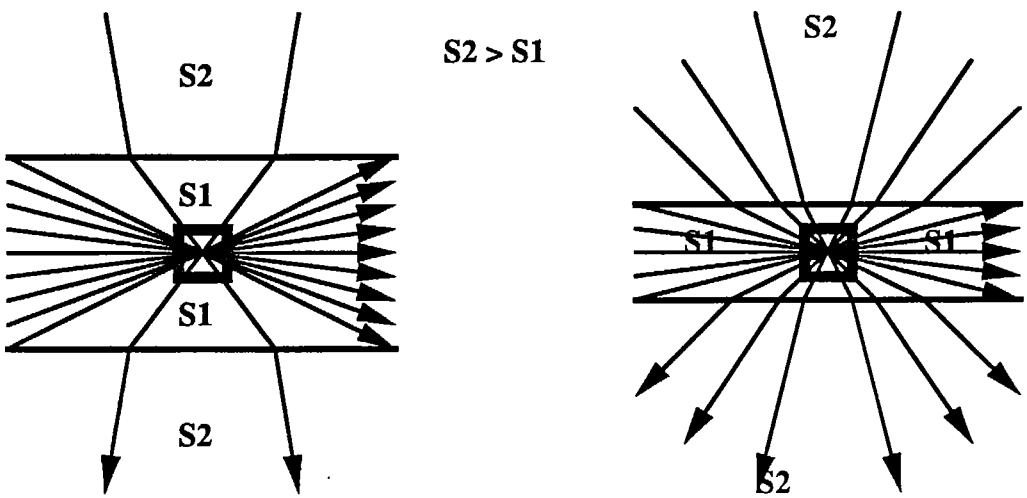

Figure 6. Two geologic situations that result in a large contrast in kurtosis values. The figure on the left has raypaths with a high concentration of slownesses at one value (unimodal), therefore having a large value of kurtosis. The figure on the right will have raypath slownesses clustered around both S1 and S2 (bimodal), therefore having a low value of kurtosis.

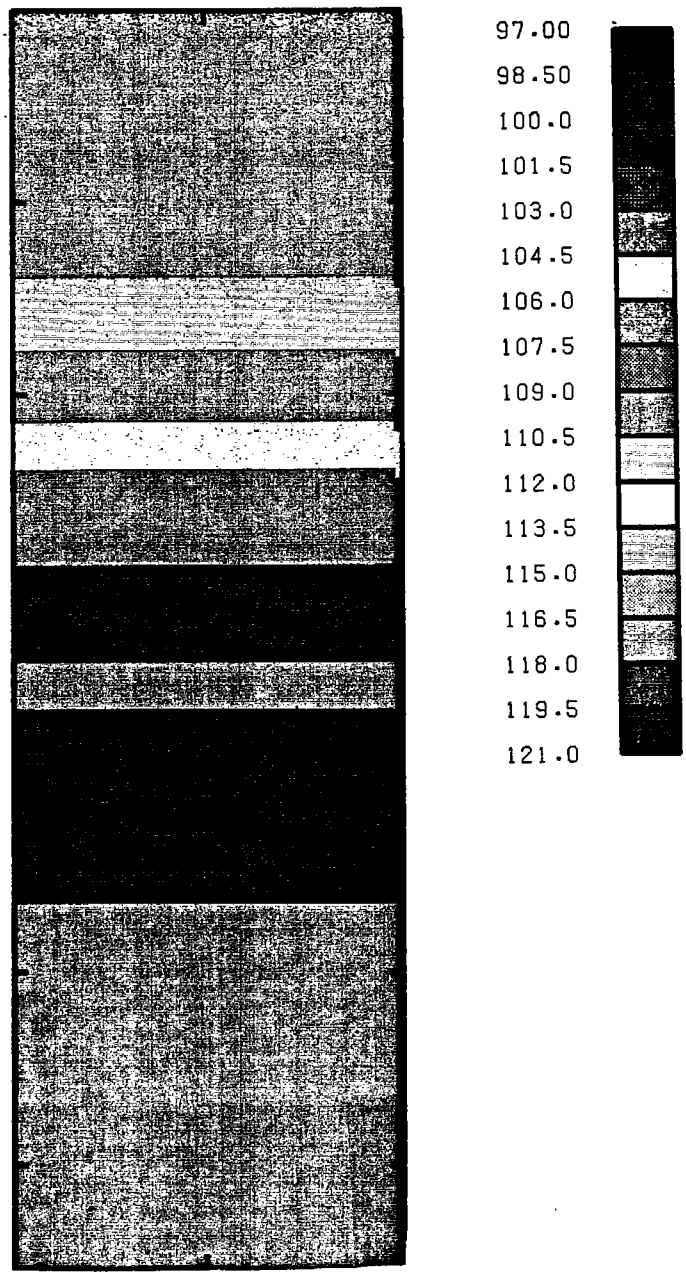

Figure 7. Slowness model used to generate synthetic measured traveltimes. Units are usec / ft.

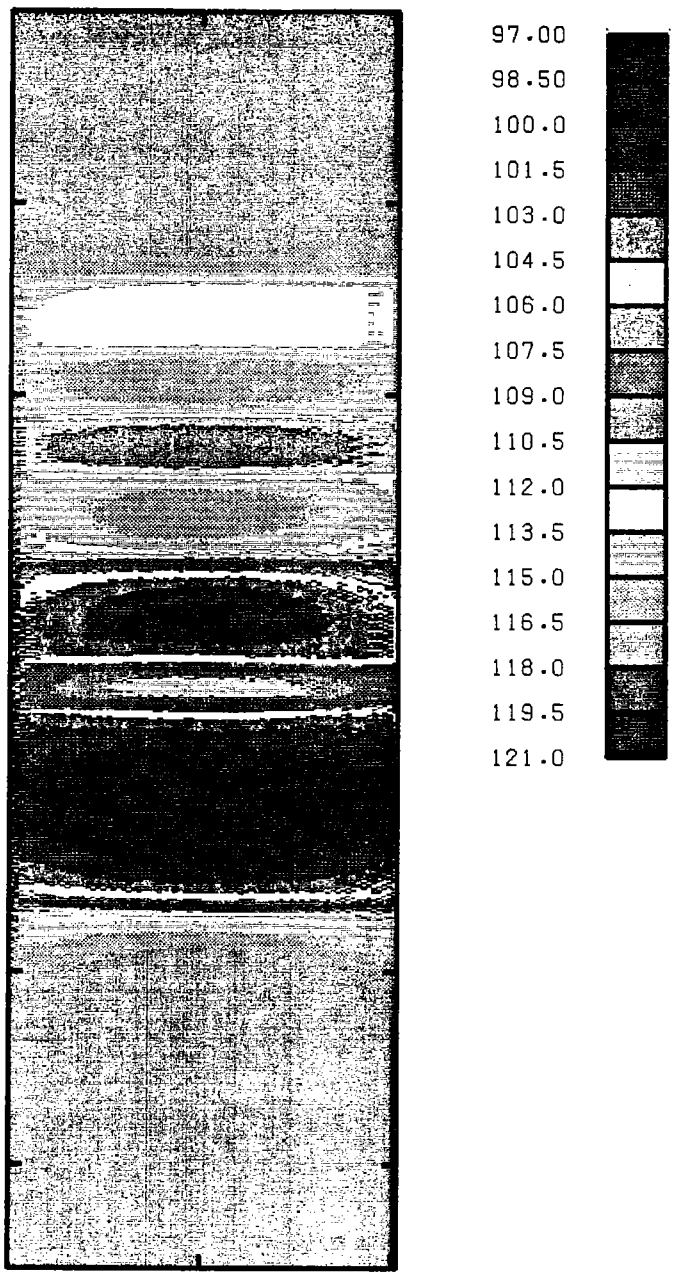

Figure 8a. Result of first iteration of string inversion. First order statistic plus input model. Units are usec / ft.

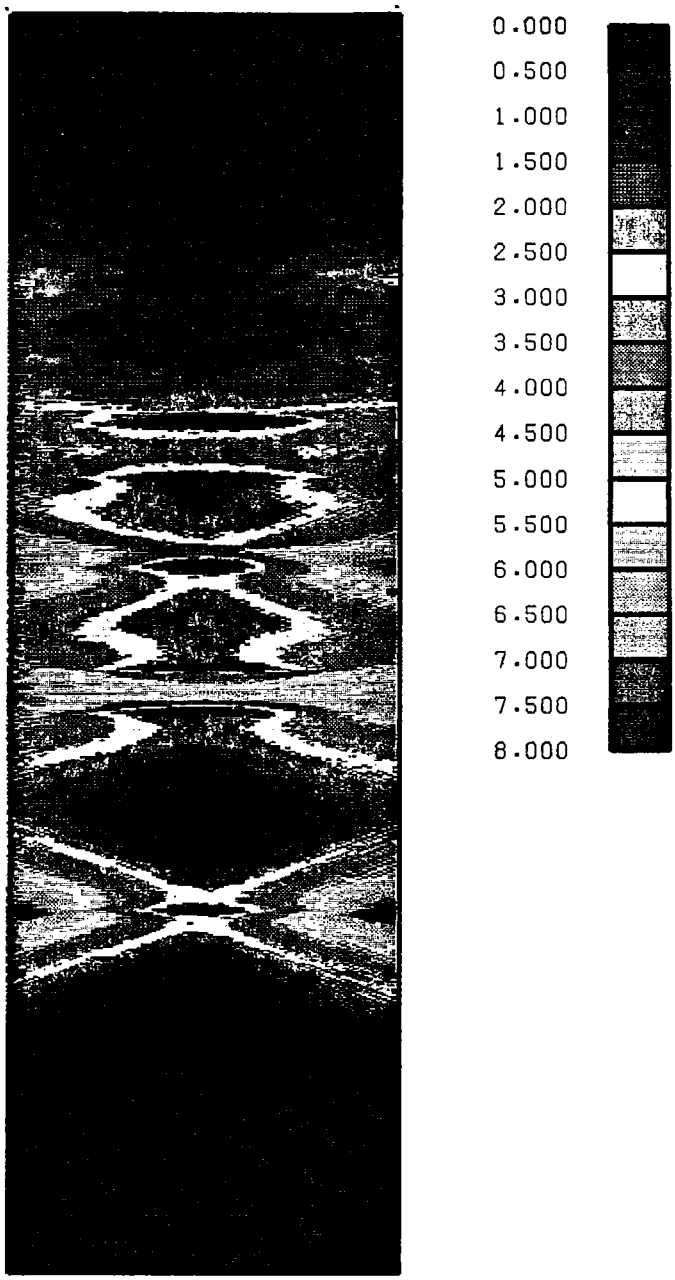

Figure 8b. Second order statistic. Units are usec / ft.

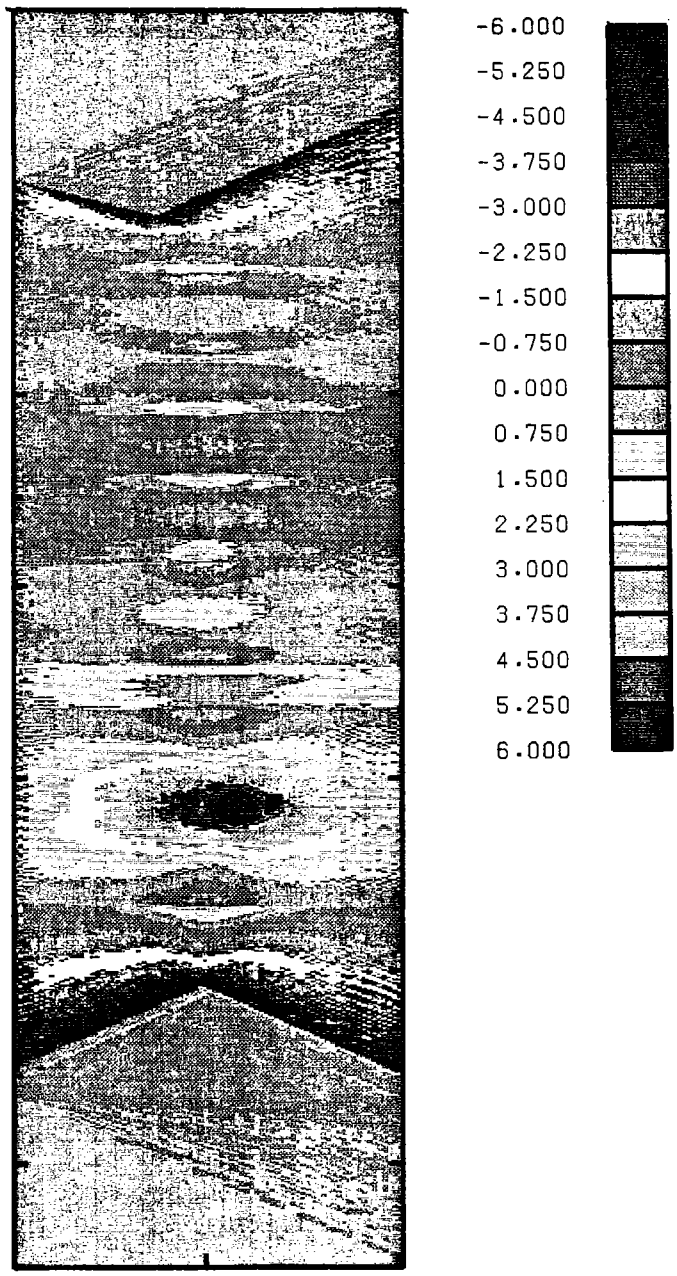

Figure 8c. Third order statistic.

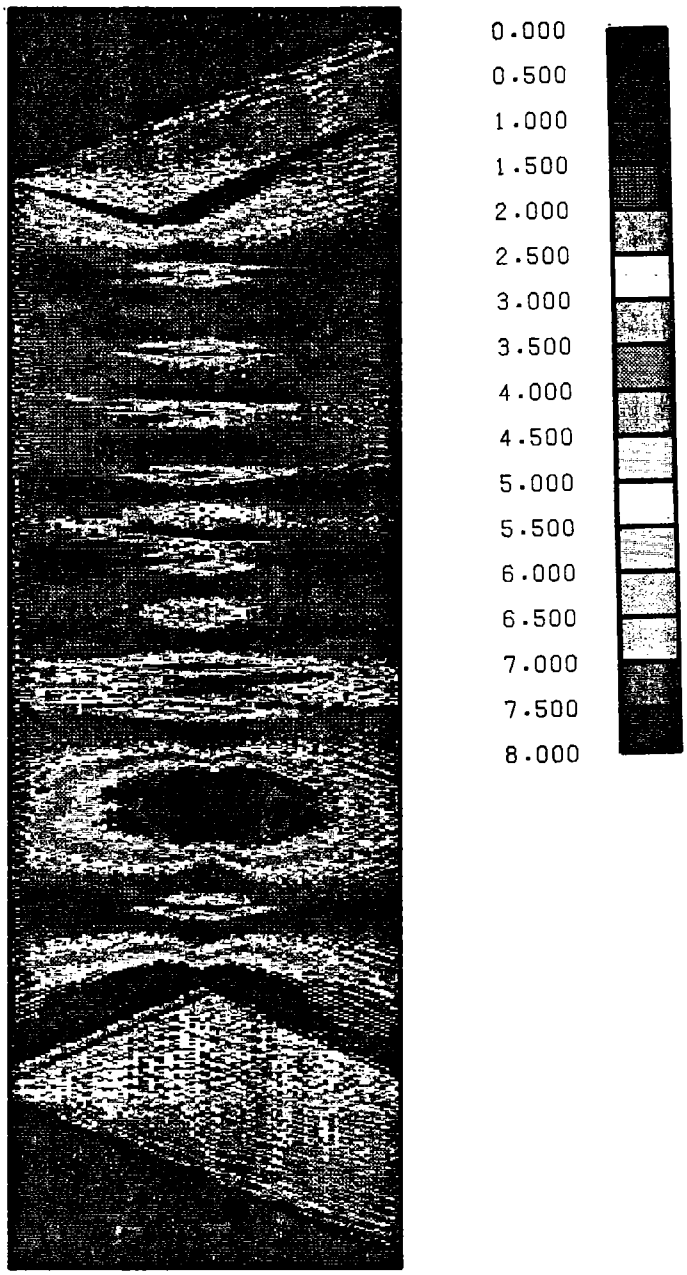

Figure 8d. Fourth order statistic.

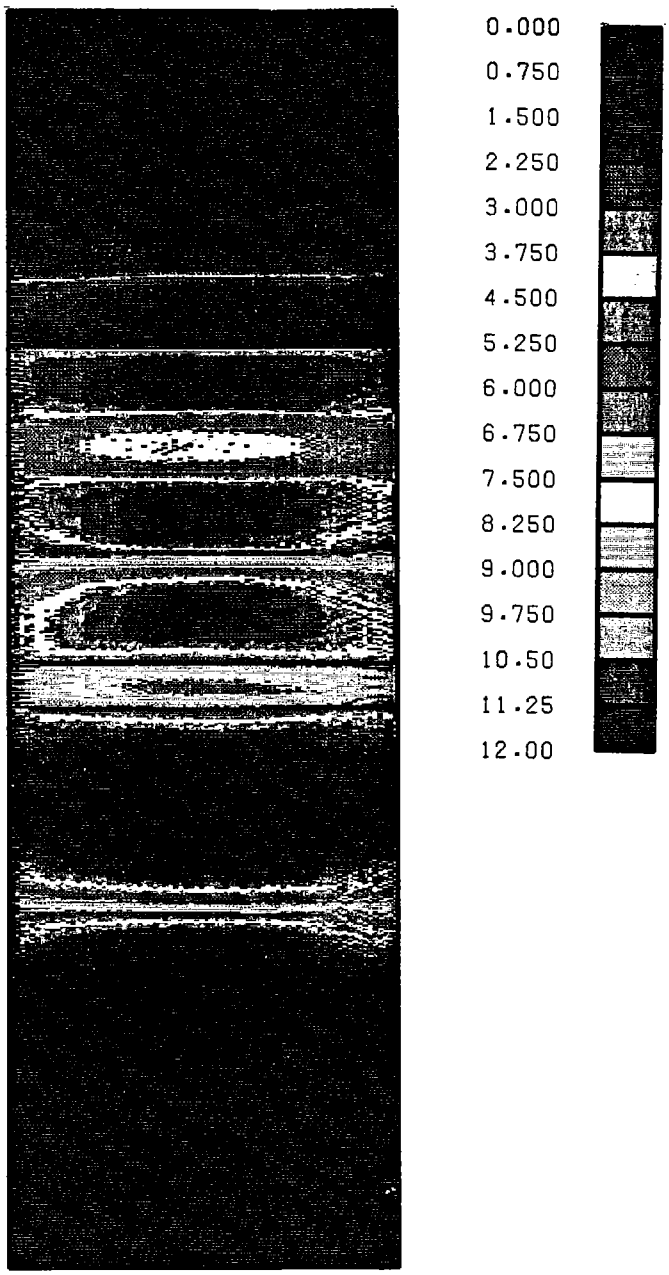

Figure 9. Error in slowness estimate after first iteration. Units are usec / ft.

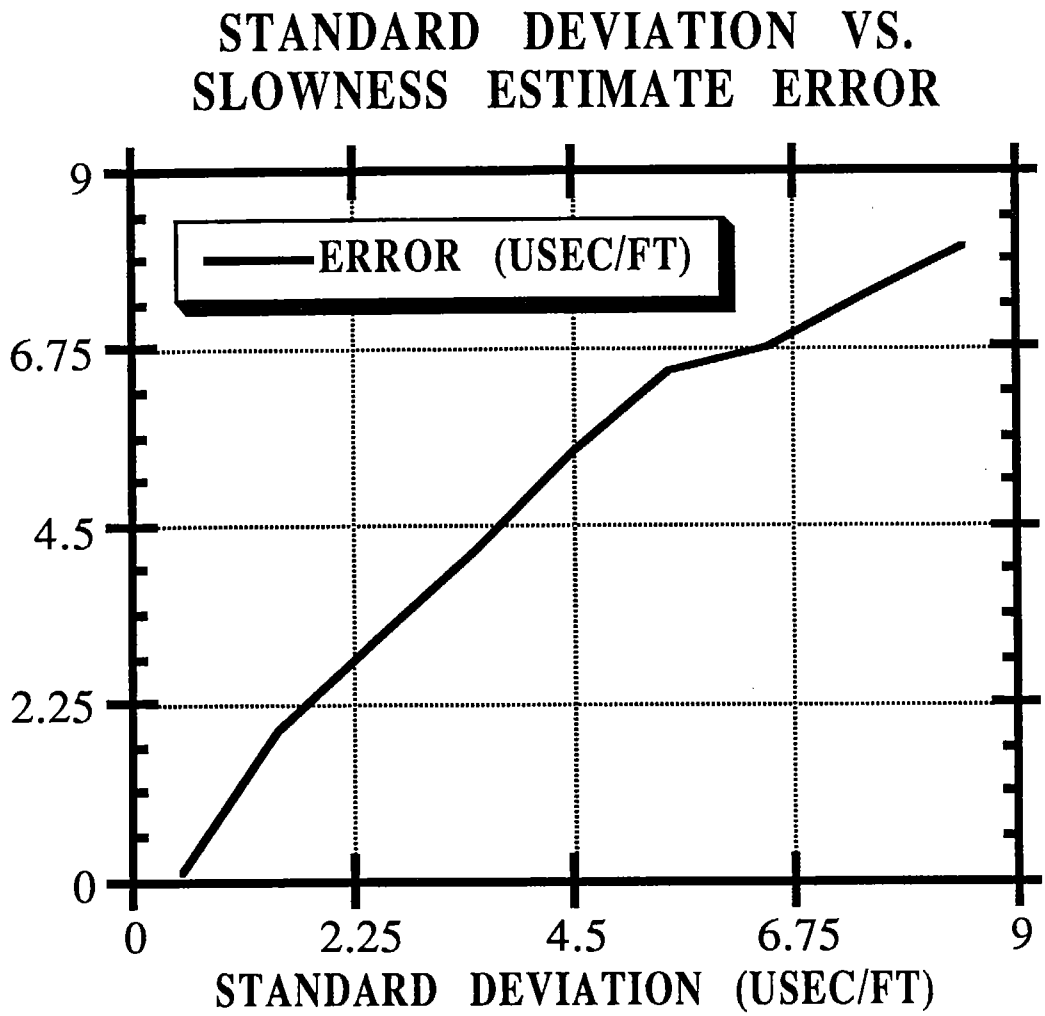

Figure 10. Standard deviation versus error. The error was calculated by averaging the errors of standard deviations within intervals of 1 usec / ft.

shows a good correlation with the error except near the middle of layer boundaries.

The third order statistic shows contrasts within layers and on the edges of boundaries. The low slowness (high velocity) layers show an increase in positive skewness towards the center. The high slowness layers (low velocity layers resulting from the input of high velocity layers in a lower velocity background) show an increase in negative skewness toward the middle. The low slowness layers will have raypaths or strings that sample areas outside the layer with higher slownesses. These raypath slownesses will be larger than the raypath slownesses that only intersect the layer of constant low slowness. This results in positive skewness with many values concentrated below the mean value and a few larger values. The opposite is true for high slowness layers resulting in negative skewness. This results in sign reversals of the skewness at the boundaries of some of the layers.

The fourth order statistic also shows transitional behavior at the boundaries and within layers. Within thin layers the kurtosis values decrease toward the middle. The thicker layers show an increase in kurtosis towards the middle. The middle of thinner layers will have raypath slownesses that are concentrated around the surrounding medium slowness and the layer slownesses (two values). The middle of thicker layers will have raypath slownesses concentrated around the layer slowness (one value). Due to the symmetry of the raypaths at the middle of layer boundaries, there will be many raypath slownesses clustered around one value. Therefore bins at these locations also have high kurtosis values.

STRING INVERSION USING THE STANDARD DEVIATION

Previously it was stated the magnitude of the standard deviation might be an indicator of the error in the slowness estimate. This would suggest the standard deviation might be used to improve subsequent iterations of the inversion. The original string inversion algorithm calculates the slowness residual from equation (2), and backprojects this correction evenly along the raypath. This can have the effect of backprojecting residual slownesses into bins that already have the correct or nearly the correct slowness value. A new algorithm was written that uses a weighted residual slowness along the raypath. The weight is a function of the standard deviation of the bins along the raypath. The average

slowness residual for the raypath was calculated using equation (2), and then was weighted as a function of the bin along the raypath using the formula,

$$\delta S_{ij} = \frac{\sigma_j^p}{\sigma_i^p} \delta S_i , \tag{8a}$$

where σ_j^p is the standard deviation to the p th power of the j th bin, σ_i^p is the average standard deviation to the p th power along the i th raypath,

$$\sigma_i^p = \frac{1}{L_i} \int_{L_i} \sigma^p (r) dl , \tag{8b}$$

δS_i is the residual slowness for the i th raypath, and δs_{ij} is the backprojected slowness as a function of the raypath and bin. The new backprojection formula has the benefits of

1) backprojecing less residual into bins with small standard deviations.

2) backprojecting more residual into bins with large standard deviations.

3) honoring the traveltime residual.

The power p in equations (8a), (8b) depends on the relationship between the standard deviation and the error in the slowness estimate. There is no exact theoretical expression for this relationship. The relationship between standard deviation and error can change as the number of iterations increases. If p is chosen to equal 1, this would backproject the slowness by assuming the standard deviation was linearly related to the error. Since standard deviation has the linear units of slowness, and is written as a linear error or deviation to the mean slowness residual, p is chosen to equal 1 for the second iteration. Continuing to use this assumed linear relationship in subsequent iterations can cause poor results in certain regions. This is due to two problems:

1) There are bins after the first iteration that have a low standard deviation but a relatively high error.

2) As the number of iterations increases, the residuals of the strings converge to certain values, often very small values. Therefore the standard deviations will become smaller. This will be true for some bins even though the errors might be approaching values that do not correlate with the small standard deviations.

Certain types of symmetry of the true inversion region relative to the input model can yield bins which have small standard deviations, but relatively large errors. All of the residual slownesses of raypaths that intersect a bin might have similar values, however the mean of these values might be quite different from the true error of the input model, ΔS . This is due to the limitations of the transmission measurement; the calculation of localized information based upon non-localized measurements. As the number of iterations increases, we are still limited by the type of measurement. This can be illustrated by equation (4)

$$\delta S = \frac{1}{L} \int_L \Delta S(r) dl .$$

As δS converges, the values of ΔS may or may not converge to the same value, since δS is an average of $\Delta S(r)$ along the raypath. Therefore as the number of iterations increase, there is the potential for an increasing number of bins to have standard deviations that are not representative of the slowness estimate error. Thus, for some bins the implication that the magnitude of the standard deviation is related to the magnitude of the slowness estimate error is violated.

Therefore by keeping p equal to one in the subsequent iterations, some bins with relatively large errors will not be corrected. This would suggest backprojecting the slowness residuals more uniformly along the raypath for subsequent iterations. If we set

$$p = \frac{1}{n-1}, \tag{9}$$

where n is the iteration number, the slowness residuals will be backprojected more uniformly along the raypath as the number of iterations increases. The average absolute

error in slowness, computed by

$$\text{AVEERR} = \frac{1}{M} \sum_{j=1}^{M} \left| S_T - S_E \right|_j , \qquad (10)$$

where $| S_T - S_E |_j$ is the magnitude of difference between the true slowness and the estimated slowness for the j th bin, and M is the total number of bins, versus iteration number is shown for p=0 (original string algorithm - uniform backprojection), p = 1, and p = 1 / n -1 in figure 11, and tabulated in table 2. The slowness estimate errors for each inversion path after 5, 9, and 18 iterations are shown in figures 12a,b,and c. The p = 1 / n -1 path showed improved convergence in the slowness estimate, and an improved convergence value. For example, the 12 th iteration of the p = 1 / n -1 path has a lower average error than the 18 th iteration of the p = 0 path. The convergence value for the p = 1 / n -1 path value is a 2.3% improvement in the error over the original strings algorithm. The convergence value was defined at the iteration before subsequent iterations stopped showing any improvement. The convergence value occured at iteration 18 for both the original string algorithm and the p = 1 / n -1 path. Figure 11 and table 2 also show for the first 6 iterations the p = 1 path provides improvement over the p = 0 path, but has a larger error for subsequent iterations.

CONCLUSION

The usefulness of string inversion theory has been demonstrated in the easy calculation of higher order statistics for the inversion region. It has shown that each of the four orders of statistics can provide us with certain information about the region of interest. It was also shown that the standard deviation can be used to provide a better convergence of the string algorithm and an improved convergence value. More research needs to be done using the third and fourth order statistics to obtain information about the inhomogeneity of the medium and improving the inversion. More research needs to be done on the relationship between the standard deviation and the slowness estimate error.

ACKNOWLEDGEMENTS

The authors would like to thank The Gas Research Institute for supporting this work. The first author would like to thank the entire STP group for helpful comments and suggestions.

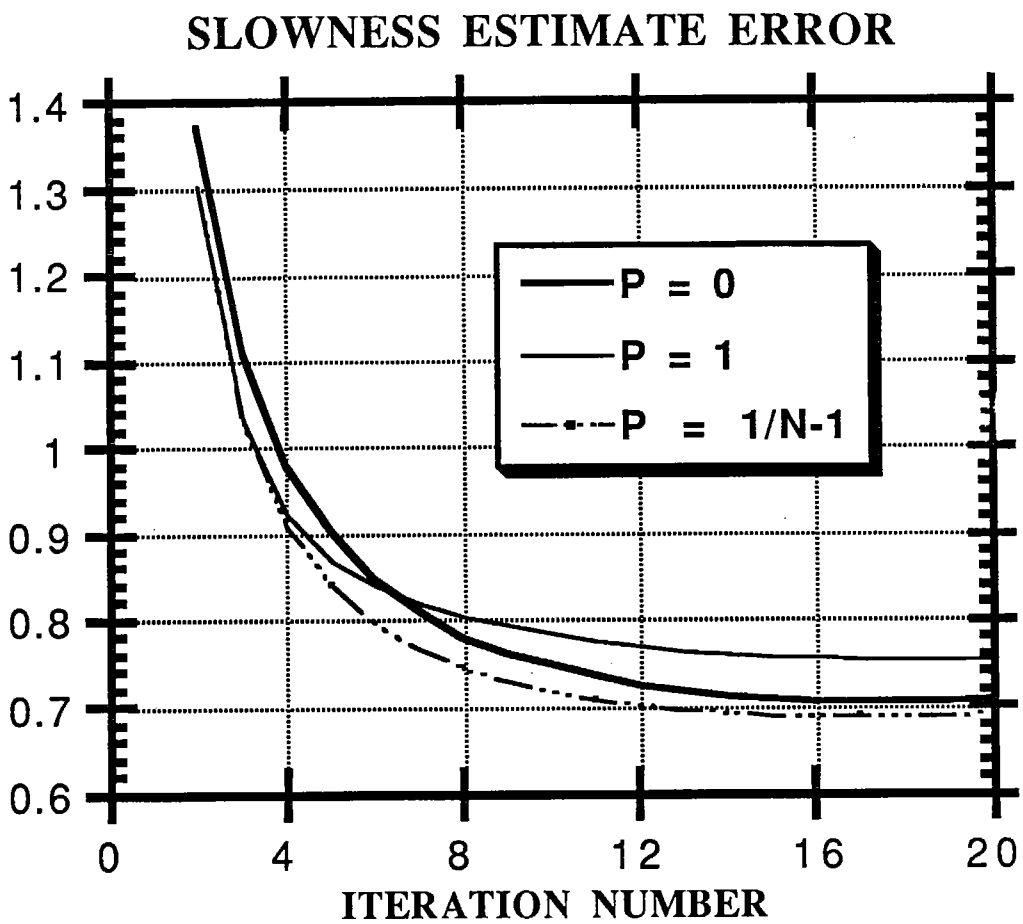

Figure 11. Average slowness estimate error for three paths of inversion. Units are usec / ft.

SLOWNESS ESTIMATE ERROR

ITERATION NUMBER (n)	P = 0 (USEC / FT)	P = 1 (USEC / FT)	P = 1 / n - 1 (USEC / FT)
2	1.3727	1.3030	1.3030
4	0.97961	0.92322	0.91013
6	0.84815	0.84021	0.79643
8	0.78089	0.80531	0.74503
10	0.74747	0.78520	0.71755
12	0.72489	0.76951	0.70239
14	0.71237	0.76052	0.69416
16	0.70651	0.75703	0.68993
18	0.70541	0.75599	0.68864
20	0.70735	0.75392	0.69197

Table 2. Average slowness estimate error for three paths of inversion. Units are usec / ft.

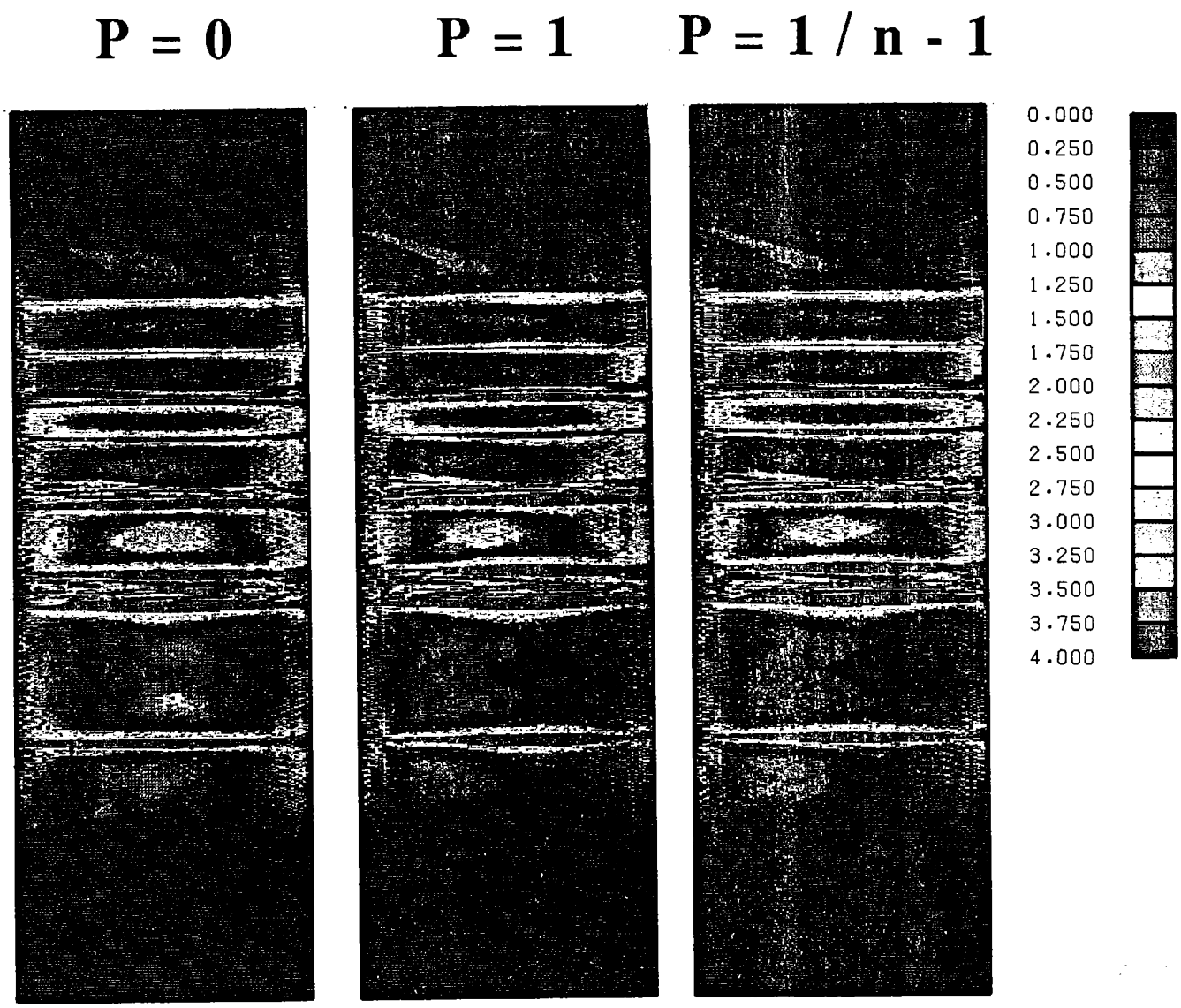

Figure 12a. Slowness estimate error after 5 iterations. Units are usec / ft.

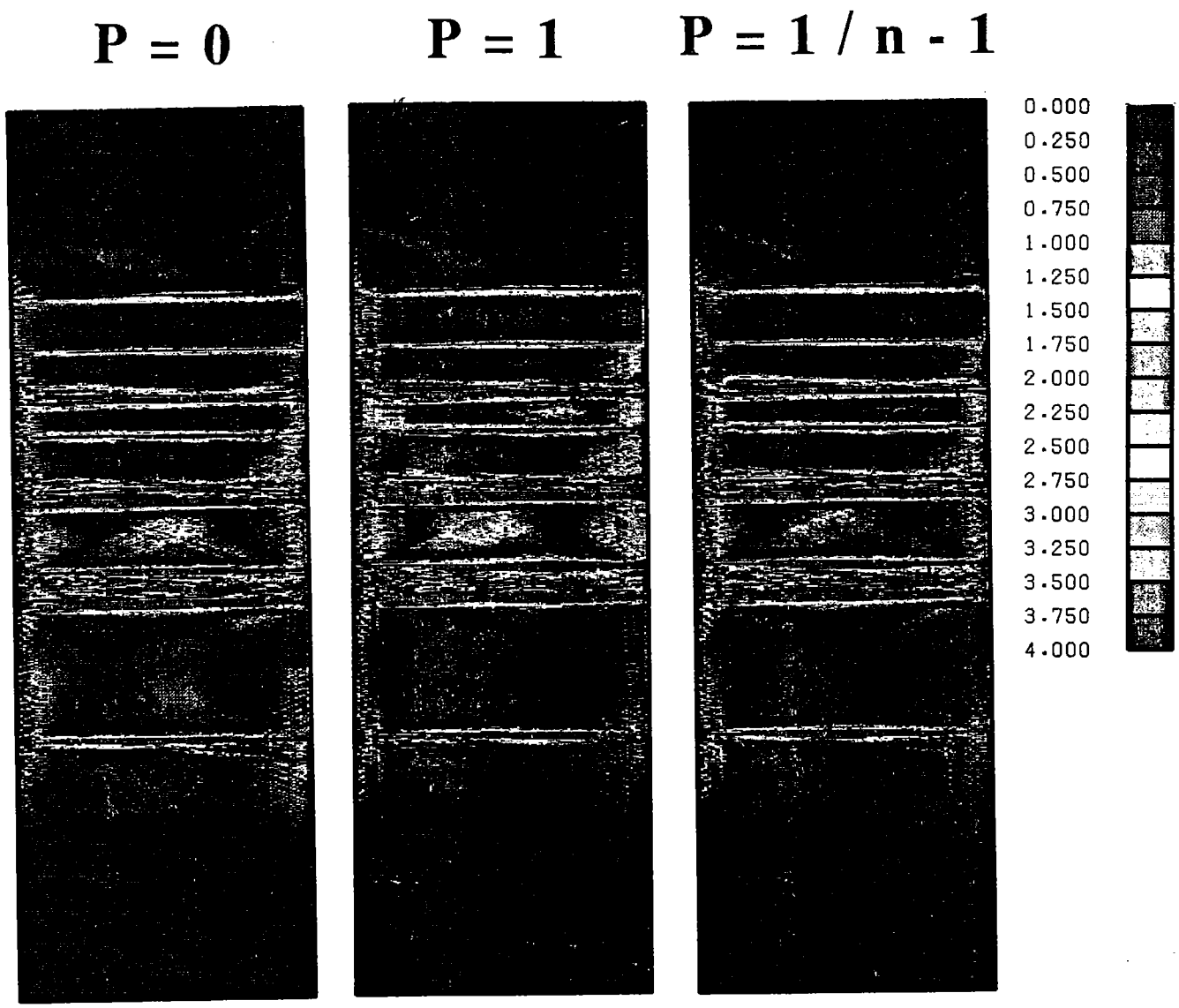

Figure 12b. Slowness estimate error after 9 iterations. Units are usec / ft.

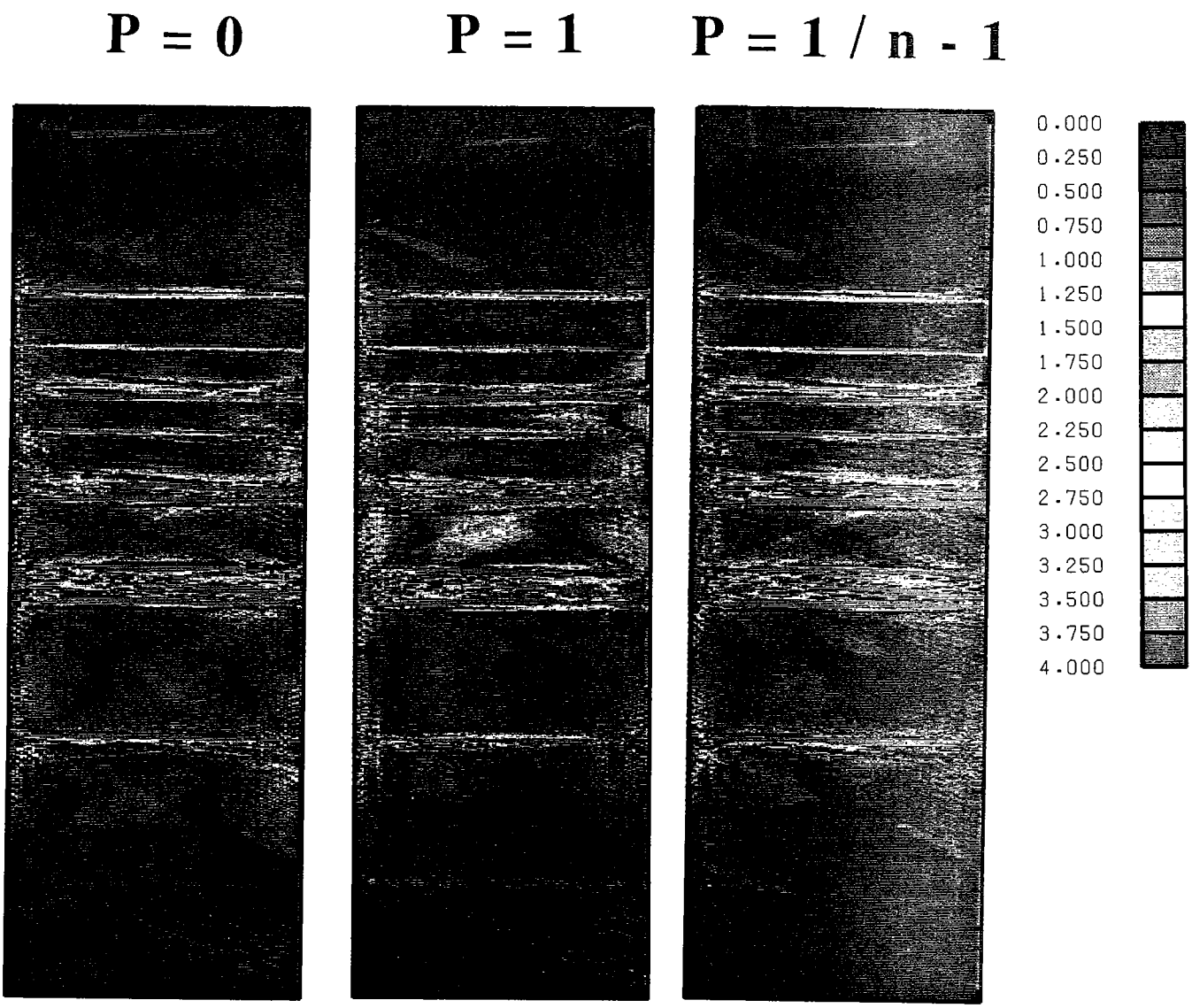

Figure 12c. Slowness estimate error after 18 iterations. Units are usec / ft.

REFERENCES

Bjerhammar, A., 1973, Theory of Generalized Matrix Inverses, Elsevier Scientific Publishing Company.

Chissom, B. S., Oct. 1970 Interpretation of the Kurtosis Statistic, The American Statistician, **24**, pp. 19-22.

Harris, J.M., Lazaratos, S., Michelena, R., 1990, Tomographic String Inversion, STP volume 1, paper B.

Sachs, L., 1984, Applied Statistics, Second Edition, Springer-Verlag.