Restoration of Missing Data by Least Squares Optimization

Jon F. Claerbout

The quality of seismic data analysis is frequently degraded by missing data. The problem can be seismic traces missing from the ends of seismic sections or from the ends of CMP gathers. It can also be spatial aliasing (interlaced missing traces) or randomly missing traces.

Ordinarily the problem of missing data is glossed over by the insertion of zeros for unknown data values. Where this causes unsatisfactory diffraction effects the data is often tapered smoothly to zero. Data tapering must be regarded as data falsification; a crude expedient. Philosophically it must be inferior to finding extensions to the data set (padding the data) which cause the processed data to be more satisfactory in some subjective way.

Least squares optimality is usually philosophically inferior to optimality in some other sense, such as entropy or the L_1 norm. But least squares has a deserved reputation for computability. The purpose of this article is to illustrate that least squares optimization can be computationally practical even when the number of unknowns is on the order of hundreds of traces, *i.e.* hundreds of thousands of unknowns. Likewise, the number of constraints, the entire set of observed data, is also a very large number, commonly in excess of 48,000, often millions.

Needed: A Transformation Between Data Space and Model Space

Prerequisite to the techniques of this paper is the existence of an invertible transformation between data space and model space. There are several ready examples of such transformation pairs. Conceptually the simplest is Fourier transformation and its inverse. Other examples are: the upward and downward wavefield extrapolation

operations; migration and diffraction; slant stack and inverse slant stack. New invertable transformations are being developed by concurrent research for transformation between the data space of a common midpoint gather and a model space of a velocity spectrum. In the practical cases considered so far the transformation pairs are unitary matrices, or approximately so.

The Optimality Criterion

In model space we have some clear ideas of what the earth does *not* look like. We doubt the existence of the semi-circular reflectors so often predicted by migration programs. We doubt the imaginary velocities (negative V^2) predicted by velocity analysis programs. Sometimes these unlikely implications of our data may be suppressible only by data falsification. This question of decomposition of the observations into true values plus noise based on acceptability of the implied model is a very deep and hazardous question. Until we achieve more technical skill than demonstrated in this paper, we will take all observations to be perfect. Only the missing data will be chosen to provide the most acceptable model.

It is easy to hold a subjective opinion about what constitutes a bad model, but not always so easy to find an optimality condition which will suppress such poor models. In the least squares framework the choice of what to optimize amounts to the choice of a weighting function. Before further discussion of the choice we will define the basic computational technique and examine some examples.

Formulation and Procedure

The data space $(\mathbf{x},\mathbf{r})^*$ is composed of two parts, the raw data $\mathbf{r}$, and values $\mathbf{x}$ to be placed in gaps. Usually there is natural ordering within the data space. The data may be two dimensional and the gaps may be interspersed arbitrarily within the data or off the ends. But for our present discussion the structure of the data space is ignored. The data space is mapped into a column vector with the known data $\mathbf{r}$ in the bottom part of the vector and the unknown part, the padding $\mathbf{x}$, in the top. Next we have a matrix premultiplier for the data vector to transform it into model space. The matrix is partitioned into a part $\mathbf{A}$ which multiplies $\mathbf{x}$ and a part $\mathbf{B}$ which multiplies $\mathbf{r}$.

$$\begin{bmatrix} \mathbf{A} & \mathbf{B} \end{bmatrix} \begin{bmatrix} \mathbf{x} \\ \mathbf{r} \end{bmatrix} \tag{1}$$

Once in model space we wish to form a weighted quadratic form. This is done by premultiplying (1) by its transpose, placing in the middle a diagonal matrix of weights **W**.

$$[\,\mathbf{x}^* \;\mathbf{r}^*\,]\begin{bmatrix}\mathbf{A}^* \\ \mathbf{B}^*\end{bmatrix}[\mathbf{W}][\mathbf{A}\;\mathbf{B}]\begin{bmatrix}\mathbf{x} \\ \mathbf{r}\end{bmatrix} \tag{2}$$

The real and the imaginary parts of $\mathbf{x}$ may be regarded as independent variables. Likewise $\mathbf{x}$ and $\mathbf{x}^*$ may be regarded as independent. Setting to zero the derivative of the quadratic form with respect to $\mathbf{x}$ gives a set of equations which is conjugate to those of setting to zero the derivative with respect to $\mathbf{x}^*$. Hence we may ignore either set. I like to consider only the set obtained by setting to zero the differential with respect to $\mathbf{x}^*$.

$$0 = [\mathbf{A}^*][\mathbf{W}][\mathbf{A}\;\mathbf{B}]\begin{bmatrix}\mathbf{x}+\mathbf{dx} \\ \mathbf{r}\end{bmatrix} \tag{3}$$

Reorganizing

$$[\mathbf{A}^*][\mathbf{W}][\mathbf{A}][\mathbf{dx}] = -[\mathbf{A}^*][\mathbf{W}][\mathbf{A}\;\mathbf{B}]\begin{bmatrix}\mathbf{x} \\ \mathbf{r}\end{bmatrix} \tag{4}$$

This equation outlines the computational algorithm. Begin with padding $\mathbf{x}$ for the data set being arbitrary, commonly zero. Compute the right hand side, a vector. Solve, or approximately solve, a set of simultaneous equations for $\mathbf{dx}$. Update $\mathbf{x}$ to $\mathbf{x}+\mathbf{dx}$. Iterate.

The left hand side matrix $\mathbf{A}^*\mathbf{W}\mathbf{A}$ is generally far too large to be inverted, or even stored in a computer. Practical problems are usually so large that exact matrix inversion is out of the question. Some kind of approximation is necessary. First note that any such approximation on the left does not affect the final solution $\mathbf{x}$ which is attained. This is because once iteration has proceeded to convergence, $\mathbf{dx} = 0$, it is quite immaterial what matrix stands to the left of $\mathbf{dx}$ in (4). The fact that $\mathbf{dx}$ vanished is a consequence entirely of the right hand side.

Next recall that the basic operators are usually unitary. For example, $\mathbf{A}$ could contain selected columns from the Fourier transform matrix. These elments are orthogonal to their counterparts in $\mathbf{A}^*$. So if the weighting function $\mathbf{W}$ is smoothly variable, it is natural to expect that $\mathbf{A}^*\mathbf{W}\mathbf{A}$ will be very close to a diagonal matrix. In my work so far I have chosen $\mathbf{W}$ to be of such a magnitude that I usually approximate $\mathbf{A}^*\mathbf{W}\mathbf{A}$ by an identity matrix. This has often given good results. But not always so. I experimented with changing the scale factor and also band matrix approximations, but

these experiments were rather specialized and gave no generally useful conclusion. Some applications seem to demand further effort with this matrix. Fortunately we may expect to see a reasonable amount of literature on the subject in the field of medical imaging. It is a subject to return to.

Treating the left hand matrix as a identity matrix, I found it particularly appealing to think of the algorithm in the following form:

$$(\mathbf{x}+\mathbf{dx},\mathbf{r}) \;=\; (\mathbf{x},\mathbf{r}) - \mathit{Select}\left[FT^{-1}\left[\mathit{Badpass}\left[FT(\mathbf{x},\mathbf{r})\,\right]\,\right]\right] \tag{5}$$

For example, suppose it is desired to add traces $\mathbf{x}$ to the side of a seismic section in such a way as to avoid unnecessary creation of evanescent energy. Initially, the traces $\mathbf{x}$ could be zero. First, the data space $(\mathbf{x},\mathbf{r})$ is two dimensionally Fourier transformed. A weighting operator *Badpass* passes evanescent energy with unit magnitude and all other frequencies with zero magnitude. After inverse transforming we have a proposed $(-\mathbf{dx},-\mathbf{dr})$. We select the $\mathbf{dx}$ and abandon the proposed perturbation to the observed data $\mathbf{dr}$. For this particular problem, iteration has been found to lead to rapid convergence (and uninspiring extrapolations of the data set).

Gaps in a Time Function

A simple one dimensional illustration of the foregoing concepts arises with gaps in a time function. Filling the gaps is a classical interpolation problem in which there is much previous experience. We are not trying to improve on previous methods here. We are just trying out new methods on an old problem.

As a test case I selected a far offset trace from a marine seismic profile. Before discarding alternate time points and also points to make up some larger gaps, the power spectrum was computed. Larger gaps were made by discarding sequential points, leaving three different sized gaps, one half wavelength, one wavelength, and two wavelengths.

The first concept for attempting to restore the missing gaps was this: Filling the gaps with zeros produced a spectrum which contained much more high frequency and much more low frequency energy than a seismogram ordinarily has. The high frequency comes from the sharp corners at the edge of the gap. The low frequency could come from a shift of the mean level owing to the unlikelihood that the lost points had exactly zero mean. The idea that these high and low frequencies should have low power in the final spectrum led to the idea of minimizing a weighted power in the final spectrum. The weight is large (≈ 1) at high and low frequencies and tapers smoothly to

zero in between. The results are shown in figure 1.

A second example of the same type makes much more use of prior knowledge of typical seismic spectra. The weight function was chosen to be

$$\mathbf{W} = \underset{\omega}{\mathbf{diag}} \frac{1}{1 + 50 \dfrac{\overline{D}(\omega) D(\omega)}{\dfrac{1}{N} \sum_{\omega} \overline{D}(\omega)\, D(\omega)}} \tag{6}$$

Here $D(\omega)$ is the discrete Fourier transform of length N of the original data before points were dropped off to make gaps. In practice you would have to use the power spectrum of a nearby seismogram. Notice that for typical values of the power spectrum, this weighting function is the *inverse* of the power spectrum. Thus frequencies which are very well represented in the original data are very weakly discriminated against by the power minimization condition. The purpose of the number 50 is to provide a floor under the spectrum so the inverse will not blow up, or exceed the unity which is desired by the iterative procedure. Comparing figure 2 to figure 1, we see that the criterion (6) is more permissive in the gaps.

Figures 3 to 6 illustrate the same concepts applied to a common midpoint gather. Here the basic idea is that any reasonable gather should focus when downward continued. The focus need not be a very narrow one, but there should not be energy left out at wide offsets. Any energy which is found at wide offsets after downward continuation is indicative of data not fitting the simple wave propagation model. For example, spatial aliasing and truncation of the data set at the inside and outside ends of the cable will cause the downward continued data to be non-zero at wide offsets.

The procedure to extend and interpolate the gather proceeds analogously to equation (5). Replacing the Fourier Transform operation in (5) by downward continuation (DC) and the inverse Fourier Transform by upward continuation (UC), we have

$$(\mathbf{x}+\mathbf{dx},\mathbf{r}) = (\mathbf{x},\mathbf{r}) - Select \left[UC \left[Badpass \left[DC(\mathbf{x},\mathbf{r}) \right] \right] \right] \tag{7}$$

In this equation the *Badpass* operator is thought of as zeroing the good information near zero offset.

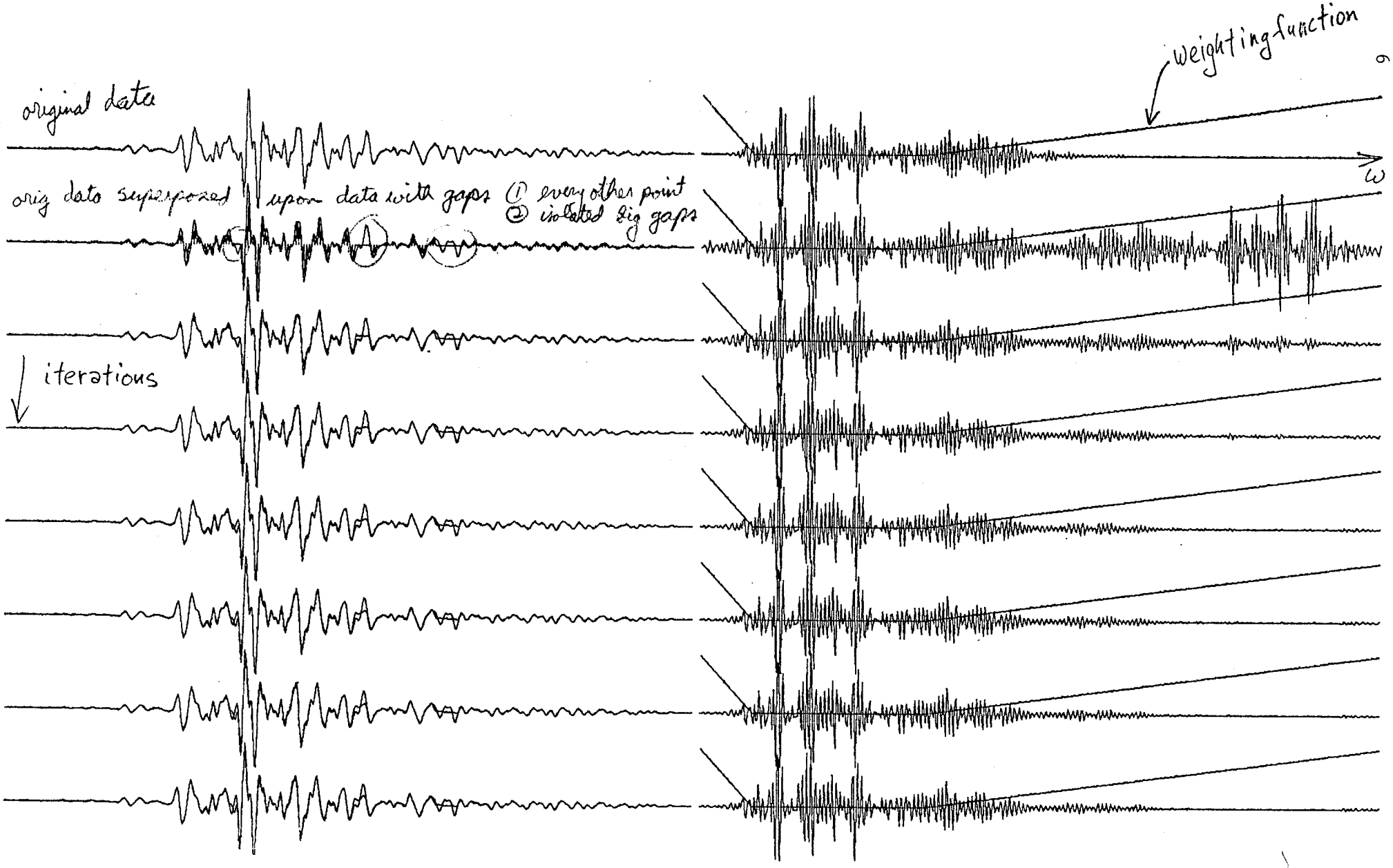

FIG. 1. Iterations to provide minimum weighted spectral power. On the left is the time function. On the right is the Fourier transform. (Real and imaginary parts are plotted as alternate points.)

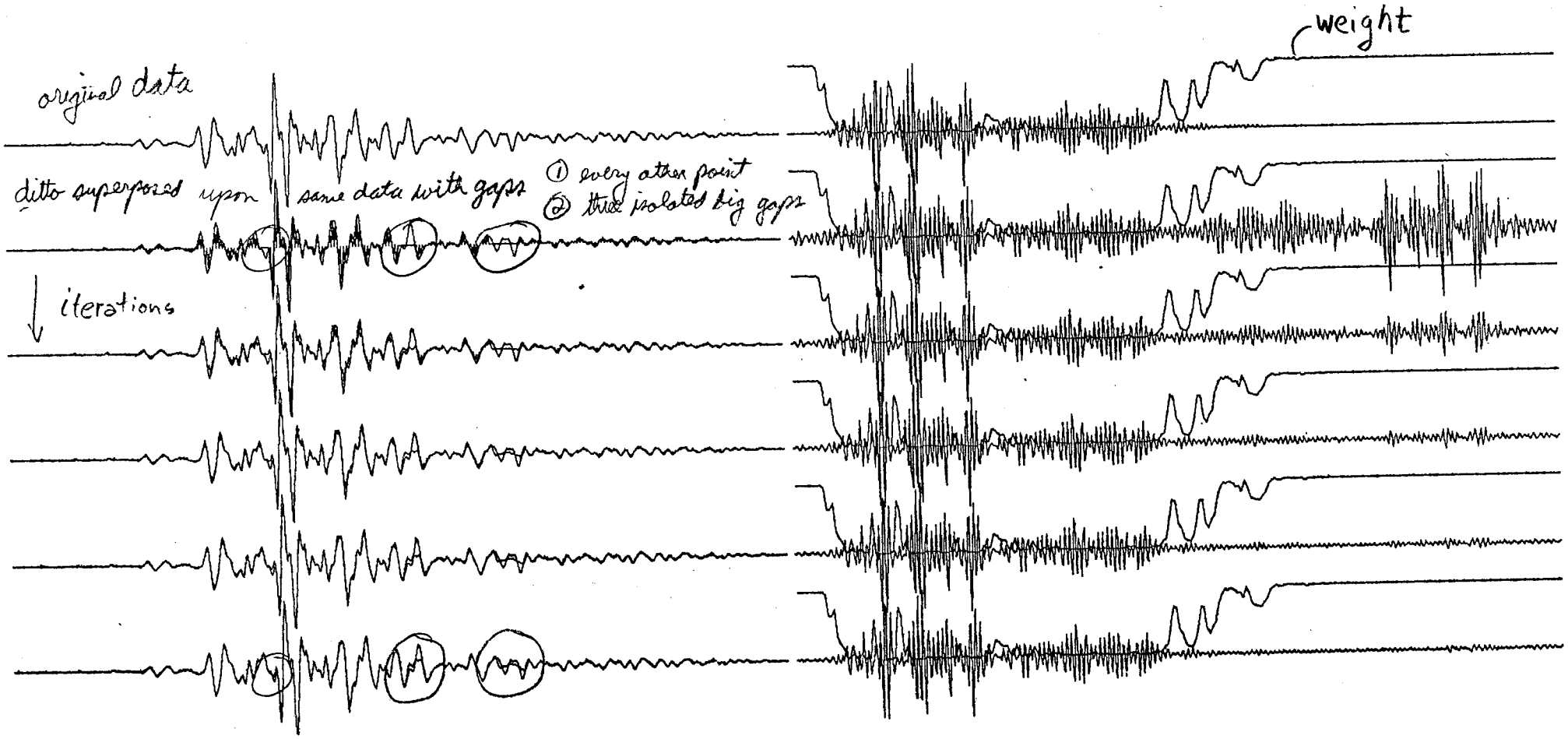

FIG. 2. This figure is similar to figure 1, but the weighting function, given by equation (6) is inverse to the expected power spectrum.

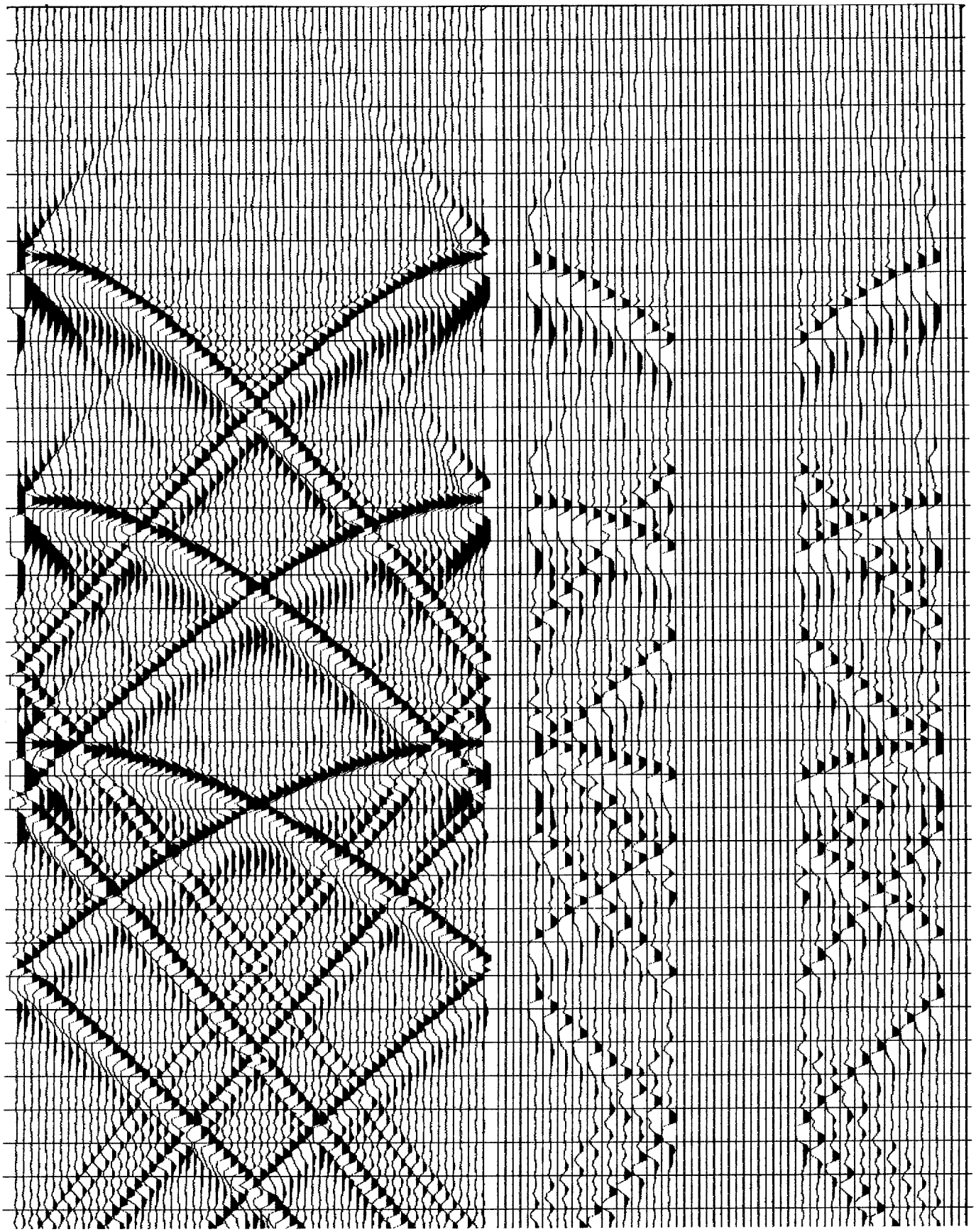

FIG. 3. Synthetic data (left) is a common midpoint gather over 3 layers. Displayed on the right, alternate traces have been removed as have traces at near and far offsets.

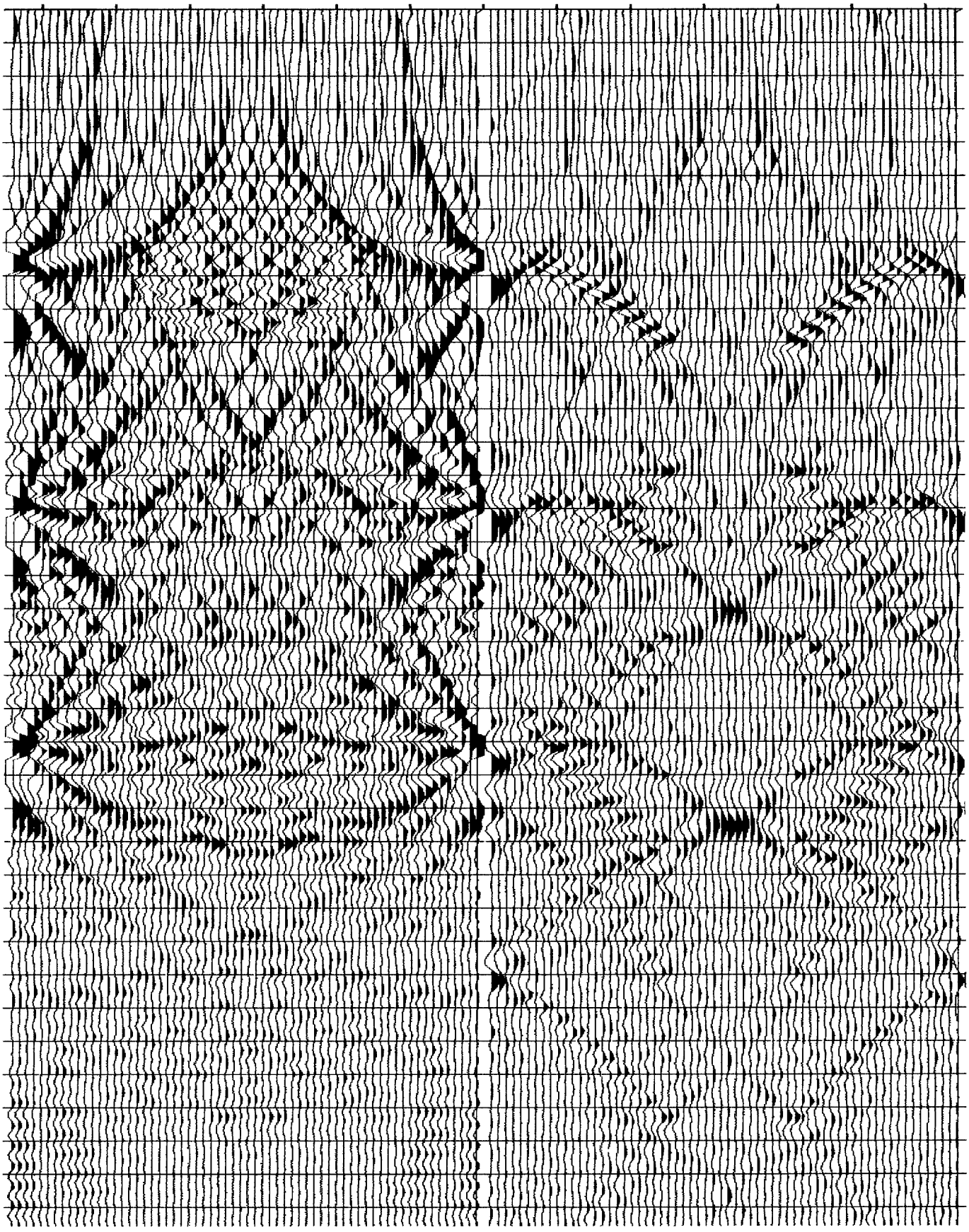

FIG. 4. Left is the downward continuation of the truncated data of figure 3. This downward continuation was zeroed about the inner traces, that is to say, a *Badpass* operation. Then it was upward continued and displayed on the right. This may then be used as the missing traces in figure 3 (right).

FIG. 5. Another iteration of figure 4.

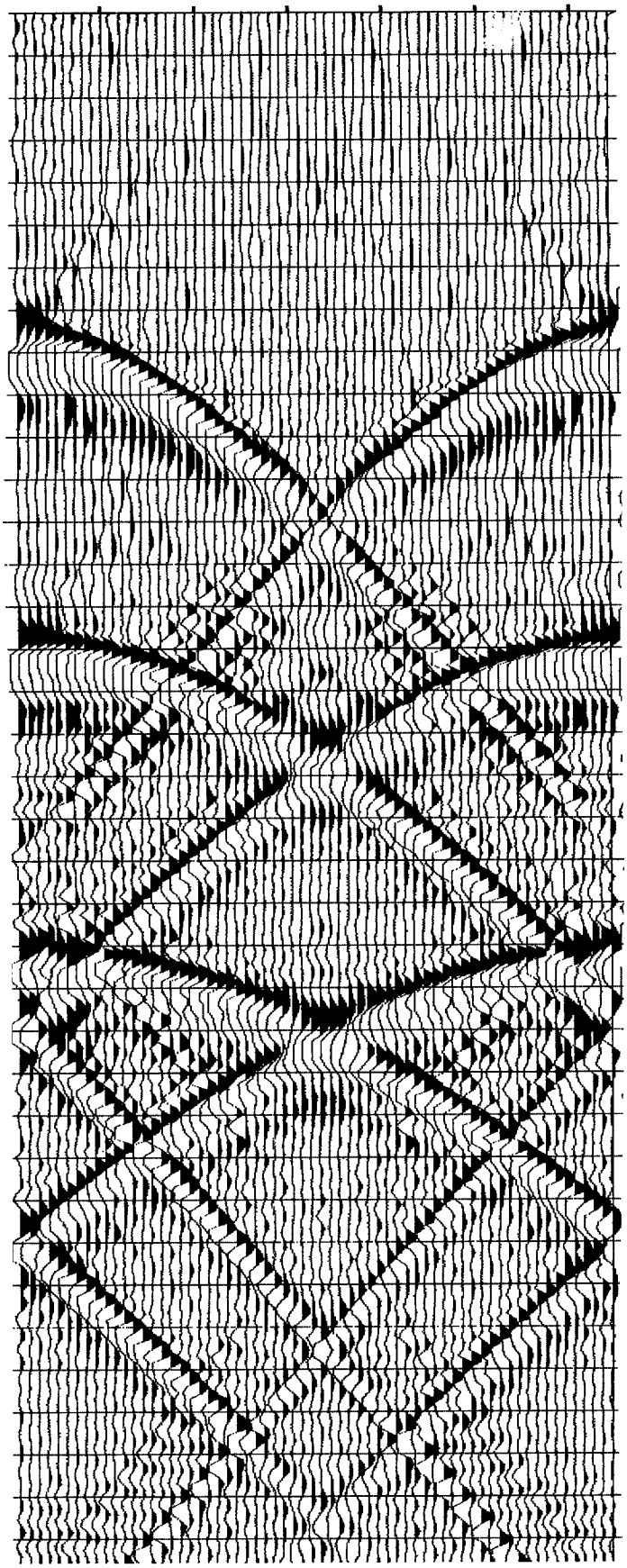

FIG. 6. Final interpolated CDP gather to be compared to figure 3.

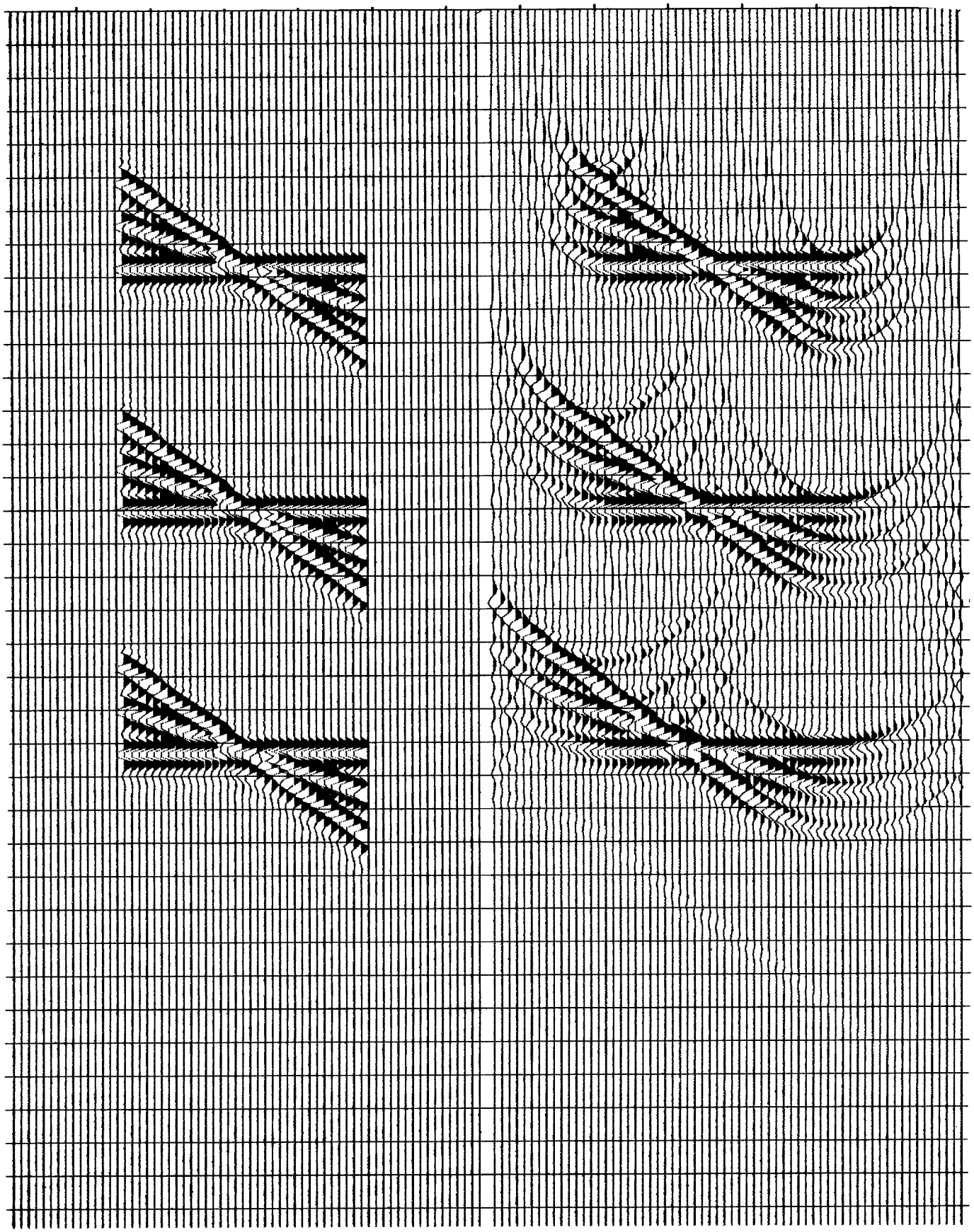

FIG. 7. Left shows a short synthetic zero offset section padded with zero traces on either side. Right shows a migration of this data. The zero traces will be modified according to the idea that there should be minimum power in the migrated data off the ends of the section.

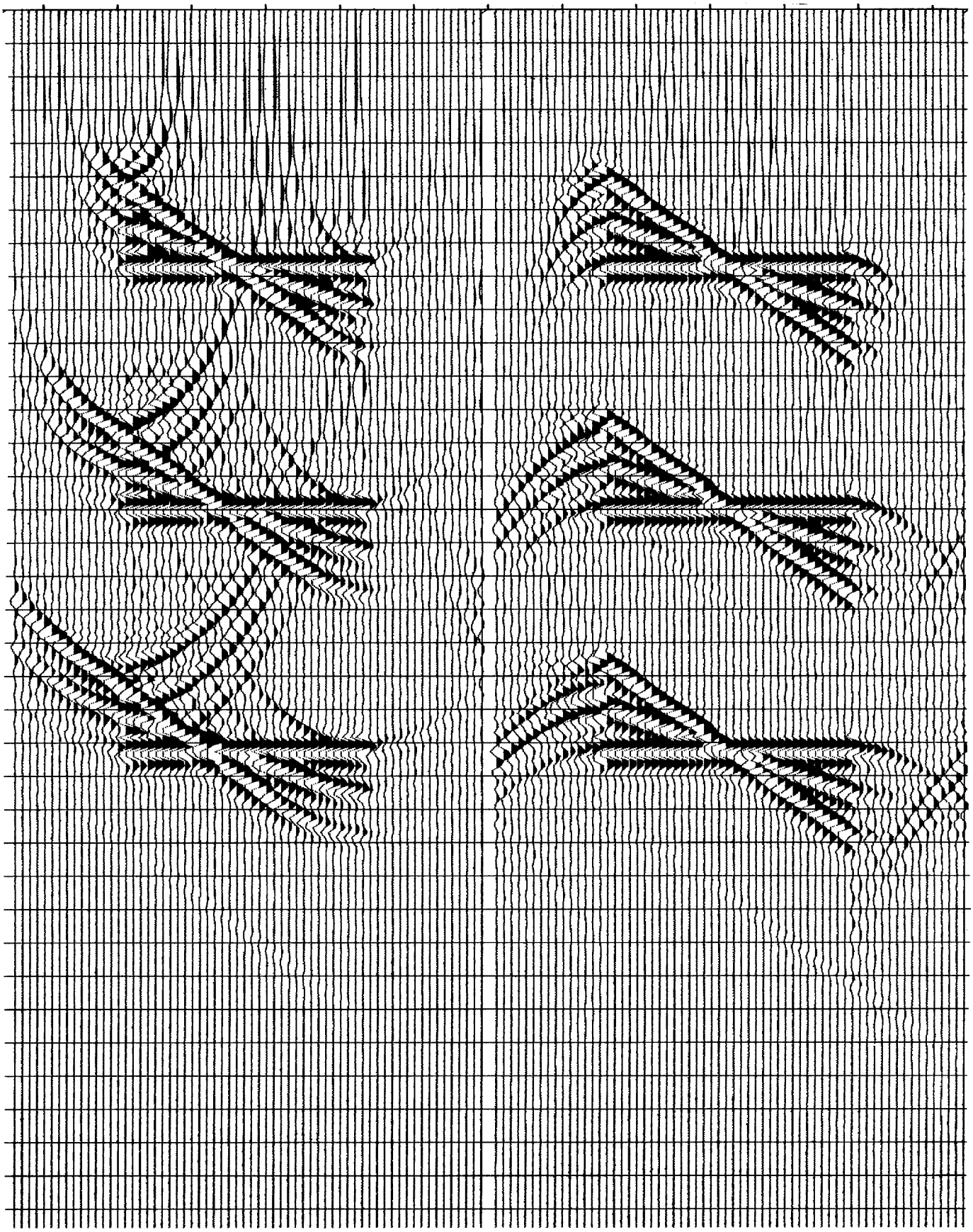

FIG. 8. Adjusting unknown data off the ends of the section so as to minimize power in the migrated section beyond the ends of the recording region we get these results. At the left is the migrated data. On the right is the surface data along with its lateral extension. Notice that the exterior semicircles have decreased as desired. But it is a disappointment to discover that the interior semicircles are now much stronger.

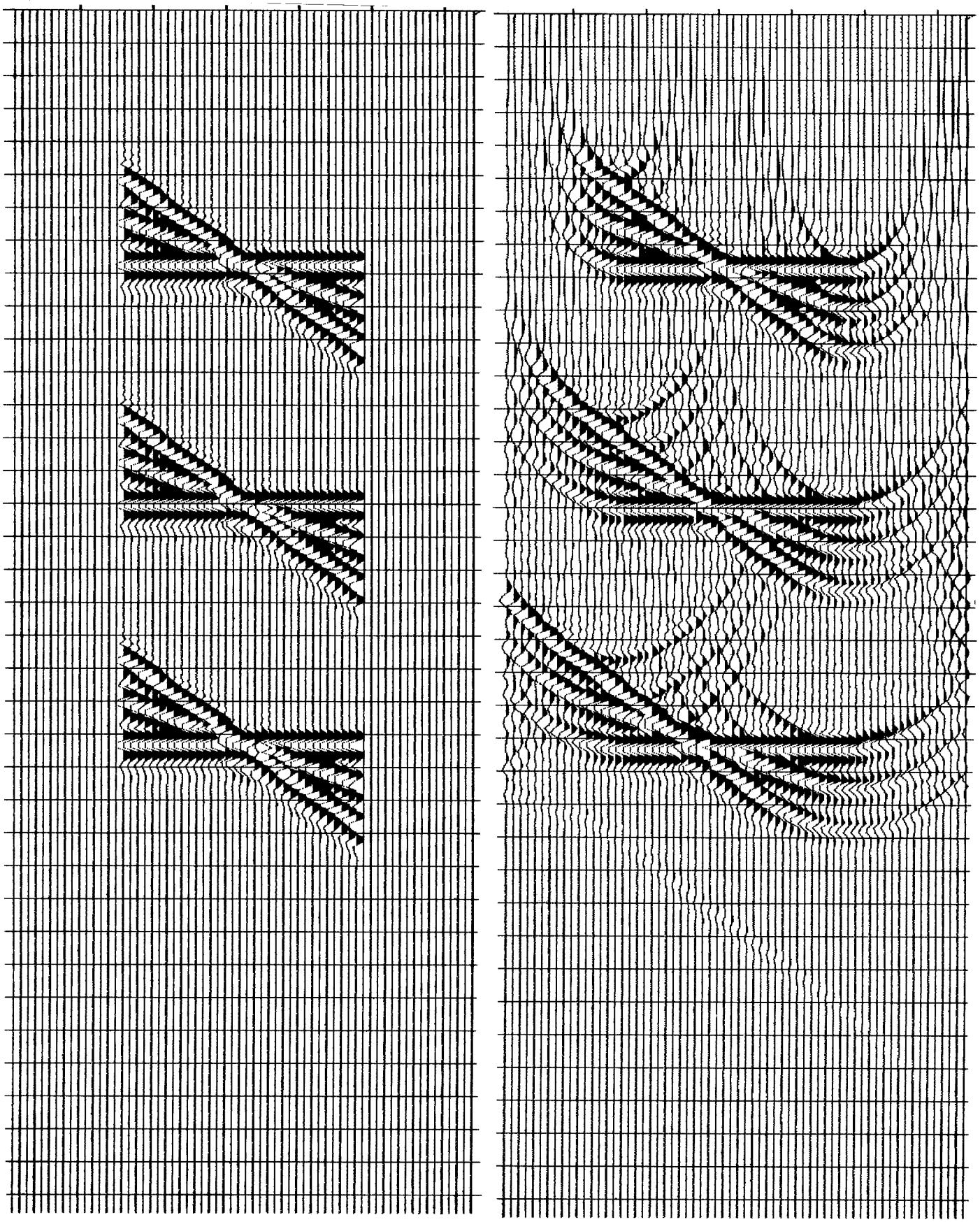

FIG. 9. A model similar to figure 7.

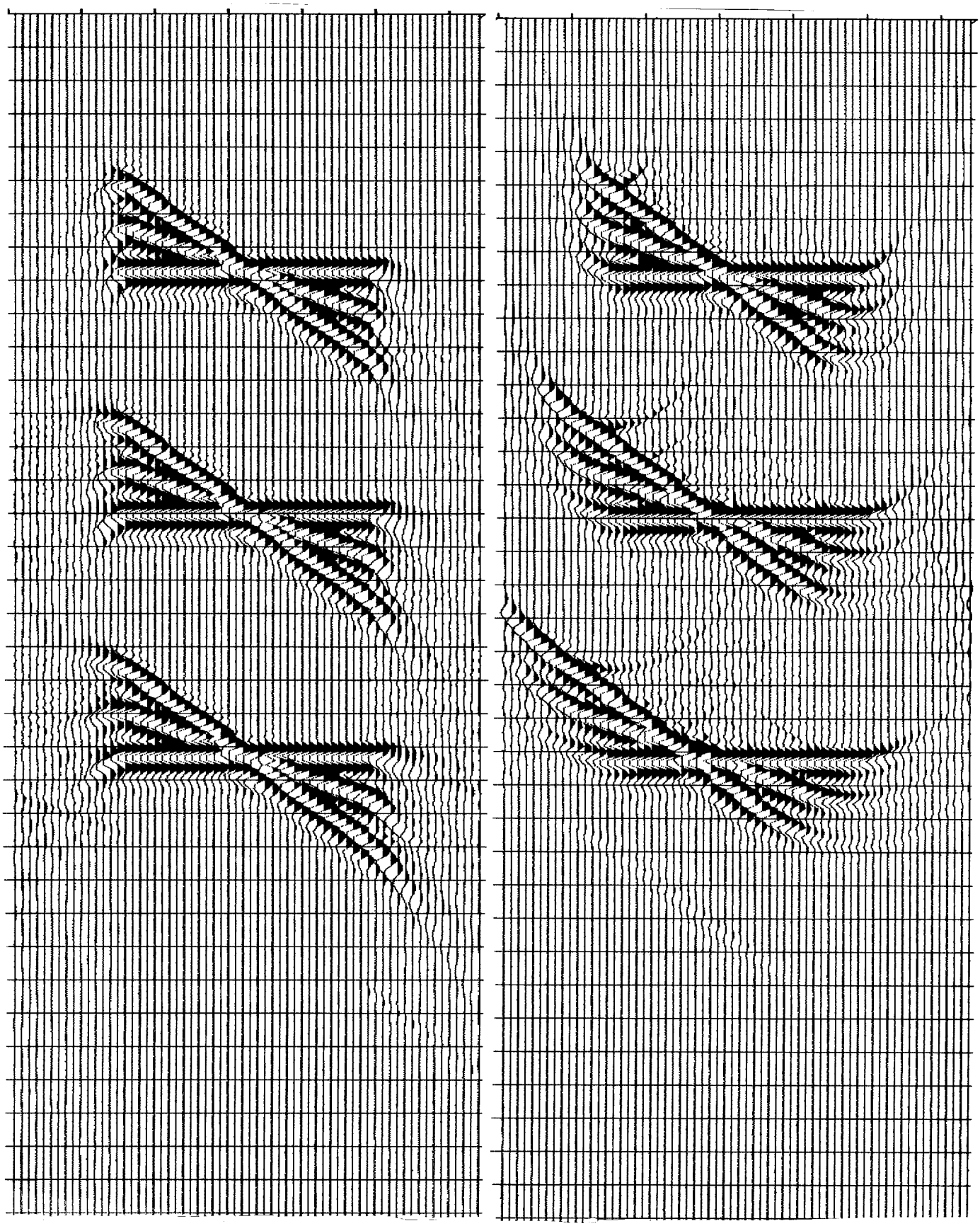

FIG. 10. Like figure 8 but the weighting function was taken to be data dependant, namely the inverse to the envelope of the data itself. Notice that the semicircles on this figure are much less than on figure 9.