Geoexploration, 20 (1982) 93—111
Elsevier Scientific Publishing Company, Amsterdam — Printed in The Netherlands

THE APPLICATION OF PRINCIPAL COMPONENTS ANALYSIS TO SEISMIC DATA SETS

DAVID C. HAGEN
ARCO Exploration Company, Dallas, Texas (U.S.A.)
(Accepted for publication June 3, 1982)

ABSTRACT

Hagen, D.C., 1982. The application of principal components analysis to seismic data sets. Geoexploration, 20: 93—111.

In the area of stratigraphic seismology, the oil company explorationist frequently encounters the problem of evaluating subtle character changes that occur within a set of essentially uniform seismic traces. Typically, the zone of interest is limited in the vertical (depth) direction to a small window relative to the overall trace length, and the seismic events are flat, or nearly so, across the set of traces. The application of principal components analysis takes advantage of the high degree of redundancy in the seismic data set to determine its statistical behavior and reduce it to its essential features. Investigations thus far indicate the information can be reduced to $\sim$10% of the original data base size. The principal component correlation coefficients were found to provide an accurate method of grouping the traces in both the supervised and unsupervised modes. If one or more well logs are available, then their geographical locations relative to the seismic data can be used to initialize the cluster centers, to which other traces are added as appropriate.

INTRODUCTION

Many exploration projects in the oil industry involve the integration of information from well logs and seismic data if these are available in an area of interest (Sheriff, 1977). The concept is to establish the lithology of the well log, and in turn relate its character to that of the seismic data. Using the much more extensive seismic coverage generally available, the lithology can be extrapolated and structural maps produced showing areas of potential hydrocarbon accumulation. This works well for finding structural traps in which permeable rocks are overlain by impermeable ones. An important element in the technique is the determination of the connection between well log and seismic data. If the well logs measure acoustic velocity information, it is possible to construct synthetic seismograms using a generating wavelet that matches the seismic data passband. Once this correlation is established a structural interpretation of the seismic data can be made.

Techniques used for seismic data acquisition and processing have continually improved over the years, and the resulting information available to

0016-7142/82/0000—0000/$02.75

the interpreter has increased in reliability and resolution. This has created a growing tendency to use seismic data for stratigraphic interpretation in which subtle changes in waveform along a reflecting horizon are related to lateral variations in composition and/or porosity within a rock layer. If well logs are available, the stratigraphic approach is similar to that used in structural interpretation, but on a different scale. Log values are adjusted in magnitude and separation until a model is found that produces a synthetic seismogram matching the seismic data after the latter has been processed in an optimal form. Additionally, or alternatively, the siesmic data can be integrated to produce a synthetic log, which is then compared to impedance data computed from the actual log.

Many stratigraphic interpretation problems involve small vertical windows of seismic data. Typically, such a data base has a high degree of redundancy in the horizontal direction, along with more subtle variations which the interpreter is actually interested in. The modeling approach described previously does not take advantage of this observation, but ultimately relies on human judgment to determine when good correlations are achieved.

For the past several years Professor Lester R. LeBlanc and coworkers at the Ocean Engineering Department, University of Rhode Island, have been working on techniques to determine the composition of sea floor sediments from acoustic reflection profiles (Milligan et al., 1978; LeBlanc and Middleton, 1980). Taking advantage of the redundant nature of the profiles, the method of principal components analysis was used to reduce the data to its essential statistical features. Each trace could then be represented by a small set of coordinates which was used to classify it into one of several possible types. This classification scheme resulted in an excellent agreement with actual sediment categories.

The principal components technique was developed in the 20's and 30's by several investigators, with perhaps Hotelling (1933) being due the most credit for the procedure. An outline of the statistical theory behind it is succintly presented by Milligan et al. (1978), which with further condensation is summarized as follows. Let a set of trace segments be represented in vector form as (χ_{ik}), where $k = 1, \ldots, N$ is the trace index and $i = 1, \ldots, M$ is the sample index. Then the mean vector and covariance matrix estimates are:

$$\bar{\chi}_i = \frac{1}{N} \sum_{k=1}^{N} \chi_{ik} \tag{1}$$

$$S_{ij} = \frac{1}{N-1} \sum_{k=1}^{N} (\chi_{ik} - \bar{\chi}_i)(\chi_{jk} - \bar{\chi}_j) \tag{2}$$

The assumption is made that the mean and covariance contain all the statistical information about the data base; i.e., it is second order. The covariance matrix is inverted to produce a set of eigenvectors and eigenvalues. The inversion process produces M eigenvectors which, as a set of vectors has the

property of being orthogonal. The contribution of each eigenvector to the overall variance is proportional to the corresponding eigenvalue. The inversion process produces eigenvectors in the direction of decreasing eigenvalues, so if the first $\widetilde{M}<M$ eigenvalues provide a sufficient amount of variance measure, then only $\widetilde{M}$ of them need be used in further analysis. The $\widetilde{M}$ eigenvectors, which contain N samples each, can then be correlated against each trace to obtain a coefficient expressing that eigenvector's contribution to the trace. Each trace then can be estimated as a linear combination of the $\widetilde{M}$ eigenvectors, or principal components.

$$\chi_{ik} \approx \widetilde{\chi}_{ik} = \sum_{m=1}^{\widetilde{M}} \alpha_{mk} Z_{mi} \tag{3}$$

where $\chi_{ik}, \widetilde{\chi}_{ik}$ = actual, estimate, ith sample of kth trace, Z_{mi} = ith sample of mth eigenvector, α_{mk} = correlation coefficient of eigenvector Z_m and trace χ_k.

The percent of the total variance of the data base which is accounted for by the first $\widetilde{M}$ principal components is given by

$$\sigma^2(\widetilde{M}) = \frac{\sum_{m=1}^{\widetilde{M}} \lambda_m}{\sum_{m=1}^{M} S_{mm}} \tag{4}$$

where λ_m = mth eigenvalue, and S_{mm} = diagonal components of covariance matrix.

To appreciate the significance of eq. 4, it might be helpful to consider that for a set of 200 good quality traces, using a 50-sample window, 3 to 4 principal components will generally express 85 90% of the overall variance. Thus the original data base of 10,000 samples can be reduced to $4 \times 50 + 4 \times 200 = 1000$ samples, indicating a typical 10-fold information redundancy in the original seismic data.

Since each trace in the data set can be estimated as a linear combination of the same principal components, the unique character of a trace is expressed by its set of correlation coefficient values (α_{mk}), $m = 1, \ldots, \widetilde{M}$. Milligan et al. (1978) and LeBlanc and Middleton (1980) used these coefficients weighted by the inverse square root of the corresponding eigenvalue as coordinates in $\widetilde{M}$-dimensional space, and allowed these to agglomerate iteratively to a pre-determined number of clusters. This non-supervised clustering procedure is initiated by defining one cluster for each of the N traces, merging the two nearest as defined by the distance measure

$$d_{AB} = \sqrt{\sum_{m=1}^{\widetilde{M}} \frac{(\alpha_{mA} - \alpha_{mB})^2}{\lambda_m}} \tag{5}$$

and recomputing the coordinate position of the merged clusters. The procedure is repeated with a reduction of one cluster at each step until only the desired number remains.

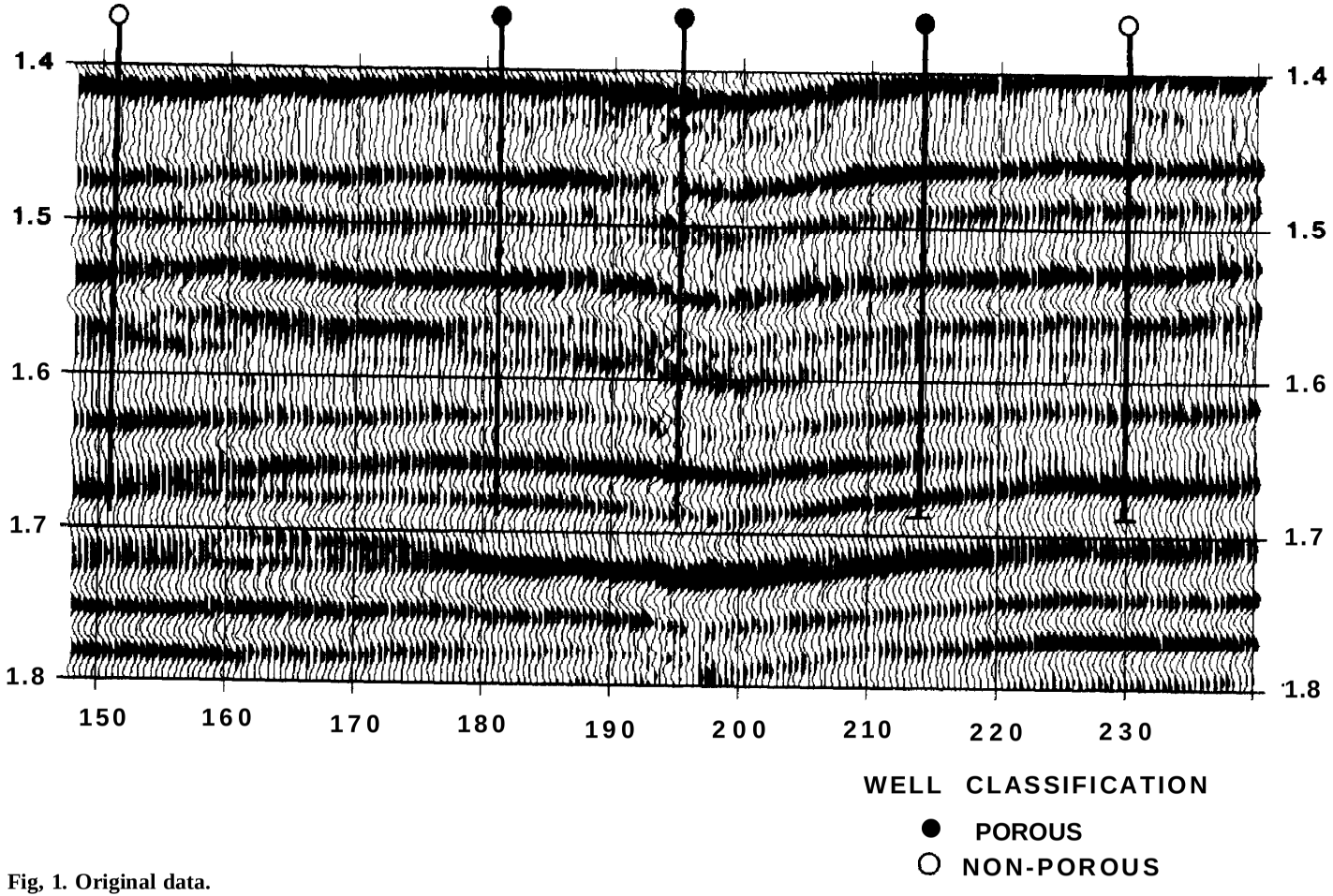
RESULTS

The stratigraphic interpretation problem is on a different scale from that of sea floor sediment identification but can be handled in a similar manner. The clustering done for the subject work primarily used the supervised mode with the correlation coefficients of the traces at known well locations serving as cluster centers.

The procedure is illustrated using the profile of seismic data shown in Fig. 1, which is a window from a common depth point stack. The exploration target was a porous zone which appeared and disappeared laterally in the zone 1.66—1.68 s. It was determined the indicator differentiating the non-porous and porous classes was the split-up of the single low-frequency cycle into two high-frequency cycles. Wells penetrated this zone at shotpoint locations 152, 181, 194, 214 and 230. The coefficients for the traces at 214 and 230 were used in the supervised classification process to represent the porous and non-porous classes, respectively. The wells at 151, 181 and 195 were not used in the classification process, and indicated non-porous, porous and porous character, respectively. The solid and hollow circles are used to symbolize the porous and non-porous wells. It should be noted that data recording difficulties were encountered around the well at 195. This led to an anamoly in the data and subsequently some difficulty in classifying the well. In Fig. 2, the data has been flattened on an event at 1.41 ms and filtered using a 6—58 Hz passband to eliminate variations unrelated to stratigraphy. The statistical analysis was done on the data in this format.

The porous/non-porous contrast is emphasized in Figs. 3 and 4, which are shade-coded renditions of the data illustrating the amplitude envelope and instantaneous frequency transformations of the traces (Taner and Sheriff, 1977). In the interest of condensing the computer time, the zone of interest was narrowed to 150 ms and every other trace was used. The behavior of the amplitude envelope did not correlate with the porous/non-porous distinction, while the frequency emphasized it very well, as might be expected from an observation of the stacked data. In general, a frequency high indicates porosity and a low is related to a lack of porosity. The high and low frequency anomalies can be detected by differences in shading patterns.

The discriminating behavior of the instantaneous frequency led to its use as input to the statistical analysis procedure. Because of uncertainty as to the number of principal components required, several examples were run and the data estimates were generated using eq. 3. The comparisons for 2, 3 and 5 principal components with the original data are shown in Fig. 5. While the 5-component case shows the best resemblance, the 3-component case was deemed a sufficiently close match.



Fig, 1. Original data.

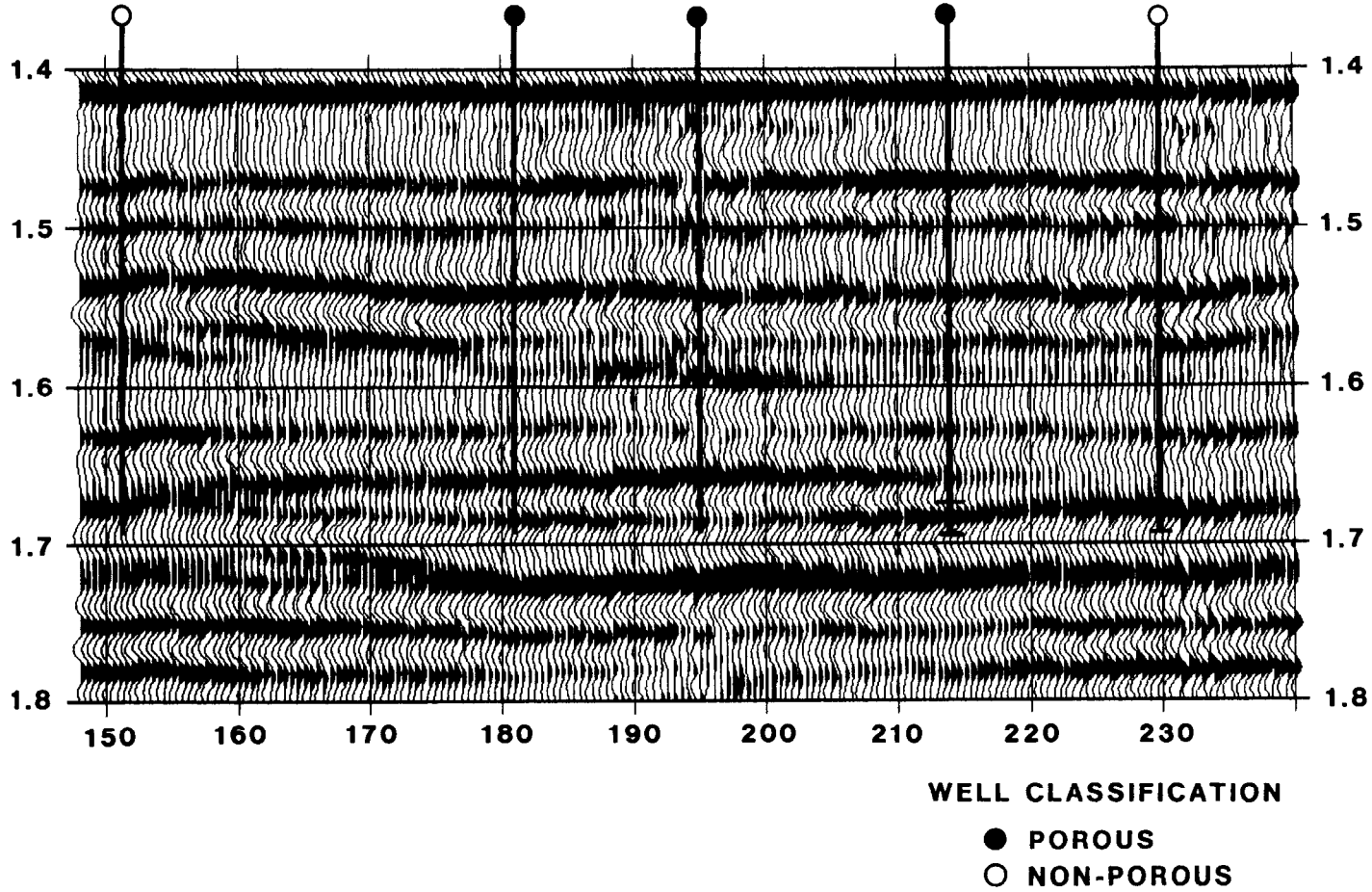


Fig. 2. Traces flattened at 1.41 ms, filtered to 6–58 Hz passband.

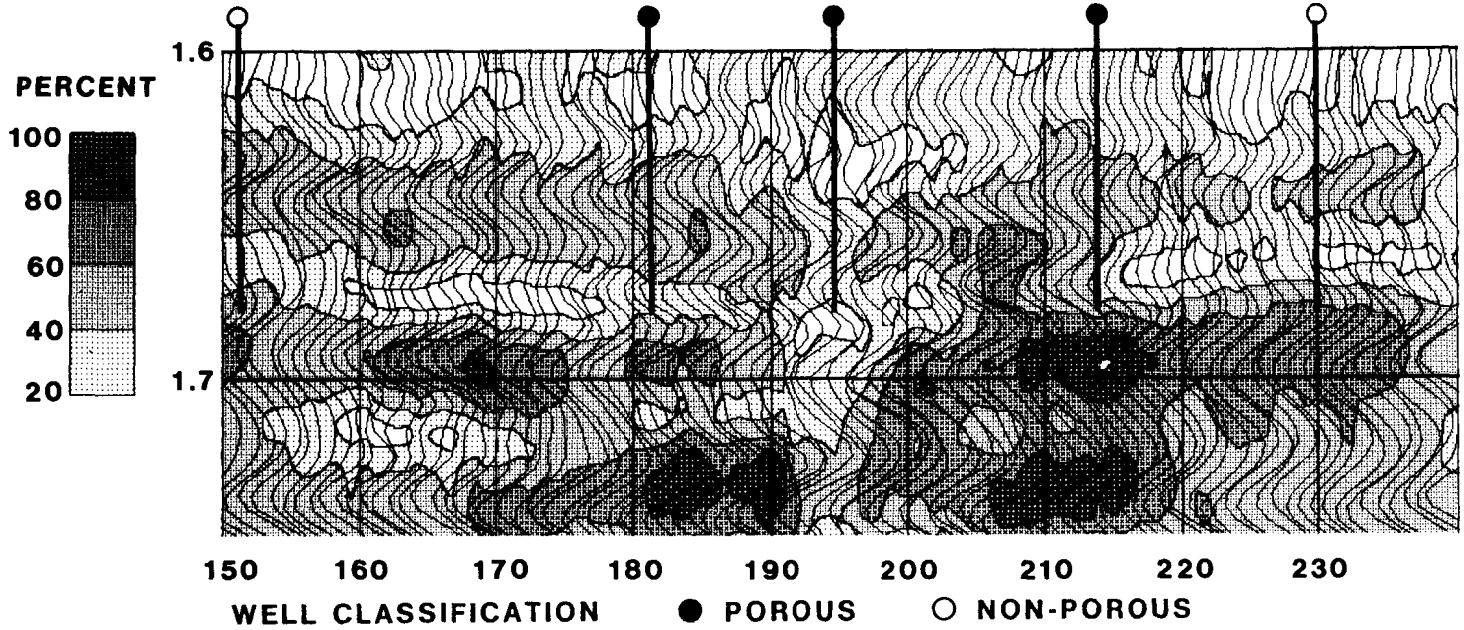


Fig. 3. Traces with amplitude envelope.

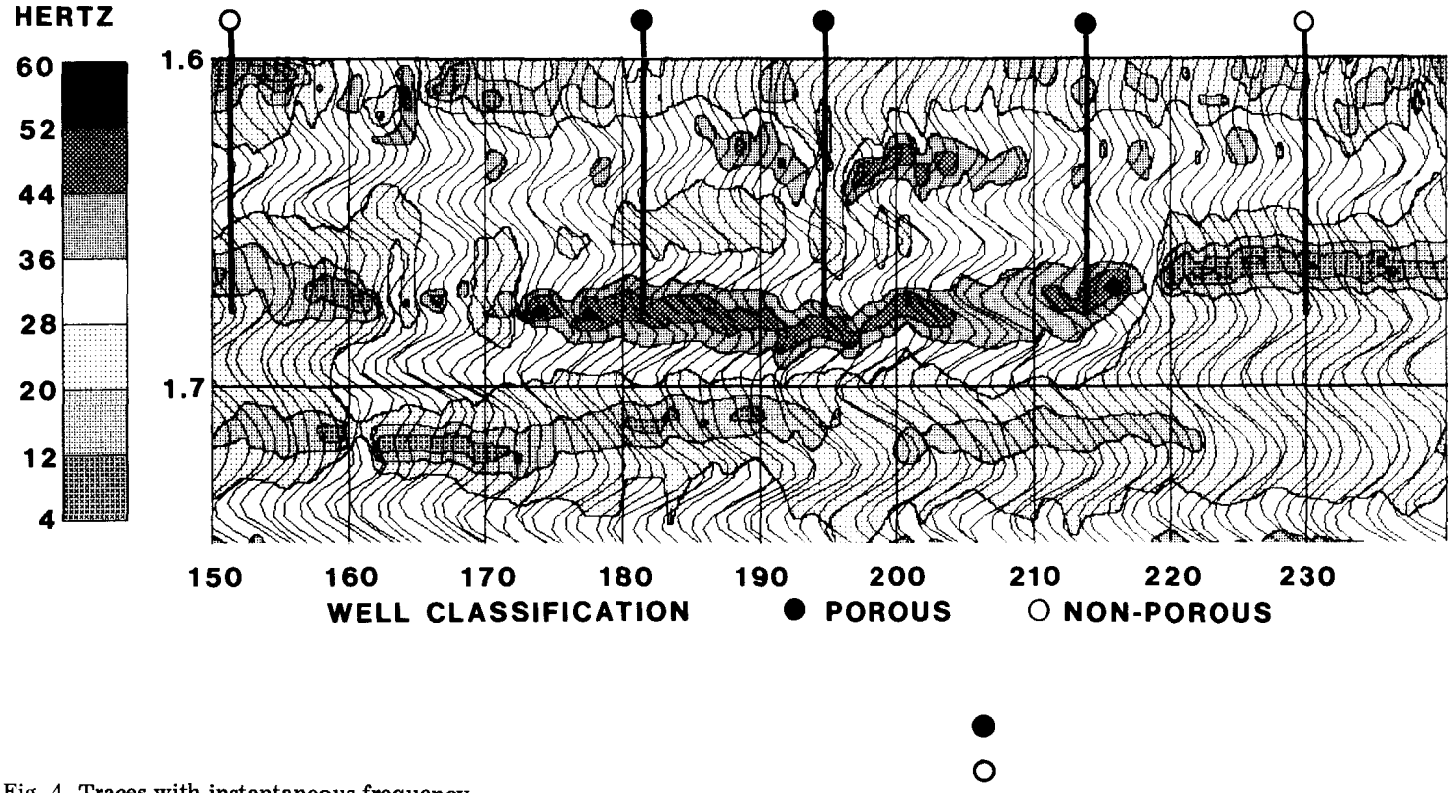


Fig. 4. Traces with instantaneous frequency.

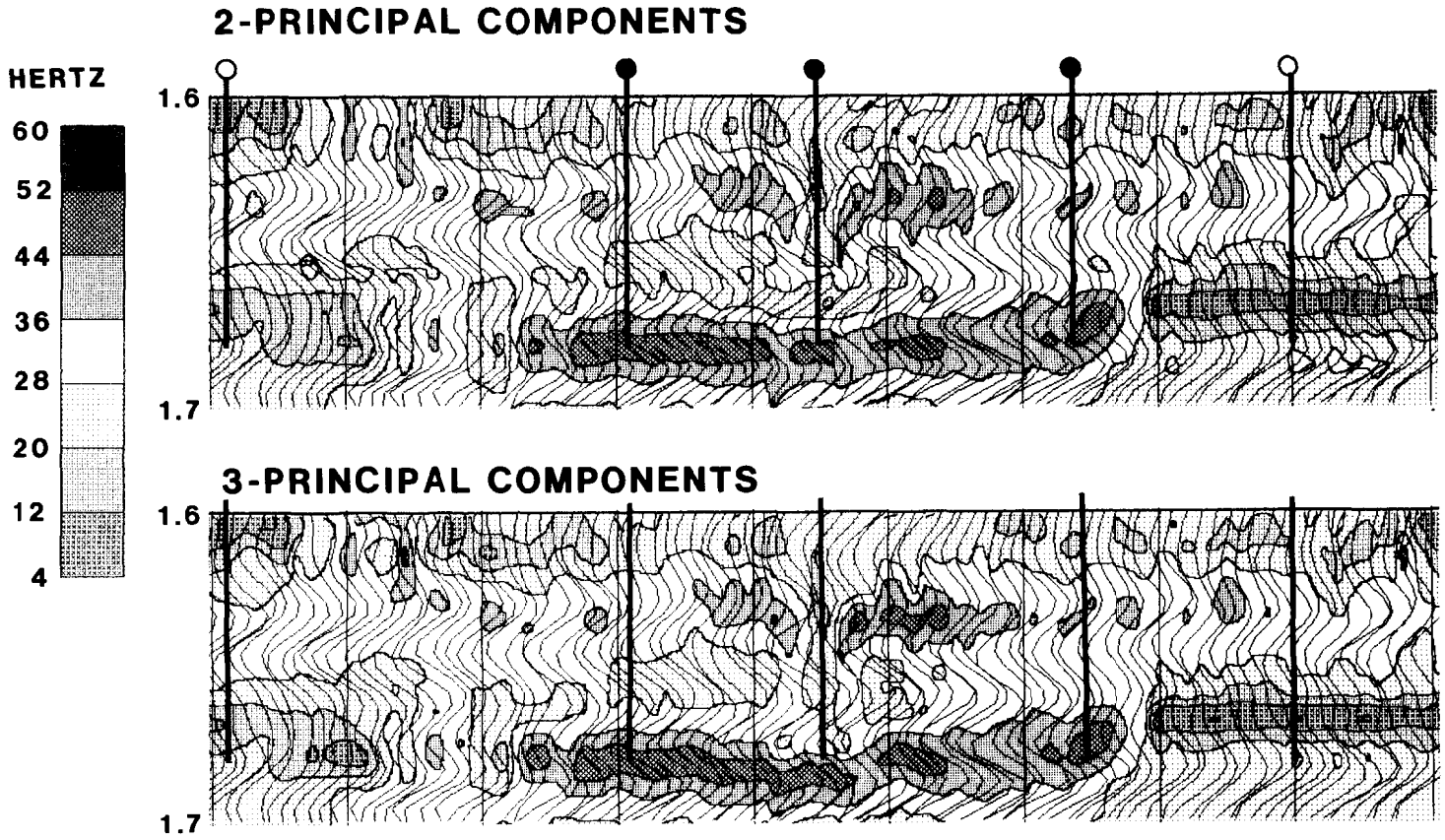


Fig. 5. (continued on p. 102)

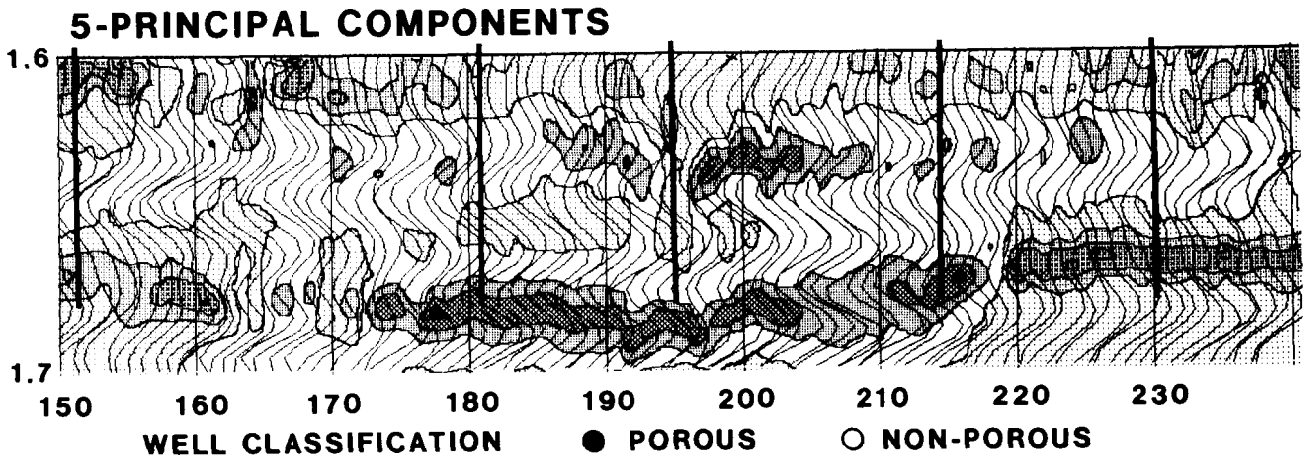


Fig. 5. Traces with instantaneous frequency reconstructed from 2, 3, and 5 principal components, resp.

The correlation coefficients for the first three components are shown in Fig. 6. The behavior reversal for components one and two between shotpoints 214 and 230 is particularly noticeable. Another observation of interest is that coefficient behavior exhibits a low frequency characteristic, which suggests coefficient values for missing traces could be determined by interpolation and used to reconstruct those traces.

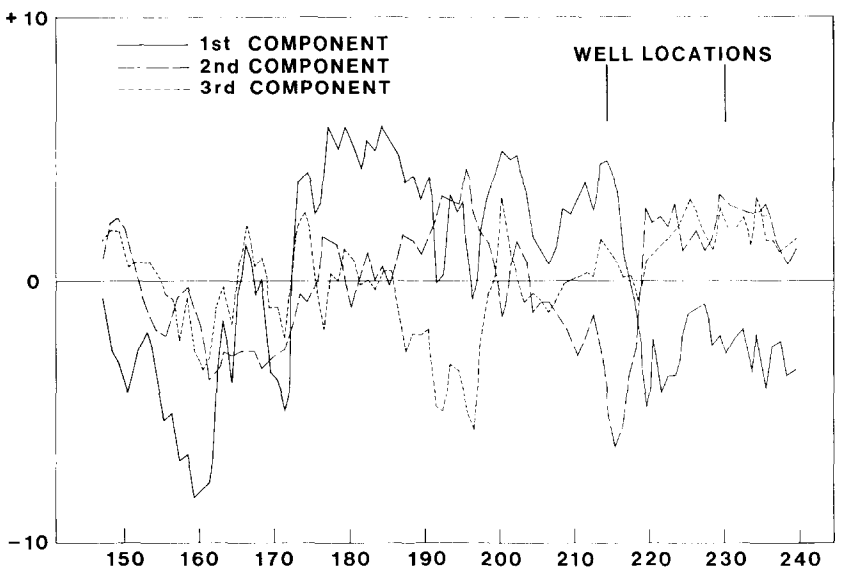
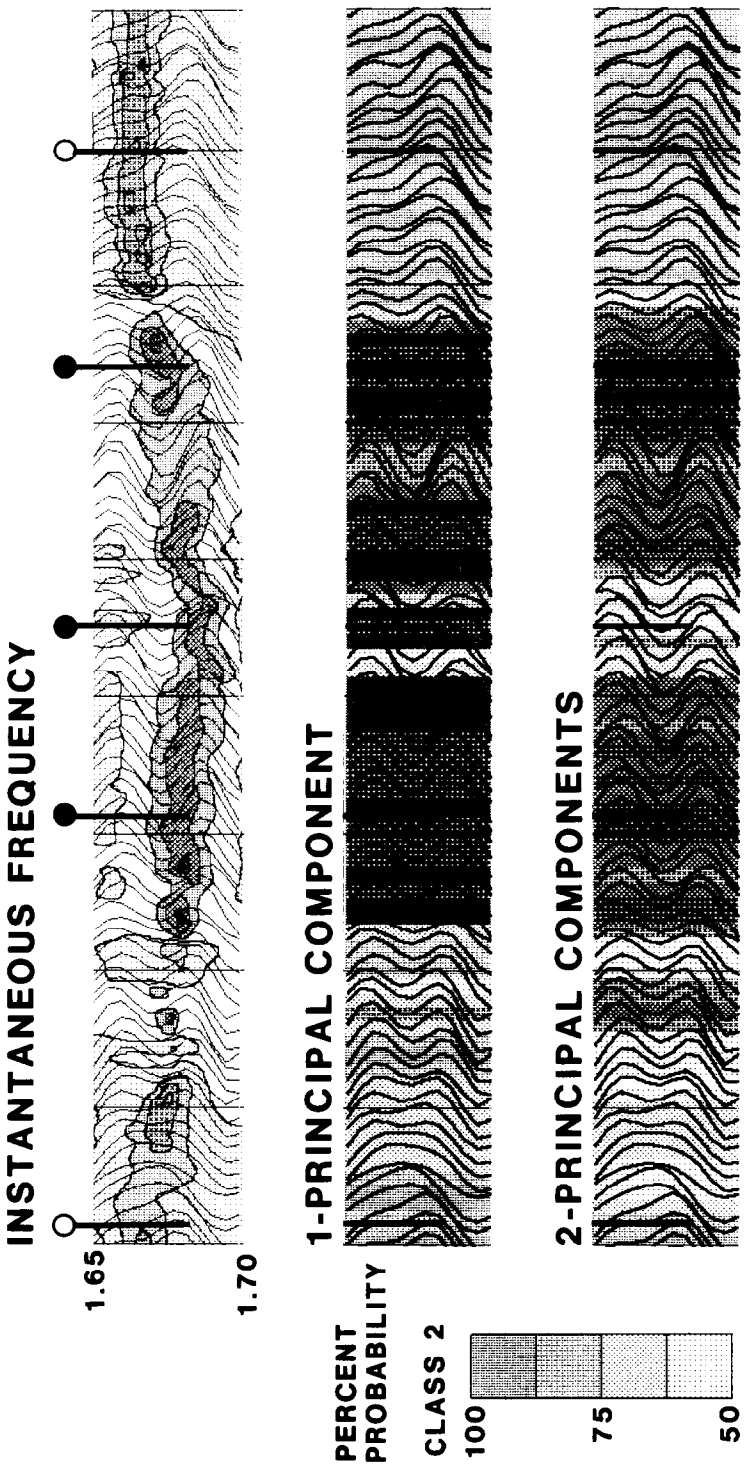


Fig. 6. Correlation coefficients for principal components.

The correlation coefficients were used to classify the data using the data at 214 and 230 as cluster centers. The distance of each trace to each center was computed using eq. 5. It was assigned to the closest class, provided it fell within a maximum acceptance radius. If it could be assigned to a class, the probability it was in the correct class was computed as the ratio of its reciprocal distance to that class and the sum of the reciprocal distance to all classes:

$$p(X \epsilon A) = \frac{\frac{1}{d_{XA}}}{\frac{1}{d_{XA}} + \frac{1}{d_{XB}} + \ldots} \tag{6}$$

Fig. 7 shows the classification results for 1, 2, 3 and 5 principal components. A dark background indicates that a trace was determined to be classified as porous, and a light background as non-porous. The heavier the shading within a particular class, the higher is the probability that it


INSTANTANEOUS FREQUENCY
1.65
1.70
1-PRINCIPAL COMPONENT
2-PRINCIPAL COMPONENTS
PERCENT PROBABILITY
CLASS 2
100
75
50

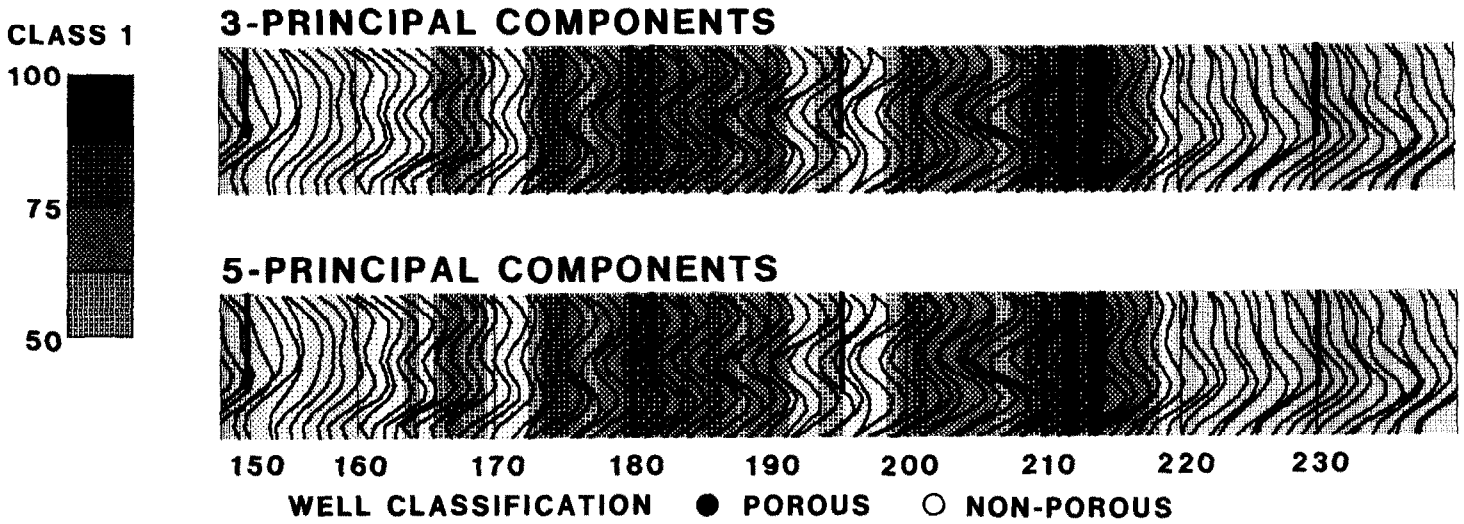
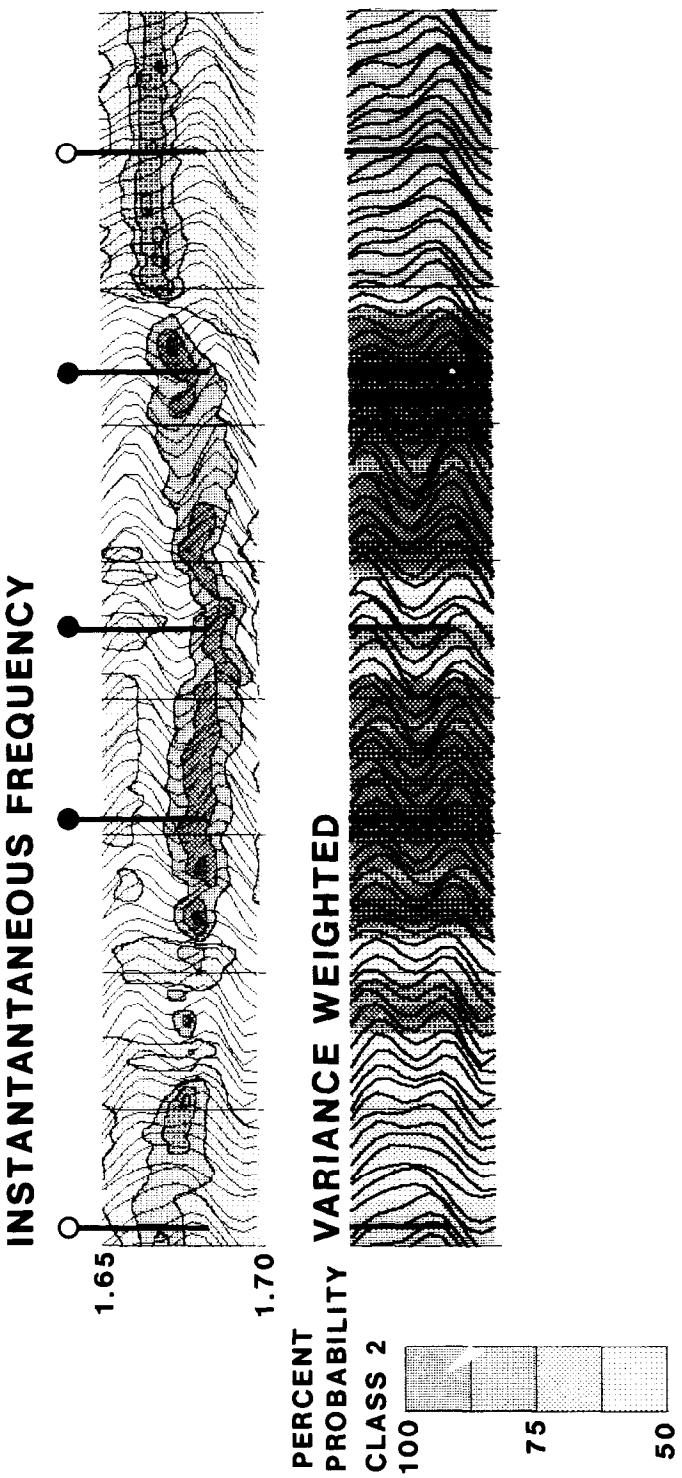


Fig. 7. Classification by instantaneous frequency.


INSTANTANEOUS FREQUENCY
1.65
1.70
VARIANCE WEIGHTED
PERCENT PROBABILITY
CLASS 2
100
75
50

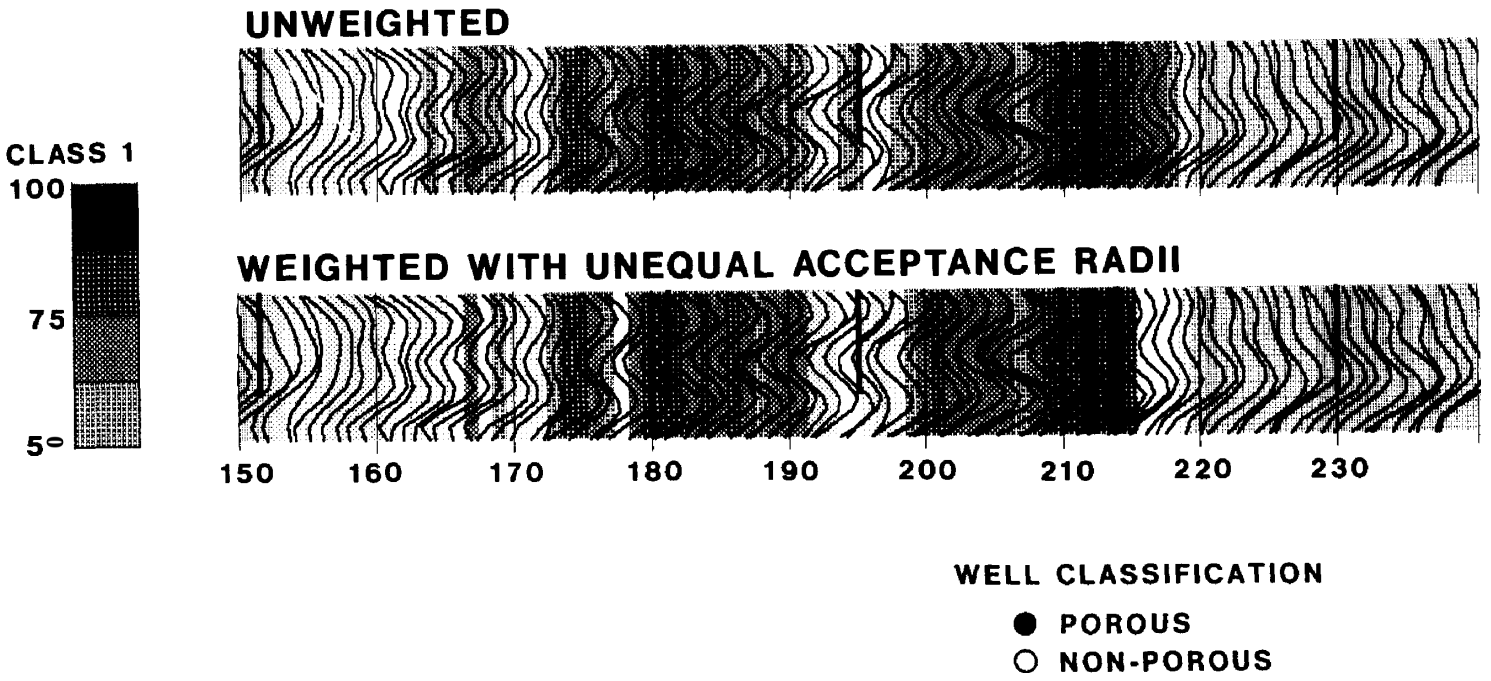


Fig 8. Classification with 3 principal components.

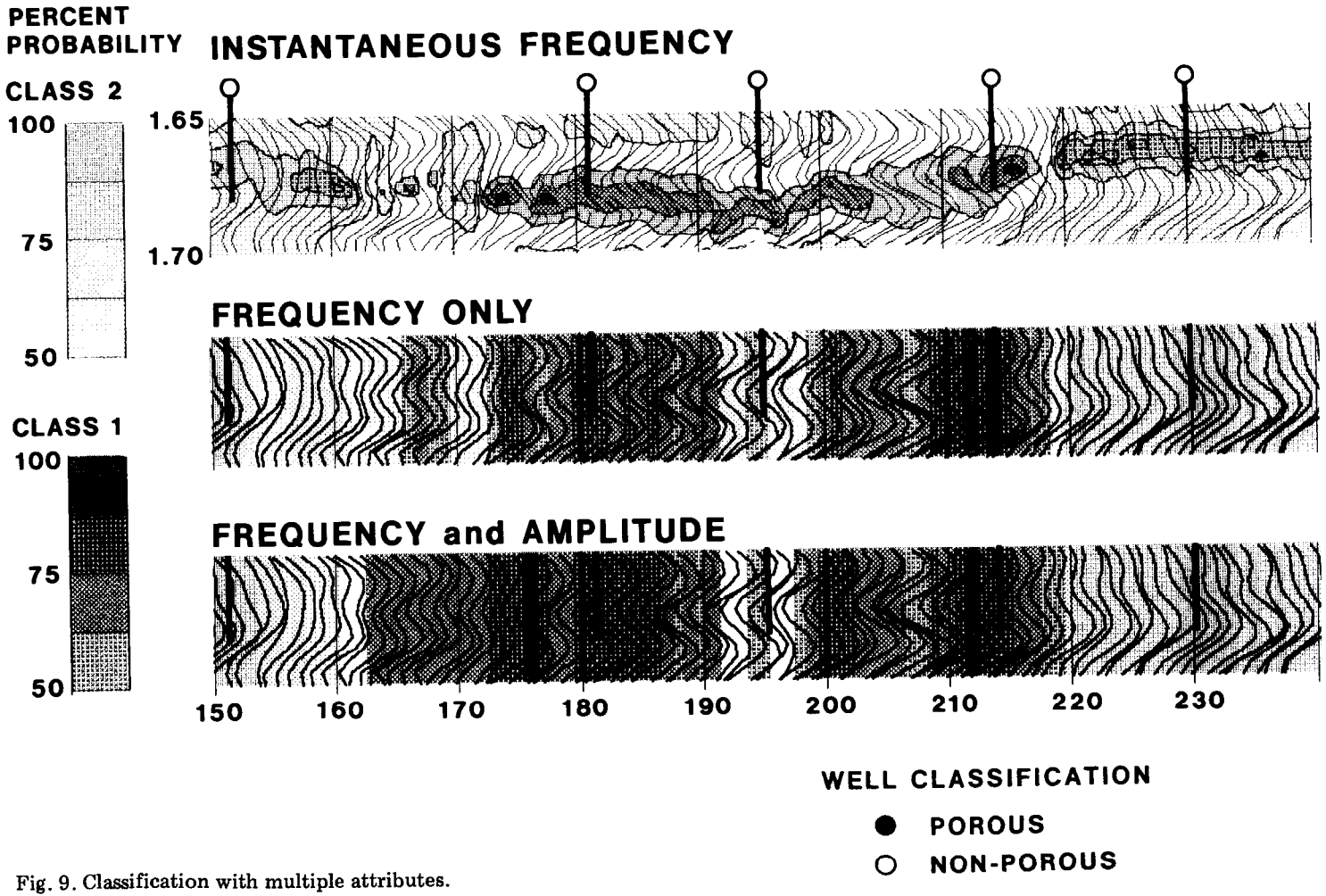


Fig. 9. Classification with multiple attributes.

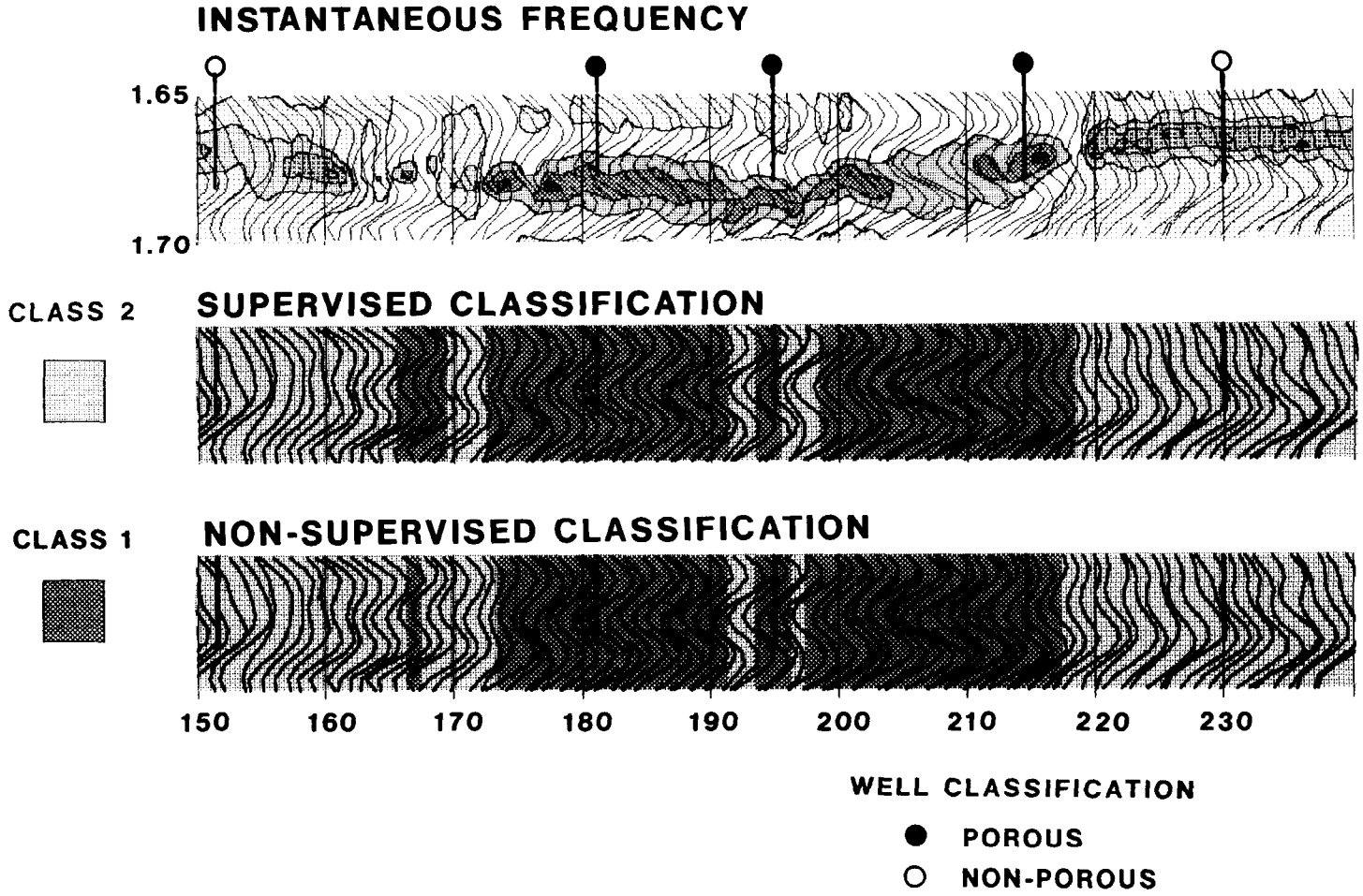


Fig. 10. Comparison with non-supervised classification.

has a stratigraphy similar to what exists at the well. As expected, the data at 214 and 230 show a high probability of belonging to the respective classes. The procedure has correctly classified the wells at 151 and 181 with a high degree of probability, but the classification of the well at 195 is less certain, probably due to the recording anomaly in the area. While all four are in general agreement, the classification becomes more discriminating for larger numbers of components.

Another comparison of classification variables is shown in Fig. 8, all for the 3-component case. The illustration showing classification for variance-weighted coordinates is shown for reference. Below it is the classification result without weighting, showing little apparent difference. The bottom figure illustrates variance-weighting with an acceptance radius for the porous class only half as large as the non-porous class. As expected, the number of traces correlatable to the porous class has diminished.

The classification procedure can readily be extended to two or more transformations of the basic data set if that is necessary to fully distinguish between different classes. Although the amplitude envelope transform did not exhibit a suitable character, as an experiment it was used anyway in conjunction with the frequency transform, and a comparison of the results is shown in Fig. 9. This classification appears to be less distinct than the frequency alone, suggesting that inclusion of the amplitude has confused the issue. The use of multiple transforms has proved helpful in other cases, particularly in combining transforms generated over dissimilar axes, such as instantaneous frequency in the time domain and power spectra in the frequency domain, or amplitude and velocity data in the time domain.

The preceding classifications have all been done in the supervised mode. Again, as a comparison, the frequency data were classified in the unsupervised mode using the previously referenced technique (Milligan et al., 1978). Fig. 10 is a comparison of the results, with the supervised classification being reduced to a binary (yes—no) decision. For the case of two classes, the results are very similar. No attempt was made in this experiment to determine a probability of class membership.

The techniques have been applied to two-dimensional data as well, in which the matrix is strung out as one long vector. If the matrix becomes sizable, however, it is difficult to invert because of roundoff problems and the build-up of computer time. For small matrices, though, the technique has proven to work quite well.

CONCLUSIONS

Although the results presented in this report refer to a single data base, the techniques of principal components analysis and clustering as described have been found to be applicable to other seismic data. If further experiments confirm the technique as being reliable and robust, it may prove useful in reducing the judgment factor in stratigraphic interpretation problems. In summary, the results presented indicate the following.

(1) A seismic data set typically has a high content of information redundancy, a fact which can be used to advantage in determining its essential features.

(2) If the appropriate pre-normalization operations are performed, this reduced information can be used to accurately categorize the data using either a supervised classification tied to well log data, or unsupervised classification in which "natural" clusters are allowed to form.

(3) The statistical analysis and classification can be done on the seismic data, or on one or more transforms of the data, depending upon which shows the discriminating character.

It appears the usage of these techniques would indeed be helpful in reducing the "manual" workload of the oil company explorationist.

ACKNOWLEDGMENTS

I want to thank the ARCO Exploration Company for permission to publish this report, and Professor LeBlanc for his time and assistance.

REFERENCES

Hotelling, H., 1933. Analysis of a complex of statistical variables into principal components. J. Educ. Psychol., 24: 417—441; 498—520.

LeBlanc, L.R. and Middleton, F.H., 1980. An underwater acoustic sound velocity data model. J. Acoust. Soc. Am., 67: 2055—2062.

Milligan, S.D., LeBlanc, L.R. and Middleton, F.D., 1978. Statistical grouping of acoustic reflection profiles. J. Acoust. Soc. Am., 64: 795—807.

Sheriff, R.E., 1977. In: C.E. Payton (Editor), Seismic Stratigraphy — Applications to Hydrocarbon Exploration. AAPG Mem., 26: 3—14.

Taner, M.T. and Sheriff, R.E., 1977. In: C.E. Payton (Editor), Seismic Stratigraphy — Applications to Hydrocarbon Exploration. AAPG Mem., 26: 301—328.