AN APPLICATION OF PATTERN RECOGNITION PRINCIPLES TO THE ANALYSIS OF WIRELINE LOGS

By

ROBERT GENE HAYES

Bachelor of Science
Oklahoma State University
Stillwater, Oklahoma
May, 1977

Master of Science
Oklahoma State University
Stillwater, Oklahoma
May, 1978

Submitted to the Faculty of the
Graduate College of the
Oklahoma State University
in partial fulfillment of
the requirements for
the Degree of
DOCTOR OF PHILOSOPHY
December, 1989

AN APPLICATION OF PATTERN RECOGNITION PRINCIPLES TO THE ANALYSIS OF WIRELINE LOGS

Thesis Approved:

Thesis Adviser

Gary F. Stewart

Keith A. Teague

C.M. Bacon

Norman N. Durham

Dean of the Graduate College

ACKNOWLEDGEMENTS

I would like to extend my sincere thanks and appreciation to my adviser, Dr. Rao Yarlagadda, for his professional guidance and counsel during my doctoral program. I have had the privilege to work with a man of great character and immeasurable patience during my graduate study. I now understand what he meant when he said that getting a Ph.D. is a 'grueling' process. I would also like to thank the rest of my graduate committee, Dr. Charles Bacon, Dr. Keith Teague and Dr. Gary Stewart. Their longevity and stability as faculty members at Oklahoma State University has been a constant encouragement to me over what seems to have been a very long time. I would especially like to thank Dr. Stewart for his geological consultations that proved ever so valuable in my efforts to understand well logs.

This research was supported by a grant from members of the Oklahoma State University Research Consortium for Enhancement of Well Log Data Via Signal Processing. The members of this consortium include AMOCO, Britoil, Mobil Research and Development, OXY Incorporated, Sohio Petroleum Company and Texaco. In particular, I would like to thank Dr. Robert Elphick of OXY Inc. for his personal interest in my work and for help in supplying the Hartzog-Draw data used in this study.

Most importantly, I commend my wife, Anita, and my children, Isaac and Corrie for enduring the trials and tribulations of my graduate work. Without their undying love and support, I would have never finished.

TABLE OF CONTENTS

Chapter Page

LIST OF TABLES

LIST OF FIGURES

Figure Page

Figure Page

CHAPTER I

INTRODUCTION

1.0 Motivation and Basic Problem

Wireline logs provide an objective continuous record of certain physical properties of the formation cut by a borehole. The log analyst uses this objective record of formation properties in conjunction with his knowledge of the local geology to assess the subsurface geological environment. This analysis helps answer questions ranging from basic geology to economics and constitutes a significant effort on the part of the oil industry. The present investigation is motivated by the continuing need for reliable automated log-analysis methods which assist the log analyst in his effort by extracting meaningful geological information from wireline logs.

This study approaches the analysis of wireline logs as a pattern recognition problem. Each log provides a continuous record versus depth of some geophysical property of the formation traversed by the borehole. The problem, simply stated, is to combine the information from each log into a composite 'picture' of the subsurface geological environment. Figure 1 shows one possible interpretation of the problem. Given the four input logs shown in Figures 1a and 1b, segment the borehole according to the naturally occurring structure of the wireline log data. One possible segmentation is shown in the center track of Figure 1 where each segment identified by a number corresponds to a certain wireline log characteristic. The basic premise of this work is: if natural structure is found in the log data, then it correlates to the geological environment represented by the log data. Fuzzy clustering algorithms provide the basic pattern recognition tool for determining natural structure in the wireline log data. The geological significance of the structure found in the logs is determined by comparing the clustering results with more conventional core descriptions of the same interval in the borehole. An underlying goal of this research is to determine what type of geological inferences might be made from the wireline-log data without prior reference to core descriptions or other special geological knowledge.

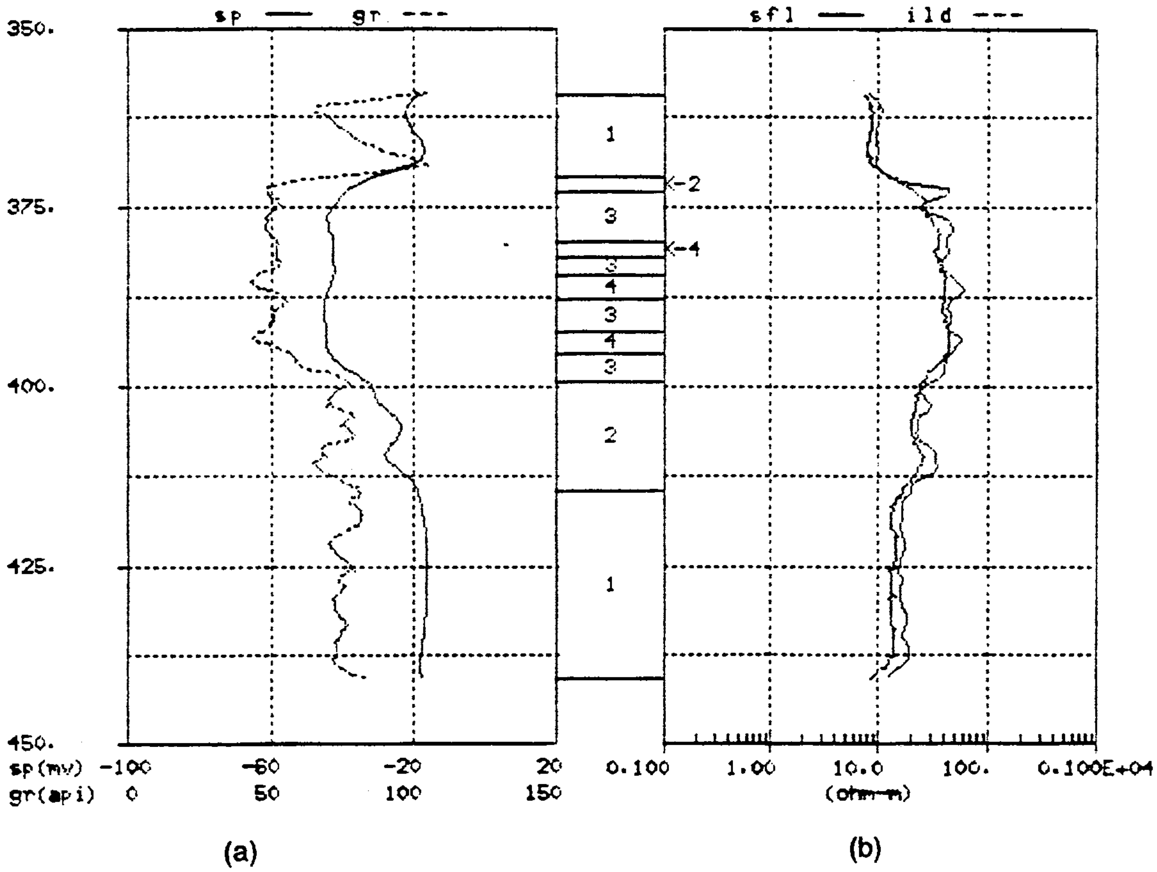


Figure 1. Borehole Segmentation Example

The remainder of this chapter deals with introductory information pertinent to a better understanding of the objectives of this work. Section 1.1 overviews basic design concepts and methodologies associated with pattern recognition systems and specifies the approach used in this study. Section 1.2 gives a simplified explanation of the physical processes represented by wireline-log data and indicates how the type of data influences the development of a particular pattern recognition model. A survey of common log analysis methods and borehole segmentation methods is given in Sections 1.3 and 1.4 respectively and finally, Section 1.5 concludes with a topical outline of the material covered in subsequent chapters.

1.1 Pattern Recognition Basics

A commonly used pattern recognition model is depicted in Figure 2. This classification model has three main elements: a sensor, a feature extractor and a classifier. The sensor either directly or indirectly measures certain physical attributes of a given physical process and this information is assimilated into a pattern vector suitable for computer processing. The feature extractor gleans presumably relevant information from the pattern vector to form a feature vector which is used by the classifier to assign the pattern vector to one of several prespecified pattern classes. Ideally, it is desired to automatically recognize and categorize incoming patterns into mutually disjoint pattern classes. Two reasons this ideal is seldom achieved in practice are: 1) the inability to choose features that completely discriminate between the various pattern classes and 2) the pattern classes are not always well defined, which results in overlapping pattern classes.

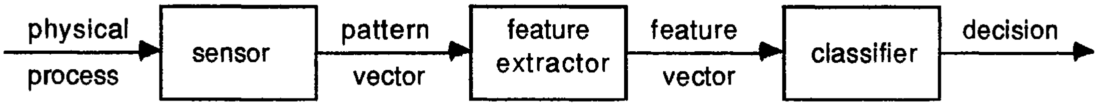


Figure 2. Generic Pattern Recognition System

Tou and Gonzalez [58] mention three design concepts for the classification model of Figure 2, namely: membership-roster concept, common-property concept and clustering concept. Each of these design concepts is motivated by how the pattern classes are characterized and defined. The following paragraphs give a brief description of these design concepts.

The membership-roster concept is the simplest design approach and is sometimes referred to as automatic pattern recognition by template matching. All possible patterns, for each class, are stored in the classifier. The classifier categorizes an incoming pattern to a particular pattern class when it matches, in some sense, one of the stored patterns belonging to that pattern class. The feature extraction step is bypassed. This design strategy is practical only when the pattern classes are very well defined.

The common-property concept is based on the idea that patterns belonging to a particular class possess common invariant attributes or features. Once these features are identified they are stored in the classifier for use in classifying unknown input patterns. Features are extracted from an incoming pattern vector and compared with the features stored in the classifier and then a decision is made categorizing the input pattern to the pattern class with similar features. The success of this approach depends on one's ability to identify a set of features that completely discriminates between the various pattern classes. The selection of features is perhaps the least scientific and most difficult part of this design approach. Feature selection is often accomplished by using a training set of pattern samples. A training set is a collection of samples representative of all the pattern classes one wishes to recognize. Each sample in the training set is labelled 'a priori' as belonging to a certain class. Once the training set is established the proposed features are computed for the samples and then these features are tested to determine which ones discriminate between the various classes. There are a number of good references that address the feature selection problem [11, 20, 58].

The third design concept is the clustering concept. The first two design concepts assumed the existence of a training set of pattern samples which were labeled to show their class membership. Procedures which use labeled training samples are said to be supervised. A more general problem involves the analysis of a collection of samples without knowledge of their classification. There are numerous unsupervised learning procedures that may be applied to unlabeled data sets. Duda and Hart [20] give an excellent introduction to unsupervised learning and clustering. In the event the system designer is able to assume: 1) the samples come from a known number of classes, 2) the *a priori* probabilities for each class are known and 3) the conditional probability densities are known, then the unsupervised learning problem becomes a parametric estimation problem. It is often difficult to satisfy these assumptions; even when these assumptions can be met the designer is still often faced with a problem of considerable computational complexity and reformulation of the problem may be desirable. Clustering procedures, which are of importance in this work, are often used when little is known about the number of classes that exist in the data and when little is known about the distribution of the data. This design concept seeks to partition the pattern space into two or more partitions according to the spatial distribution of the unlabeled data samples within the pattern space. Cluster validity measures objectively evaluate clustering performance and help determine the number of clusters that best fit the data. The value of such a method depends on how it relates back to the physical process represented by the data. Such clustering procedures may or may not result in clusters with known physical meaning. If the clusters are determined to have physical significance, then they may be considered pattern classes that provide a first step toward the design of a classifier to recognize similar patterns. The complexity of the pattern recognition system depends on the spatial distribution of the pattern classes in the pattern space. When the classes are characterized by compact, well separated clusters, then simple recognition schemes such as minimum-distance classifiers generally yield good results. When the pattern class clusters overlap, then more sophisticated methods for partitioning the pattern space must be employed. This design concept is of primary importance in this study and is discussed in detail in Chapter II.

Three basic methodologies exist for the implementation of each of the above mentioned design concepts: heuristic, mathematical and syntactic [58]. Typically, a combination of these methodologies is used for a given pattern recognition system.

The heuristic approach uses a set of *ad hoc* procedures, based on human intuition and experience, developed for specialized recognition tasks. This approach is an important part of pattern recognition system design, but little can be said about general principles in this area, since each problem requires the application of specific design rules. The success of a heuristic system is largely dependent upon the experience and expertise of the system designer.

Syntactic systems are designed using primitive elements or subpatterns and the relationships between the subpatterns. A pattern can be described by a hierarchial structure of subpatterns analogous to the syntactic structure of languages. This permits the use of formal language theory to the pattern recognition problem. These systems still require a good bit of cleverness on the part of the designer to identify the proper primitive elements and the interconnecting structure [58].

This study will focus on the mathematical approach, because this approach is better suited for analysis. Also, this approach is often a good first step toward more complex classification schemes that might involve syntactic grammars. The mathematical approach derives classification rules based on certain mathematical properties possessed by the pattern vectors, and differs from the heuristic approach which is based on a set of *ad hoc* rules. The mathematical approach may be divided into three categories: deterministic, statistical and fuzzy. Bezdek [11], does a good job outlining the differences between deterministic, statistical and fuzzy modelling techniques. Distinction among the three categories depends upon the source of uncertainty of the physical process being observed. A process is deterministic if its outcome can, with absolute certainty, be predicted upon replication of the circumstances defining it. Any uncertainty associated with a deterministic process arises from an inability to monitor the process exactly. Statistical uncertainty arises when the process under consideration is believed to be random. In this case, there is an element of chance concerning the outcome of the process, which is distinct from any imprecision in monitoring the process. The source of uncertainty in a physical process is important because it dictates the assumptions supporting the mathematical structure of the model chosen to represent the process. The point is that deterministic and statistical models transmit different types of information about the processes they represent. Fuzzy models are sometimes used when the observed process is judged to be neither deterministic nor random. Bezdek [11], motivates the plausibility of fuzzy models with the following example.

Consider the question, "is the person x nearly two meters tall?" The use of the word "nearly" introduces a source of nonstatistical uncertainty or fuzziness as to the proper response to the question. Obviously, a "yes" or "no" response is expected to the question. There are several ways to model this problem in a mathematical framework. Let X be a sample of n people and define A to be a subset of X such that,

$$A = \{ x \in X \mid 1.995 \leq h(x) \leq 2.005\} \qquad (1)$$

where h(x) is the observed height of x. This deterministic approach equates membership in A with being nearly two meters tall by defining a tolerance of 0.005 meters. The characteristic function for set A is given by:

$$u_A(x) = \begin{cases} 1; & x \in A \\ 0; & \text{otherwise} \end{cases} \qquad (2)$$

By observing the height of x one can make a "yes" or "no" decision based on the value of the characteristic function in Equation (2). A statistical approach the problem would try to determine the probability that $x \in A$, $Pr(x \in A)$. Let X be part of a larger population S and define the random variable, h: X --> $(0, \infty)$ as the height of x. Consider the event $1.995 \leq h(x) \leq 2.005$. This event is identical to set A in Equation (1) but is stated in a statistical setting. It follows that by a suitable set of experiments on population S and by the proper statistical inference a probability can be assigned to each $x \in X$, of being in A. This leads to an estimate of the probability of the stated event: $Pr(1.995 \leq h(x) \leq 2.005) = Pr(x \in A)$. This type of model tells us the chance of any particular element of X being in set A. For example, suppose $Pr(x \in A) = 0.95$. This is not satisfactory for responding to the question in a "yes" or "no" manner, because $Pr(x \in A) = 0.95$ does not preclude the possibility that h(x) is far removed from 2. One difficulty with this approach is that an element of chance is attached to the phrase "nearly two meters" when this is not warranted. A third approach allows for the natural fuzziness in the stated question. Since set membership is key to the decision making process let

$$B = \{ x \mid x \text{ is nearly two meters tall}\}. \qquad (3)$$

Since B is not a conventional set, there is no set theoretic realization for it; however, it is possible to visualize a function theoretic representation. Let u_B: X --> [0,1] be a function whose values, $u_B(x)$, give the grade of membership of x in the fuzzy set B. This is a natural extension of the set theoretic relationship given for set A in Equation (2). In this example, $u_B(x)$ indicates the degree to which h(x) is close to the value 2. There are many possible functions which would do this and Bezdek [11] uses the discrete function given in Equation (4). This approach gives more quantitative information than

$$u_B(x) = \begin{cases} 1.00; & 1.995 \leq h(x) \leq 2.005 \\ 0.95; & 1.990 \leq h(x) < 1.995 \text{ or } 2.005 < h(x) \leq 2.010 \\ & \\ \cdot & \\ \cdot & \end{cases} \qquad (4)$$

the first two approaches about how close h(x) is to 2 and lets the user answer the question "yes" or "no" based upon the value of $u_B(x)$. Bezdek [11] points out that once a particular model is developed, its usefulness and capabilities vary and different models give different types and amounts of information about the process being investigated.

Bezdek [11] cites several good references for a more complete discussion of uncertainty and the plausibility of fuzzy models.

The previous example serves to illustrate how one's perception of the physical process being investigated influences the choice of mathematical model used to describe the process. The present investigation is concerned with the segmentation of a borehole based on the corresponding wireline log responses. It is expected that the resulting segments will have physical meaning inasmuch as the wireline logs reflect the geological environment in the borehole. The approach taken is a very general one. No assumptions are made concerning the number of different geological environments which might exist in the borehole or the statistical distribution of the wireline log data. The notion of a fuzzy model seems to be appropriate for this geological application. The number of environments that might be encountered in a particular borehole is seldom well defined and many transitions from one environment to the next environment are gradational. For example, consider the description of an interval within a borehole which is primarily a sandstone-shale sequence. Provided prototypes representing the sandstone and shale classes have been identified, the fuzzy model would allow each pattern vector to have a membership distributed between the sand and shale classes rather than classifying the pattern vector as belonging entirely to one class or the other, based on probability of membership. This approach identifies pattern vectors that possess attributes of both the sandstone and shale classes. In general, the fuzzy model is not restricted to a fixed number of prespecified pattern classes, but can identify pattern vectors that have attributes of several classes. This is a desirable characteristic for the application at hand. Since there is a certain amount of nonstatistical uncertainty associated with geological classifications, pattern recognition using fuzzy objective function clustering algorithms is used for the automatic segmentation of a borehole based on wireline log responses.

1.2 Wireline Logs

To develop a pattern recognition model that identifies meaningful structure in a data set it is helpful to understand the physical processes involved in generating the data. This section gives a simplified discussion of open-hole logs, with emphasis given to the formation properties that have the greatest effect on logging measurements. Several good references give the detailed theory of operation for the various logging tools and interpretation principles for the curves recorded from these tools [3,5,6,32,51,60].

The Society of Professional Well Log Analysts defines a wireline log as the product of a survey operation that provides one or more physical measurements as a function of depth in a well bore [25]. Ordinarily, the survey operation is conducted shortly after the completion of drilling activity. A logging tool or sonde is attached to the end of a wireline and lowered into the borehole. As the sonde is pulled out of the borehole, its response is sent via the wireline to a logging truck, where the signal is conditioned, digitized and recorded for display in a readable log format [6]. The digitized log values represent samples taken at six-inch intervals. Figure 3 shows the logs: spontaneous potential(SP), gamma ray(GR), spherically focused(SFL), deep induction(ILD), neutron porosity(NPHI) and interval transit time(DT) in a typical log format. These logs are representative of the data used in this work and a brief description of each kind follows.

The first track in Figure 3 displays the SP and GR curves. Notice the similarity between the two curves even though the respective logging tools measure entirely different properties of the formation. The SP measures the difference in electrical potential between an electrode in the borehole and a surface electrode whereas the GR measures the naturally occurring gamma [5].

The SP readings are given in millivolts and provide a crude indication of formation permeability. Opposite impermeable shales the SP is relatively constant and is referred to as "the shale baseline". In permeable zones the direction and magnitude of deflection of the SP curve depend primarily on the relative ion content of the formation water and the drilling fluid. The SP works best where the drilling fluid('mud') is fresher than the formation water. In such cases the SP curve deflects to the left opposite permeable formations and permits easy sand-shale discrimination. These deflections give qualitative information concerning permeability, since there is no definite correlation between the amplitude of the curve and the degree of permeability of the formation [5,51].

The GR is recorded in American Petroleum Institute(API) Gamma Ray units. The detector-measurement systems of all primary service companies are calibrated to this standard unit in the regulation API test pit at the University of Houston. The three common radioactive elements found in nature are uranium, thorium and potassium 40, with potassium being the most abundant in the earth's crust [5]. In clay materials potassium is abundant, as compared to potassium in other sedimentary rocks. Clay, when compacted, forms shale; therefore the GR log generally reflects the shale content of sedimentary formations. Shale-free or 'clean' sandstones and carbonate rocks normally

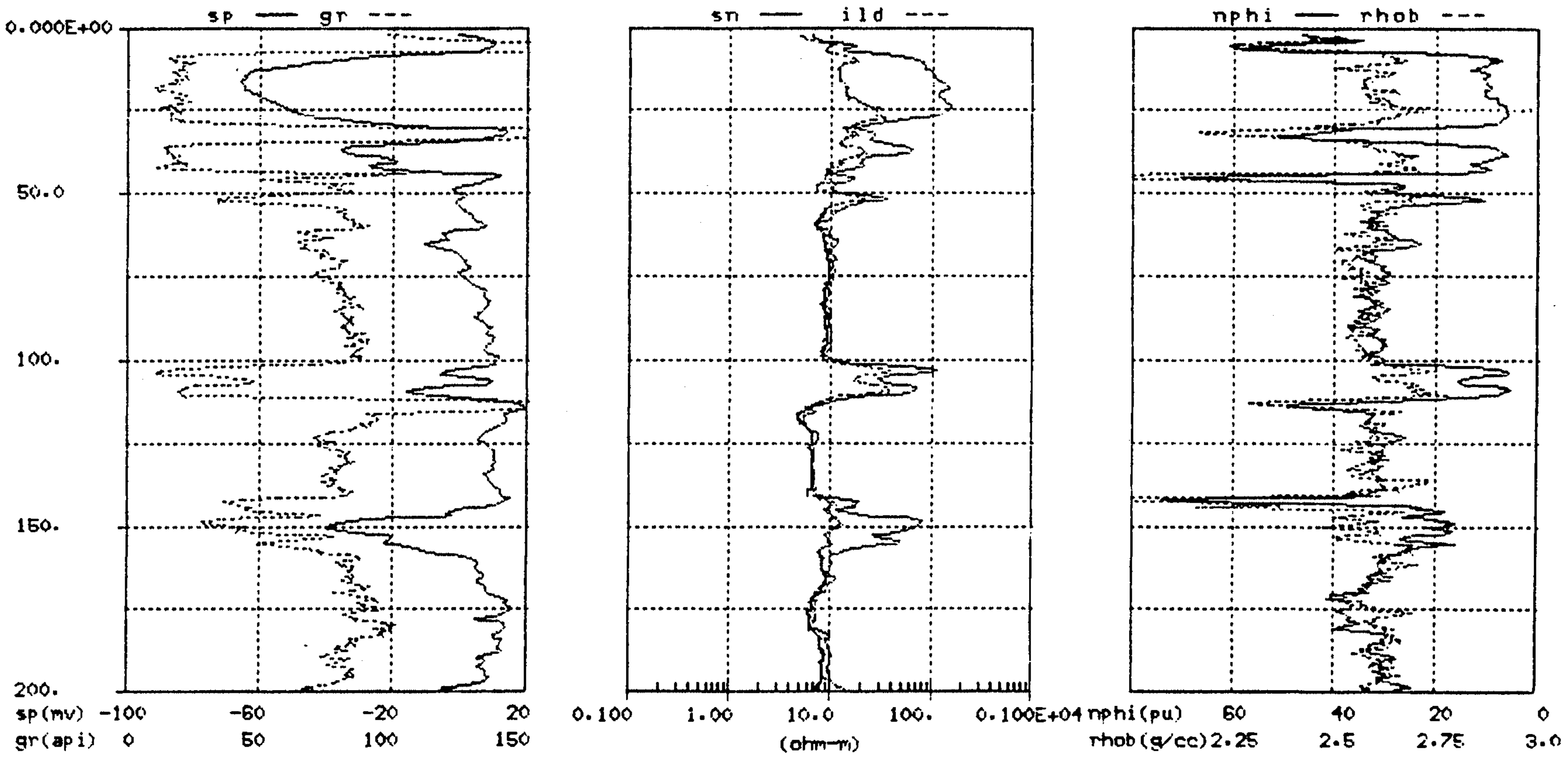


Figure 3. Typical Log Format

exhibit low GR response. One primary application of the GR log is lithology identification [60].

The second track in Figure 3 displays two resistivity logs, the SN and ILD measured in ohm-meters. Two applications of resistivity logs are to determine hydrocarbon-bearing versus water-bearing zones, and indicate permeable zones [3]. The resistivity of any formation is a function of the amount of water in that formation and the resistivity of the water. Ion-bearing water is conductive, whereas the rock matrix and hydrocarbons act as dielectrics. The various resistivity-logging tools record the resistivity at different depths of investigation into the formation. For example, the SN measures resistivity about one foot into the formation, but the ILD measures the resistivity several feet into the formation [32]. In tight impermeable formations the resistivity curves tend to read similar values, but in permeable formations there is separation between the SFL and ILD curves due to invasion of drilling mud into the formation [3]. In the case of fresh drilling mud the SFL will read a higher resistivity than the ILD. By far the most important application of the resistivity logs is the detection of hydrocarbon-bearing zones. In formations with 100% water saturation, the resistivity is at a minimum for a given porosity and rock structure. Any increase in the amount of hydrocarbons within the pore space will increase the respective resistivity readings. If two porous, permeable zones exist within a formation, one showing appreciably higher resistivity readings than the other, and all else being equal, then the higher resistivity is most likely due to the presence of hydrocarbons.

The last track in Figure 3 shows the neutron porosity(NPHI) and bulk density log(RHOB) porosity logs. The interval transit time(DT) is a third porosity log not shown in Figure 3. All three porosity logs are primarily responsive to formation porosity, yet other formation characteristics also influence these measurements. Each of the logging tools responds differently to the effects of lithology as well as to the amount and type of fluids in the pores. The differences among the porosity log measurements allow them to be used in various combinations to determine specific lithologies, porosity and, under certain circumstances, type and amount of fluid in the pores [60]. It should be noted that unusually large porosity indications may be caused by washouts in the borehole,.and should not be attributed to the formation.

Neutron-porosity logging devices respond to hydrogen atoms in the formation pore space. Since the only hydrogen in a clean(shale free) formation is due to the presence of water or hydrocarbons in the pore space, there is a relationship between the response of the neutron logging device and the formation porosity. Both water and oil contain about the same amount of hydrogen per unit volume and the neutron tool does not permit differentiation between them. Gas has a lower hydrogen density and is characterized by a low neutron-porosity reading. Neutron readings will indicate a higher porosity than actually exists in formations which contain hydrogen in the rock matrix or as dispersed solids in the pore space. In formations containing significant amounts of clay or shale the NPHI values will be inflated, due to the hydrogen content of the bound water contained in the shale. This is a limitation when using the neutron log alone, but in conjunction with the other porosity logs it is sometimes useful to identify mixed lithologies [3,5,60]. Neutron readings are also affected by lithology, and since the lithology is usually not known, the neutron device is run assuming a limestone matrix and porosity is recorded in.limestone porosity units(p.u.). Standard procedures exist to correct the NPHI values when the matrix is known to be something other than limestone.

The sonic porosity tool is based on the interval transit time(DT)-- the time required for a compressional sound wave to travel one foot through the formation. DT is measured in microseconds per foot(μs/ft) and depends on the lithology, porosity and fluid type of the formation. In general, dense formations with small amounts of porosity have small travel times, and increasing travel times indicate increasing porosity for a given lithology and fluid type. DT represents the shortest travel time through the formation and indicates primary formation porosity, which may be the same as total porosity. Comparison of DT with NPHI and RHOB will help clarify whether secondary porosity is undetected by the sonic porosity tool [3,5,60].

In summary, logging tools respond primarily to the chemical nature of the rock matrix and the pore fluids. One or more log responses are affected by: lithology, porosity, permeability, shale volume and water and hydrocarbon saturations. Logs provide formation data not directly accessible by means other than coring and can be used as an exploration tool to describe local stratigraphy, structure, and environments of deposition [60].

1.3 Log Analysis Methods

Log data constitutes a 'signature' of the formation that provides valuable geological information, assuming that there exists well defined relationships between what is measured by the logs and formation parameters of interest to the geologist and reservoir engineer. Two basic assumptions are implied in log analysis: 1) a significant change in any geological characteristic will manifest itself in a physical parameter detectable by one or more logs; and 2) any change in log response indicates a change in at least one geological parameter[3,32]. Two exceptions to these assumptions are noted below.

First, a log response may reflect a change in borehole conditions, rather than a change in some formation parameter. For example, any contact device, such as a density porosity tool, requires good contact between the detector pad and the borehole wall for accurate readings; therefore borehole rugosity directly affects the measured values of such a device. Certain borehole conditions are correctable with the aid of service company chart books; however, the value of such correction charts is limited because of the difficulty in satisfying the assumptions for their use [32]. When log values unduly influenced by the borehole environment are detected, and can not be corrected, then they should be eliminated from the log-analysis procedure.

A second exception occurs in sharp transitions between beds and in thin beds due to limitations in vertical resolutions of the various logging tools. Vertical resolution refers to the minimum thickness of formation that can be distinguished by a logging tool under operating conditions [25]. Thin beds influence the values recorded by a logging tool, but are not thick enough to yield discrete signatures of beds and representative measurements of attributes. In the absence of core information or other special knowledge, it is unreasonable to characterize the geological attributes of thin beds based solely on wireline log responses. Yet, in spite of their limitations, wireline logs provide an objective quantitative measure of the subsurface geological environment.

Wireline logs have been a valuable geological tool ever since their inception in 1927 by Conrad Schlumberger. Some often used methods for extracting geological information from wireline logs include crossplotting techniques, discriminant analysis, cluster analysis and principal component analysis. These methods are discussed briefly below.

1.3.1 Crossplot Techniques

Crossplot techniques plot key log responses against each other and then the analyst seeks manually to correlate significant clusters of points or significant trend lines with particular geological characteristics. Crossplot techniques have a variety of applications but tend to be subjective and applicable in limited situations.

Since Savre [49] suggested the use of sonic, neutron and density logs for more accurate determination of porosity and mineralogy in complex lithologies, the crossplotting of porosity logs has become a standard interpretation technique for describing porosity and lithology, particularly in carbonates [60]. This is perhaps the most common use of the crossplot technique. The scope and limitations of crossplotting porosity logs are outlined in lessons 15, 16 and 17 of the Dresser-Atlas Home Study Course [60]. Another application of crossplotting is cited by Almon [2] and involves work done by Bedwell [8] and Carloss [13], who used crossplot techniques to identify depositional environments. Almon characterizes this work as subjective and lacking sufficient detail for good discrimination among the major depositional environments [2]. Crossplot techniques are easy to use and are often a first step in trying to extract sedimentological information from wireline logs. Priisholm and Michelsen [44] use porosity logs and crossplot techniques as part of their method for lithology determination, lithostratigraphy and basin analysis in the Norwegian-Danish basin. Watney [59] uses gamma ray-neutron crossplots to facilitate the understanding of the sedimentological variation in the Missourian sequences of northwestern Kansas. When crossplotting yields good results for a particular application, it often leads to the development of more objective analytic methods, such as discriminant function analysis. For example, Meyer and Nederlof [39] use various porosity/resistivity crossplots for the identification of source rocks. Specifically, sonic transit time/resistivity and density/resistivity crossplots were used as a basis to discriminate between source rocks and non-source rocks. These crossplots were used to develop linear discriminant functions to distinguish between the two rock classes; they are discussed in Section 1.3.2 [39].

1.3.2 Discriminant Analysis

Discriminant function analysis is a multivariate statistical means of differentiating among members of various groups or classes, based on statistical observations of the members within the respective classes. The success of this classification scheme depends on the form of the discriminant functions and one's ability to determine the coefficients for these functions [20,58]. This method of classification requires a training set of sample patterns. Meyer and Nederlof [37] provide a simple two class example of discriminant analysis. One hundred sixty-nine rock samples were divided into two classes based on geochemical analyses: class 1 were petroleum source rocks(71 samples), and class 2 were non-source rocks(98 samples). A crossplot of these samples, shown in Figure 4, uses sonic transit time and resistivity log values to determine a decision boundary separating the two classes by using psuedoregression.

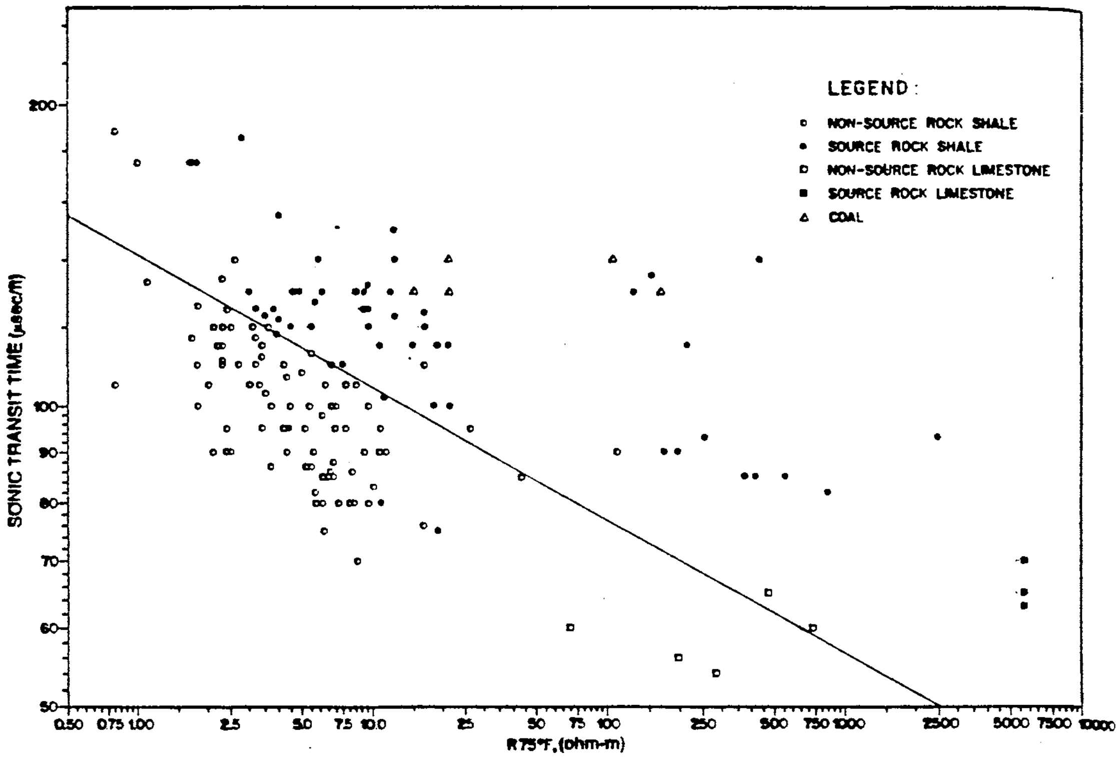


Figure 4. Example of Discriminant Function Analysis for Two Classes (taken from Meyer and Nederlof [39])

The decision boundary is called a discriminant function and has the form:

$$D = w_1(\text{sonic log value}) + w_2(\text{resistivity log value}) + w_3 ,$$

where w_1,w_2 and w_3 are regression coefficients and D is the discriminant score. The discriminant function transforms the sonic transit time and resistivity values into a single number, the discriminant score. In this example, positive discriminant scores indicate source rocks and negative scores indicate non-source rocks. A discriminant score of zero is indeterminate. Meyer and Nederlof [39] followed a similar procedure using a density/resistivity crossplot. It is interesting to note that a 91% correct classification rate was attained when the discriminant function was used to classify the 169 samples in the training set. It is not difficult to visualize situations in which the sample patterns are not linearly separable as in the above example. The development of generalized decision functions is covered in the literature [58]. Once the form of the decision function has been specified, then the problem becomes one of determining the coefficients for the function. This is typically done using a training set of labeled sample patterns. The important point is that discriminant analysis assumes knowledge of the classes in order to construct a discriminator for the classification of future observations. Almon [2] applied this method in an attempt to discriminate among six sedimentary facies on the basis of wireline log responses. Almon's [2] study is of particular interest because it used data from the Shannon Sandstone, Hartzog Draw Field, Wyoming. Similar wireline log data is used in Chapter IV of this work and certain comparisons are made to Almon's [2] results. A training set of 89 core samples and corresponding log values was taken from three wells and used to generate three linear discriminant functions, which effectively separated the six classes of data. Almon [2] used as few as three samples and as many as 34 samples to characterize the respective data classes. This seems a rather modest training set for a statistical analysis method. It should also be noted that no attempt was made to standardize the wireline log data from well to well. The three discriminant functions were applied to the training set data and resulted in a 98 percent correct classification of the data. These same discriminant functions were then applied to wireline log data from eight other wells in the Shannon Sandstone and Almon [2] claims a 94 percent correct classification of the data from these wells. A footnote should be added at this point to indicate that two of the six facies types were not present in any of these eight wells, a third facies was nominally present in two of the eight wells and this third facies went undetected by the discriminant function analysis. The success of Almon's[2] method is tempered by the fact that it was effectively demonstrated on three of the six classes of data. This work at least alludes to the potential of a multivariate statistical method such as discriminant analysis. Discriminant function analysis has also been applied to other geological problems such as uranium exploration and the determination of clays in shale [7,42,43].

1.3.3 Cluster Analysis and Principal Component Analysis

The analytic methods of cluster analysis and principal component analysis are considered in tandem since they are the primary analyses used in this study. These methods are not new; they have been applied to geological problems for the past two decades [17,19,38,41,53,56]. Cluster analysis has been mentioned briefly in Section 1.1 as a means to characterize the various data classes by their clustering properties in the pattern space. Principal component (PC) analysis can be regarded as a dimension-reducing tool, asking the question: "Are there a few functions of the many original variables which in some sense capture the essential variability in the data?" [1]. Principal components are nothing more than eigenvectors of a variance-covariance or a correlation matrix [18,40]. The interpretation of principal components is subjective and should be tested properly with an appropriate independent data set. Some geological literature uses the term 'factor analysis' in place of 'principal component analysis' and what are called 'factors' in one article might be called 'components' in another article [19]. Chapter II outlines the specific application of cluster analysis and principal component analysis to discrete wireline log data.

There are many software packages used by log analysts to aid them in their analysis and interpretation of wireline logs. One of particular interest is Faciolog, a Schlumberger product. This software package is introduced because its analytic tools include PC analysis and cluster analysis [63]. These analysis tools parallel those outlined in Chapter II of this thesis, but the present study is an independent effort and differs from Faciolog at several significant points. Also of importance is a multiwell Faciolog evaluation of wells in the Hartzog-Draw Field, which will provide a basis of comparison for some of the analyses performed in Chapter IV of this study [61]. The following paragraphs contrast the Faciolog technique with the methodology used in this study.

Faciolog uses principal component analysis and cluster analysis to zone a well into 'electrofacies' [63]. An electrofacies is defined "as a set of log responses characterizing a sediment" [63]. A set of input logs is chosen and corrected for environmental effects.

The input logs are then weighted by the user. This weighting process is done largely by trial and error. The logs are then normalized taking into account their respective standard deviations over the interval of interest. New orthogonal axes, called principal component axes, are defined in the space created by the normalized logs. The origin of the PC axes is the center of gravity of the normalized log data and the PC axes are oriented such that PC axis 1 is in the direction of maximum variation of the normalized log data, PC axis 2 is in the direction of next greatest variation of the data and so on for the remaining PC axes. PC logs are then derived by projecting the normalized log data on these PC axes. Once the PC logs are computed, a process of finding small clusters or local modes takes place. The open literature is not clear on the exact nature of this clustering process. These small clusters are then manually grouped into larger clusters, which are identified as electrofacies. The Faciolog results are then presented in a suitable display [63].

The methodology used in this study is outlined at the beginning of Chapter III but the basic methodology is given here and contrasted to the Faciolog technique. First, each input log is scaled using a scale factor equal to the largest excursion from the mean of the original log. This differs from the normalization process used in Faciolog. Second, the PC logs are derived by applying the Karhunen-Loeve Transformation(KLT) to the scaled input logs. This may or may not differ from the Faciolog process described in the literature. The exact mathematical process is not given by Schlumberger, but is described in general terms. Third, there is no weighting of the input logs by the user; the weighting is done automatically by the KLT. Fourth, the clustering process is done using a Fuzzy-C-Means(FCM) clustering algorithm and validity measures are used to indicate the number of clusters which best fit the data. There is no manual grouping of clusters.

1.4 Segmentation Methods

Hawkins and ten Krooden [31] review a variety of segmentation techniques as they apply to various univariate and multivariate geological signals. Competitive algorithms are evaluated on the basis of statistical optimality and numerical computational requirements. The segmentation methods given in order of statistical optimality are: maximum-likelihood, hierarchic optimization and split-moving window. This order may be reversed when evaluating the methods on the basis of numerical computational requirements. The basic notation and models for each of the above mentioned methods is given here for univariate data with the extension to multivariate data given by Hawkins and ten Krooden [31]. All the methods given in this section assume that the number of segments is known and that the data within each segment is normally distributed. Therefore, the problem is one of estimating the mean and variance of each segment such that within segment variance is small and between segment variance is relatively large. This is analogous to unsupervised learning using parametric estimation methods.

Let $\{X_1, X_2, \dots , X_N\}$ denote a set of N discrete scalar values taken from the geological signal of interest. The problem is to segment these N data values into k distinct homogeneous segments. A segmentation of this data into k homogeneous segments consists of determining breakpoints, $0 = b_0 < b_1 < \dots < b_k = N$, such that the $\{X_j\}$, $b_{i-1} + 1 \leq j \leq b_i$, are in some sense homogeneous. All of the methods reviewed by Hawkins and ten Krooden [31] define homogeneity with respect to a normal statistical model.

$$X_j \sim N(\xi_i, \sigma^2{}_i), \qquad j = b_{i-1} + 1, \dots , b_i \qquad (5)$$

It is generally assumed that the means, ξ_i, differ significantly from one segment to its neighbor. In terms of the variances there are two broad classes of problems, the homoscedastic problem and the heteroscedastic problem. The former problem assumes the variances are the same for all segments and the latter problem allows for the possibility that the variances differ from one segment to another. Hawkins and ten Krooden [31] show that the numerical computation for the homoscedastic and the heteroscedastic models is approximately the same and therefore the heteroscedastic model is more appealing on the grounds of greater generality. The drawback of the heteroscedatic model is its greater sensitivity in departures from the assumed normal distribution of the data. All of the methods use the following quantities.

$$\bar{x}_{l,m} = \sum_{i=l+1}^{m} X_i/(m-l) \qquad \text{(sample mean)}$$

$$S_{l,m} = \sum_{i=l+1}^{m} (X_i - \bar{x}_{l,m})^2 \qquad \text{(sample variance)}$$

$$Q_{l,m} = \begin{cases} S_{l,m} & \text{for a homoscedastic model} \\ (m-l)\ \log\ (S_{l,m}) & \text{for a heteroscedastic model} \end{cases}$$

The maximum-likelihood segmentation method determines the set of breakpoints, $\{b_1, \ldots, b_{k-1}\}$, that maximizes the likelihood function given in (5). This is done by minimizing

$$\sum_{i=1}^{k} Q_{b_{i-1}, b_i} \qquad (6)$$

for all possible sets b_i [31]. It is shown that this minimization process may be performed using the optimization method of dynamic programming. From a statistical point of view the maximum-likelihood method is recognized as the best general estimation method; therefore this method may be considered statistically optimal provided there are adequate computer resources to carry out the minimization. This method requires on the order of N(N+k) computations [31]. If the number of sample points is large, then a suboptimal method which takes less computational time may be desirable.

The hierarchic optimization methods are considered the next best from a statistical perspective and three such methods are considered by Hawkins and ten Krooden [31]. The hierarchic disaggregative, hierarchic aggregative and the stepwise method are discussed with respect to their methodology and relative computational requirements. These methods are similar to hierarchic clustering methods.

For the hierarchic disaggregative method, suppose that at some iteration one has identified the changepoints $b_1, \ldots, b_i$, and wishes to determine whether any additional changepoints are present. For each segment j, $1 \leq j \leq i$, determine

$$\max_{m \text{ in segment } j} \{(Q_{b_{j-1}, b_j}) - (Q_{b_{j-1}, m}) - (Q_{m, b_j})\}, \qquad b_{j-1} < m < b_j \qquad (7)$$

and then determine the maximum of these i maxima. If this maximum is sufficiently large, then declare the corresponding m to be a breakpoint and add it to the set of breakpoints, $\{b_1, \ldots, b_{k-1}\}$.

The hierarchic aggregative method involves the deletion of breakpoints rather than the addition of breakpoints done in the disaggregative method. Suppose the set of breakpoints consists of $\{b_1, \ldots, b_i\}$, then for each breakpoint determine

$$\min_{1 \leq j \leq i} \{(Q_{b_{j-1}, b_{j+1}}) - (Q_{b_{j-1}, b_j}) - (Q_{b_j, b_{j+1}})\} \qquad (8)$$

and if the minimum is sufficiently small then segments on either side of breakpoint, b_j, are merged.

The stepwise method is a combination of the aggregative and disaggregative methods. At each iteration,

1) the most similar pair of adjacent segments are merged if they are not sufficiently different according to criterion (8), otherwise

2) the existing segments are tested for possible subdivision according to criterion (7) and the least homogeneous segment is split if (7) exceeds a specified threshold.

In terms of computational requirements the disaggregative method requires computations on the order of N log k, while the aggregative and stepwise methods require computations on the order of N+k [31]. N is the number of discrete points taken from the signal and k is the number of segments to be determined.

One other method which is the least attractive in a statistical sense is the split-moving window method. This method attempts to locate one change point at a time and requires the user to prespecify the window width, 2h. Then for every i, $h \leq i \leq N-h$, compute the value: $(Q_{i-h,i+h} - Q_{i-h,i} - Q_{i,i+h})$. If for any i this value exceeds some preset threshold, then i is concluded to be a breakpoint. One major disadvantage of this method is the specification of the window width, 2h. It is desirable to have the window size as large as possible but, if the window size is too large then more than one breakpoint may be contained within the window and performance can be inhibited badly. The computational requirements for this method are relatively minor and are on the order of N [31].

Hawkins and ten Krooden [31] conclude that the maximum-likelihood method is best provided the number of points, N, is moderate. For large N, the hierarchic methods yield, in general, more reliable results than the split moving window method. One disadvantage associated with all the above mentioned methods is that the number of segments must be at least approximated prior to any analysis.

1.5 Thesis Overview

The remainder of this thesis is concerned with the development and testing of a segmentation algorithm. Chapter II describes the mathematical tools of cluster analysis and principal component analysis that are used in the development of the segmentation algorithm. These analytic tools are demonstrated on a well known botanical data set for illustration purposes, and to validate the software being used. Chapter III begins with an outline of the methodology used in the segmentation algorithm. A series of examples is presented in Chapter III to evaluate the performance of the borehole segmentation procedure. The physical significance of the segments is evaluated by comparing borehole segmentation results with a core description of the same interval. Chapter IV extends the segmentation process to a multiwell environment using data from the Hartzog-Draw Field, Wyoming. The investigation in Chapter IV is motivated in part by this question: "Is it possible to automatically find geologically similar segments between wells in the same field when such are known to exist?" If so, is it then possible to design a classifier to reliably identify similar segments in other wells in the same field? Chapter V summarizes the results of this investigation and enumerates possible extensions of the work done to date.

CHAPTER II

PATTERN RECOGNITION TOOLS

2.0 Introduction

This chapter outlines the basic mathematical tools that are used in the development of a borehole segmentation algorithm. Section 2.1 gives a brief overview of cluster analysis methods and then proceeds to outline the differences between hard and fuzzy clustering algorithms. An example using a famous botanical data set demonstrates the character of a Fuzzy-c-Means(FCM) clustering algorithm. This same example also demonstrates the use of cluster validity measures to objectively evaluate clustering results. Section 2.2 outlines how the Karhunen-Loeve Transform(KLT) generates a set of principal component logs from a set of scaled wireline logs. For continuity of presentation, the data set of Section 2.1 is used again to demonstrate how the KLT can be used to represent a discrete data set in terms of its principal component features. Also illustrated in Section 2.2 is the effect of two linear scaling procedures on the derivation of principal components. The chapter concludes with a summary of the various tools which will go into the development of a borehole segmentation algorithm.

2.1 Cluster Analysis

The history, general philosophy and many specific techniques of cluster analysis may be found in a number of good texts [11,20,58]. The bibliographical and historical remarks given by Duda and Hart [20], along with their listed references, are especially helpful for locating specific topics in cluster analysis. This discussion will be limited to information pertinent to subsequent sections of this thesis and draws heavily from Bezdek [9,11].

Clustering methods can be categorized according to: 1) axiomatic bases, 2) clustering criterion and 3) similarity measures [11]. The axiomatic basis categorizes clustering methods into deterministic, stochastic or fuzzy methods whereas the clustering criterion subdivides the methods into hierarchic, graph-theoretic or objective functional methods. The choice of similarity measure further subdivides clustering methods. The way in which similarity between two sample points is measured directly affects the shapes of the resulting clusters. Fuzzy clustering methods are the preferred method for the present application, as stated at the end of Section 1.1. The choice of clustering criterion is dependent on the geometrical structure of the data set being investigated. This structure is a function of the physical processes generating the data. Some insight into which clustering criterion might be best for wireline log data is gained by reviewing past applications for the three criteria mentioned above. A discussion of similarity measures is delayed until Section 2.1.1.

Hierarchic clustering methods had their origin in biological taxonomic studies, where much of the early work in clustering was done [20]. There are agglomerative and divisive techniques. Examples of both techniques are given in Section 1.4. Hierarchic methods have the characteristic of nested clusters and for every hierarchic clustering there is a corresponding tree, called a dendrogram, that shows how the sample patterns are grouped. It is easy to see that these clustering methods are well suited for biological taxonomy where individuals are grouped into species, species into genera and so on.

Graph-theoretic clustering methods consider the data set under investigation to be a set of isolated nodes or points. These methods tend to use some measure of connectivity or bonding between groups of nodes in the clustering procedure. Such techniques are preferred when the data are believed to have a linear or a psuedolinear structure. At present, there is no uniform way of formulating clustering problems as graph theory problems and use of these ideas is still very much an art [20].

Objective function clustering methods allow the most precise mathematical formulation of the clustering problem [11]. The quality of a particular partitioning of the data is measured by an objective or criterion function. The "optimal" clustering of the data is achieved when the objective function is extremized. Most clustering methods of this type have either explicitly or implicitly accepted some type of minimum-variance objective. All the methods discussed in Section 1.4 use a minimum-variance criterion to measure the quality of the resulting segments or clusters of data. Objective function algorithms, using some type of minimum-variance criterion, are believed to work best when the data form essentially compact clouds that are relatively well separated from one another [11,20]. One of the pitfalls of minimum-variance methods is that the best clusters, as measured by the objective function, do not necessarily have a good physical interpretation. Numerous examples in the literature illustrate this shortcoming and most often this problem arises when there is a large disparity in the number of samples in different clusters [11,20]. Three alternatives for attacking this problem are: the choice of a different objective function, the addition of heuristics for splitting and merging clusters or the reformulation of the clustering problem.

The geometrical structure of wireline log data lends itself best to the objective function clustering criterion. The overview of wireline logs in Section 1.2 states that logging responses are primarily affected by lithology, porosity, permeability and pore fluids of the formation penetrated by the wellbore. Where these characteristics are relatively constant over an interval the corresponding digitized log values tend to cluster in clouds of points. Changes in these geological characteristics can be either abrupt or gradual depending on the forces at work at the time of deposition. When the changes in geological environment are gradual the delineation of where one environment stops and another environment begins is unclear; therefore the clusters of log values representing this changing environment will not be well separated. This complicates the cluster analysis problem. However, the initial method of choice for the present application is still a fuzzy objective function algorithm. The remainder of Section 2.1 introduces necessary notation and theory leading to the Fuzzy-c-Means(FCM) algorithm which will be a basic pattern recognition tool used in the analysis of wireline log data.

2.1.1 General Notation and Hard Algorithms

Let the data set $\mathbf{X} = \{\underline{x}_1, \underline{x}_2, \dots, \underline{x}_N\} \subset \Re^p$, be a finite subset of real p-dimensional Euclidean space with cardinality equal to N. Each $\underline{x}_k = (x_{k1}, x_{k2}, \dots, x_{kp}) \in \Re^p$ is a pattern vector of data set $\mathbf{X}$, with x_{kj} being the j-th observation of the k-th measured characteristic of members of some physical population being investigated. This leads to the following definition for a hard c-partition of data set $\mathbf{X}$.

Definition 1 [11]

A conventional hard c-partition of data set $\mathbf{X} = \{\underline{x}_1, \underline{x}_2, \dots, \underline{x}_N\} \subset \Re^p$ is represented by a matrix $\mathbf{U} = [u_{ik}]$ when and only when:

1. $u_{ik} \in \{0,1\}$; $\quad 1 \leq i \leq c$, $1 \leq k \leq N$

2. $\sum_{i=1}^{c} u_{ik} = 1$; $\quad 1 \leq k \leq N$

3. $0 < \sum_{k=1}^{N} u_{ik} < N ; \qquad 1 \leq i \leq c.$

Matrix **U** has: elements that are either 0 or 1, columns that sum to 1 and rows that sum to a value strictly between 0 and N. Condition 3 assures each partition has at least one member. The ik-th element of **U**, u_{ik}, represents the membership of the k-th sample point in the i-th partition.. If $u_{ik} = 1$, then the k-th sample point is a member of the i-th partition or i-th cluster. For example, if $\mathbf{X} = \{\underline{x}_1, \underline{x}_2, \underline{x}_3\}$, then there exists only three possible hard 2-partitions of **X**. $\mathbf{U}_1$ partitions the data set **X** such that $\underline{x}_1$ and $\underline{x}_2$ are in one partition and $\underline{x}_3$ is in a second partition. $\mathbf{U}_2$ partitions $\underline{x}_1$ and $\underline{x}_3$ from $\underline{x}_2$ and $\mathbf{U}_3$ partitions $\underline{x}_2$ and $\underline{x}_3$ from $\underline{x}_1$. Notice that permuting the rows of **U** simply reorders the partitions and does not represent a different partitioning of the data set, **X**

$$\mathbf{U}_1 = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \qquad \mathbf{U}_2 = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix} \qquad \mathbf{U}_3 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 1 \end{bmatrix}$$

In general, the number of possible hard partitons for a given data set is extremely large. Recall that the data set **X** has N elements and let M_c be the set of admissible solutions for the conventional(hard) cluster analysis problem. The magnitude of M_c given in Equation (9) is the number of ways to partition data set **X** into c nonempty subsets and is quite large for all but trivial values of c and N. The discreteness of M_c endows it with certain analytical and algorithmic intractabilities. An exhaustive search of which hard partition best fits the data is impractical[11].

$$\left| M_c \right| = \frac{1}{c!} \left[\sum_{j=1}^{c} {c \atop j} (-1)^{c-j} j^{N} \right] \tag{9}$$

In the usual context, one has a data set **X** and hopes to infer, by some clustering method, structure in the represented physical population. The selection of the clustering criterion is a key step in any clustering algorithm. Specifically, what mathematical properties possessed by the members of the data should be used to identify clusters in **X**? No single criterion will be universally applicable and selection of a particular criterion is at least partially subjective and subject to question.

Objective function methods allow the most precise, but not necessarily the more valid, formulation of the clustering criterion. In particular, minimum-variance clustering is a popular choice for defining clusters in **X**. The within-group-sum-of-squared-error(WGSS) objective functional is a classical minimum-variance criterion which generates hard clusters in **X**. Toward this end let the objective function $J_1(\mathbf{U},V)$ be defined by:

$$J_1(\mathbf{U},V) = \sum_{k=1}^{N} \sum_{i=1}^{c} u_{ik}(d_{ik})^2 \tag{10a}$$

where

$$d_{ik} = d(\underline{x}_k,\underline{v}_i) = \| \underline{x}_k - \underline{v}_i \| =$$

$$\left[\sum_{j=1}^{p} (x_{kj} - v_{ij})^2 \right]^{\frac{1}{2}} \tag{10b}$$

$$V = \{\underline{v}_1, \underline{v}_2, \dots, \underline{v}_c\}, \quad \underline{v}_i \in \Re^p \text{ for every i} \tag{10c}$$

and

$$\mathbf{U} = [\, u_{ik} \,] \in M_c \text{ is hard.} \tag{10d}$$

The set V contains the c cluster centers, $\underline{v}_i$ is the cluster center for the hard cluster $u_i \in \mathbf{U}$, $1 \le i \le c$. In general, d_{ik}, is some measure of similarity between the k-th sample point, $\underline{x}_k$, and the i-th cluster center, $\underline{v}_i$. Typically d_{ik} must satisfy the following requirements:

$$d_{ik} \text{ is defined as } d(\underline{x}_k,\underline{x}_i) > 0 \tag{11a}$$

$$d_{ik} = 0 \text{ if and only if } \underline{x}_k = \underline{x}_i \tag{11b}$$

$$d_{ik} = d_{ki} \tag{11c}$$

Functions that satisfy (11) are called measures of dissimilarity because the larger the value for d_{ik} the less similar $\underline{x}_k$ is to $\underline{x}_i$. The dissimilarity function is often some measure of distance in $\Re^p$. Of particular interest for clustering purposes are inner product norms induced via matrix **A** in Equation (12).

$$d_{ik} = d(\underline{x}_k, \underline{v}_i) = (\underline{x}_k - \underline{v}_i)^T \mathbf{A} (\underline{x}_k - \underline{v}_i) \tag{12}$$

Recall that $\underline{x}_k$ and $\underline{v}_i \in \Re^p$ and **A** is any p x p positive-definite matrix. In the case of J_1 defined in Equation (10a) the measure of dissimilarity is the Euclidean norm metric and **A** is the p x p identity matrix. In general, the measure of dissimilarity need not be a metric, merely positive-definite and symmetric on $\Re^p$. The choice for **A** directly influences the shape of the clusters determined by a given clustering algorithm.

It is desired to find the pair $\mathbf{U}^*, V^*$ such that $J_1(\mathbf{U}^*, V^*)$ is minimum. Since an exhaustive search of M_c is impractical, the following Hard c-Means(HCM) algorithm is a popular way to approximate minima of J_1.

Hard c-Means Algorithm [11]

step 1: fix c, $2 \le c \le N$, and initialize $\mathbf{U}^{(0)} \in M_c$, then at iteration r: $r = 0,1,2, \ldots$

step 2: calculate the c mean vectors $\{\underline{v}_i^{(r)}\}$ using $\mathbf{U}^{(r)}$ and

$$\underline{v}_i = \frac{\sum_{k=1}^{n} u_{ik} \, \underline{x}_k}{\sum_{k=1}^{n} u_{ik}} \tag{13a}$$

step 3: update $\mathbf{U}^{(r)}$ for every i and k

$$u_{ik}^{(r+1)} = \begin{cases} 1 & , \; d_{ik}^{(r)} = \min\limits_{1 \le j \le c} \{d_{jk}^{(r)}\} \\ 0 & , \; \text{otherwise} \end{cases} \tag{13b}$$

where d_{ik} is the Euclidean norm between sample point, $\underline{x}_k$ and the i-th cluster center, $\underline{v}_i$

```
step 4:  compare $U^{(r)}$ to $U^{(r+1)}$ in a convenient matrix norm

            if $\| U^{(r)} - U^{(r+1)} \| \leq \varepsilon$   then,  stop
            else r = r + 1, go to step 2
            end if
```

Note the expression for $\underline{v}_i$ in step 2 is merely the sample mean for the sample points in the i-th cluster. This minimizes J_1 for the given hard partition. A new partition is then calculated in step 3 based on the new cluster centers. This process proceeds iteratively until the difference between successive partition matrices, using some convenient matrix norm, is less than some predefined tolerance, ε . Hard algorithms such as HCM have no general proof of convergence but yield acceptable results in certain data cases [11]. The success of a clustering algorithm depends upon its ability to identify meaningful substructure in data set **X**. Some of the mathematical difficulties of hard algorithms are overcome by allowing the elements of the partition matrix to be continuous variables rather than discrete variables. This leads to a fuzzy version of the HCM algorithm.

2.1.2 Fuzzy Algorithms and Cluster Validity

Let V_{cn} be the set of real c x n matrices: $2 \leq c \leq N$, then the fuzzy c-partition space, M_{fc} for **X** is defined below.

Definition 2 [11]

$$M_{fc} = \{\mathbf{U} \in V_{cn} \mid u_{ik} \in [0,1] \text{ for every i and k;}$$

$$\sum_{i=1}^{c} u_{ik} = 1 \text{ for every k; and}$$

$$0 < \sum_{k=1}^{N} u_{ik} < N \text{ for every i} \}$$

Note that each column sum of **U** is still one, but it is possible for each $\underline{x}_k$ to have a distributed membership among the c fuzzy partitions. One of the primary advantages of fuzziness is the differentiability of the u_{ik}'s over M_{fc}. This is not the case for J_1 over M_c. Differentiability of the u_{ik}'s often allows first-order necessary conditions to be found on the gradient of the fuzzy objective function. This provides the theoretical basis for approximation of local minima of the fuzzy objective function by the gradient method [11,47]. Consider the following fuzzy c-means algorithm(FCM).

Fuzzy c-Means Algorithm [11]

Let $$J_m(\mathbf{U},V) = \sum_{k=1}^{N} \sum_{i=1}^{c} (u_{ik})^m \, (d_{ik})^2 \qquad (14a)$$

be a fuzzy WGSS objective function where d_{ik} is some measure of dissimilarity as in (11) and m is a weighting exponent indicating the degree of fuzziness. For $m = 1$, J_m reduces to J_1 in Equation (10a), defined for the HCM algorithm. It is desired to minimize J_m.

```
step 1: fix c,  $2 \le c \le N$;
        choose a measure of dissimilarity, $d_{ik}$;
        fix m,  $1 \le m < \infty$;
        and initialize $\mathbf{U}^{(0)} \in M_{fc}$, then at iteration r:  $r = 1,2, \ldots$

step 2: calculate the c fuzzy cluster centers $\{\underline{v}_i^{(r)}\}$, using $\mathbf{U}^{(r)}$ and

        $\underline{v}_i = \dfrac{\sum_{k=1}^{n} (u_{ik})^m \, \underline{x}_k}{\sum_{k=1}^{n} (u_{ik})^m}$     for every i     (14b)

step 3: update the fuzzy partition, $\mathbf{U}^{(r)}$ ,

        if $d_{ik}(\underline{x}_k, \underline{v}_i) = 0$  then
            $u_{ik} = 1$;  $u_{jk} = 0$  for  $1 \le j \le c$, $j \ne i$
        else
            $u_{ik} = \left[ \sum_{j=1}^{c} \left( \dfrac{d_{ik}}{d_{jk}} \right)^{2/(m-1)} \right]^{-1}$     (14c)
        end if
```

step 4: use a matrix norm to check for convergence;

```
if $|| U^{(r)} - U^{(r+1)} || \le \varepsilon$    then,
        stop
else
    $r = r + 1$, go to step 2
end if
```

Practically, the FCM algorithm is useful since the Steps 1-4 outlined above are easily implemented. Given an initial partition matrix, $U^{(0)}$, the algorithm iteratively generates a sequence $\{(U^{(r)}, V^{(r)})\}$ by selecting $V^{(r+1)}$ to satisfy Equation (14b) using $U^{(r)}$ and by selecting $U^{(r+1)}$ to satisfy Equation (14c) using $V^{(r+1)}$. The expression for u_{ik} in Step 3 is derived by fixing the cluster centers $\{\underline{v}_i\}$ and applying Lagrangian multipliers to the variables, $\{u_{ik}\}$, of the objective function J_m in Equation (14a) [11]. Theoretically, it has been shown that $\{(U^{(r)}, V^{(r)})\}$ converges, at least along a subsequence to a pair (U^*, V^*) that satisfies necessary but not sufficient derivative conditions for minimizing J_m in Equation (14a) [12]. The original theory concerning descent and convergence properties of FCM stated that (U^*, V^*) is a local minimizer of J_m, but both Sabin [12] and Tucker [12] have produced counterexamples illustrating that (U^*, V^*) may indeed be a saddle point of J_m. This modified convergence theory brings into question the underlying mathematical rationale for using FCM for exploratory data analysis and classifier design. The original theory accepted U^* as a reasonable explanation of the substructure of **X** based on the belief that the clustering criterion J_m would be minimized locally when data points in **X** pack tightly around their prototypes $\{\underline{v}_i^*\}$. The fact that the modified convergence theory can not guarantee that $J_m(U^*, V^*)$ is a local minimum has had little impact on the practical significance of the FCM algorithm in applied research. Over the past decade there have been applications of FCM in agriculture, engineering, astronomy, chemistry, geology, image analysis, medical diagnosis, shape analysis and target recognition [10]. FCM and associated algorithms have been used for vector quantization which is an important aspect of data compression or coding [37,48]. The computational experience of these applications has shown that when FCM is applied to real data it almost always terminates and the data is partitioned into relatively high density clusters of points. Even if the convergence theory could guarantee that the terminal value of the sequence generated by the FCM algorithm was an estimate of a local minimizer of J_m, this does not assure that the resulting partition is 'meaningful' when related back to the physical process generating the data. The fact that FCM behaves well in practical applications makes it useful, but there is still the need to evaluate clustering results in light of any additional information the investigator may have about the data.

An especially important application of the FCM algorithm is the design of a nearest prototype classifier for an unknown distribution function using a training set of observations. This application will be explored in Chapter 4. The prototypes for the classifier are generated by running the FCM algorithm on the observations in the training set. Hopefully, these prototypes can be used to classify observations effectively outside the training set. It is important to note that the FCM algorithm makes no explicit assumptions concerning the distribution of the data represented by the training set of observations. The algorithm is driven solely by the observations in the training set. The intent is to find existing structure in the data rather than impose structure on the data by making unwarranted assumptions concerning the distribution of the data being investigated.

Once the data set **X** has been partitioned using an algorithm such as FCM, then it is desirable to have an objective way to evaluate the resulting fuzzy partition **U**. One such measure of cluster validity is given by:

$$F(\mathbf{U},c) = \frac{tr(\mathbf{U}\,\mathbf{U}^t)}{N} \tag{15}$$

F is called the partition coefficient of **U** and provides a scalar measure of the amount of unshared membership of **X** in the c fuzzy partitions designated by **U** [9,11]. F is bounded by $1/c \le F \le 1$, and F maximizes as unshared membership increases, with F equal one, if and only if **U** is a hard partition of **X**. If **X** really has distinct substructure, then fuzzy partitioning algorithms should produce relatively hard partitions as measured by F. Since F maximizes to one for every hard partition it can only be used to evaluate fuzzy partitions.

A second measure of cluster validity is made with respect to a hard c-partition of **X**. The following definition allows a simple way to determine the nearest maximum membership(MM) hard partition for a given fuzzy partition.

Definition 3 [11]

If **U** is a fuzzy partition of **X**, then the nearest hard c-partition of **X** in the sense of maximum membership is the partition $\mathbf{U}_{MM}$, whose ik-th element is

$$u_{ik} = \begin{cases} 1, & u_{ik} = \max\limits_{1 \leq j \leq c} \{u_{jk}\} \\ 0, & \text{otherwise} \end{cases}$$

Using Definition 3, the separation coefficient of $\mathbf{U}_{MM}$ is the scalar:

$$G(\ \mathbf{U},V,c,\ \mathbf{X},d) = 1 - \max_{i+1 \leq j \leq c} \{ \max_{1 \leq i \leq c-1} \{(r_i + r_j)/d_{ij}\}\} \qquad (16)$$

where r_i represents the radius of the smallest closed ball centered at $\underline{v}_i$, that contains hard cluster u_i and d_{ij} denotes the distance of separation between cluster centers $\underline{v}_j$ and $\underline{v}_i$ [11]. Notice that G depends not only on **U** and c, as did F, but also on the cluster centers, $V = \{\underline{v}_i\}$, $1 \leq i \leq c$, the data **X** and the measure of distance, d. G has the following properties:

1. $0 < G < 1$ if and only if no pair of closed balls intersect one another;
2. $G = 0$ if and only if the closest pair of closed balls are exactly tangent;
3. $G < 0$ if and only if at least one pair of closed balls intersect one another.

G depends on the compactness(r_i) and separation(d_{ij}) of the closed balls containing the respective hard clusters. Larger values of G are indicative of better substructure in **X**. G can not be used directly for fuzzy cluster validity but via Definition 3 an indirect evaluation can be made for fuzzy clusters.

A third measure of cluster validity is based on the difference in magnitude between the hard objective function, J_1, and the fuzzy objective function, J_m, given in Equations (10a) and (14a) respectively. The objective function coefficient is defined by:

$$\Delta J = |\ J_m - J_1\ | \qquad (17)$$

ΔJ is a relative measure of how close the fuzzy partition of **X** is to the corresponding maximum membership hard partition of **X** in terms of within-group-sum-of-square-error. Recall that the underlying rationale of fuzzy clustering is that fuzzy clustering works best when the resulting clusters are reasonably "hard", and this condition is reflected by relatively small values of ΔJ.

The purpose of any validity measure is to point to the value of c which best fits the data for a given measure of dissimilarity. Fuzzy clustering and the use of validity measures F, G and ΔJ are better explained via an example using a well known data set.

2.1.3 An Example Using Anderson's Iris Data

Anderson's Iris data was chosen to illustrate a modified form of the FCM algorithm and validity measures F, G and ΔJ because it is a well defined data set that provides a point of reference to other classification schemes. The Iris data has been used as a test set by numerous authors, including Backer [4], Bezdek [9,11], Duda and Hart [20], Fischer [22], Friedman and Rubin [23], Kendall [36], Scott and Symons [50] and Wolfe [62]. The Iris data consists of 150 four-dimensional vectors, each of which gives the sepal length, sepal width, petal length and petal width, all measured in centimeters [22]. The 150 samples represent 50 samples each from three subspecies of irises(i.e., setosa, veriscolor and virginica). Figure 5 shows the Iris data plotted using the petal features which are the two most discriminating features. Bezdek [9] used the Iris data to contrast the performance of the FCM algorithm with the performance of several other clustering methods [23,36,50,62]. If it is assumed *a priori* that the Iris data consists

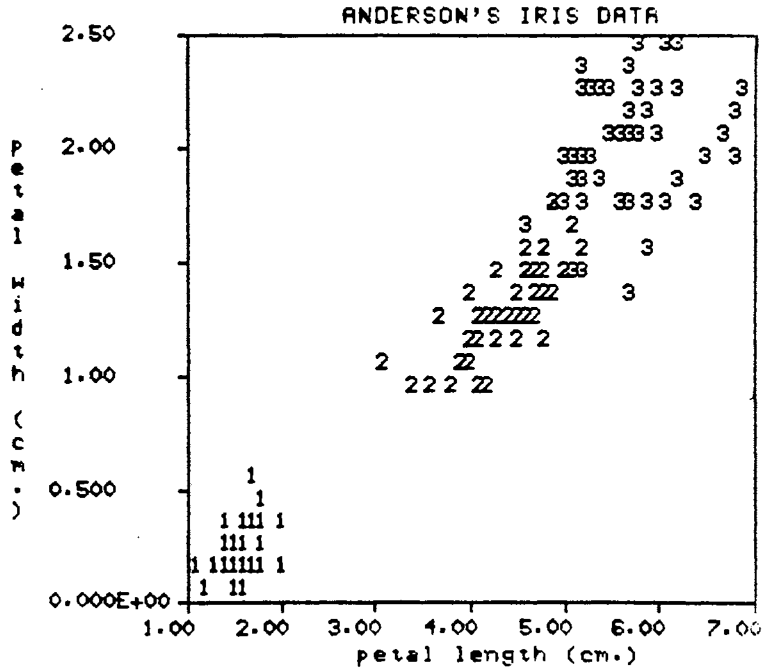


Figure 5. Anderson's Iris Data plotted using Petal Features
(1 = setosa, 2 = veriscolor and 3 = virginica)

of three subpopulations, then the FCM algorithm was outperformed by the methods of Wolfe [62] and Friedman and Rubin [23] in terms of each method's ability to classify the Iris data correctly . The primary advantage of the FCM algorithm, demonstrated by Bezdek [9], is its ability to decompose the Iris data without the *a priori* knowledge of the number of clusters being sought. This advantage, plus the fact that the FCM algorithm makes no explicit assumptions about the distribution of the data, make the FCM algorithm an excellent tool for exploratory data analysis.

The initial example uses the original Iris data; no scaling or transforming of the data has been done. There are several algorithmic parameters associated with the FCM algorithm given in (14), namely: c, m, $\mathbf{U}^{(0)}$, d and ε. For this example, c = 2,3,4,5,6 and d is the Euclidean metric that is implemented by letting **A** in Equation (12) be the 4 x 4 identity matrix. The matrix norm in Step 4 of the FCM algorithm is a simple element-by-element comparison of the two matrices; $\mathbf{U}^{(r)}$ and $\mathbf{U}^{(r+1)}$, and the maximum difference between corresponding elements is compared to $\varepsilon = 0.01$. A modified version of the FCM algorithm was applied to the Iris data with $V^{(0)}$ being specified instead of $\mathbf{U}^{(0)}$ and the iterative loop beginning at Step 3 rather than Step 2. The initial cluster centers, $V^{(0)} = \{\underline{v}_i^{(0)}\}$, are selected by using c equally spaced points along each of the coordinate axes. Finally, the weighting exponent m must be specified. There are heuristic guidelines for choosing m, but there exists no theoretical basis for the optimal choice of m [11]. This initial example uses m = 1.25, 1.5 and 2.0 to show how m impacts the cluster validity measures that are used to evaluate the clustering results.

A few observations about the weighting exponent before proceeding with the example. In general, as m becomes larger, the 'fuzzier' are the cluster membership assignments, and as m approaches 1 from the right the FCM solutions become hard. Theoretically, as m approaches infinity the u_{ik} in Equation (14c) approach 1/c and all the fuzzy cluster centers approach the centroid of the data. Again, according to theory, as m approaches one from the right FCM converges to a 'general' hard solution [11]. Thus, m controls the extent of membership sharing between fuzzy clusters in the data set being investigated. One can artificially influence the FCM solution by choosing extreme values for m. It is desirable to choose m large enough so that if the resulting FCM solution yields relatively 'hard' clusters then this is a good indication of substructure in the data. However, choosing m too large essentially eliminates the possibility that the resulting clusters will have good structure, as measured by the cluster validity measures.

Table I shows the partition coefficient, F(**U**,c), the separation coefficient, G(**U**,V,c,**X**,d), and the objective function coefficient, ΔJ for the original Iris data, using the Euclidean norm and $\varepsilon = 0.01$. Maximum F and G are attained when c = 2 for all three values of m. Recall that larger F values are indicative of better substructure. Also, recall that G is a relative indicator of separation for the maximum membership hard clusters derived using Definition 3. G indicates the worst case separation for the pair of hard clusters with the least spatial separation. In contrast to F and G, ΔJ indicates that five is a good choice for c for all three values of m.

Based on the validity measures, one could reason that two is the best value of c with five being the next best value. Notice how the various validity indicators in Table I are affected by the different values for m and c. Increasing m increases the amount of shared membership in the fuzzy clusters and this is reflected by lower values of the partition coefficient F. F also exhibits a tendency to decrease as c increases. The only exception to F's decreasing tendency occurs when m = 1.25 and c increases from 4 to 5. The decreasing tendency in F, as c increases, is explained in part by the changing lower bound for F (i.e., $1/c \leq F \leq 1$) which results in a larger possible range for F. Also, as c increases there are more opportunities for an individual sample point to have its membership divided between two or more clusters, thus lowering F. The fact that all the F values are relatively high for m = 1.25 is an indication that m may be too small, which forces the resulting clusters to be fairly 'hard' clusters. The fact that F increases when c increases from 4 to 5 in the case of m = 1.25, and F decreases only sightly for the same situation in the case of m = 1.5, is another indication that five is a reasonable, but perhaps not the best, choice for c. In this example, the overlap of the maximum membership hard clusters, as measured by G, is slightly greater between clusters as m increases. This is evidenced by comparing the values of G in Table I(c) with the values of G in Tables I(a) and I(b). It is worth noting that for c = 2, G is the same for all three values of m and for c = 5, G is the same for m = 1.25 and m = 1.5 and slightly worse for m = 2.0. The objective function coefficient yields larger values as m increases, but for all three cases shown in Table I the minimum ΔJ is achieved when c = 5.

A value of 1.5 is viewed as a good nominal value for m and Figure 6 shows the FCM maximum membership clusters for m = 1.5 when c = 2 and c = 5. It is somewhat disappointing that the FCM clusters in Figure 6(a) do not correspond exactly to the two visually obvious clusters. A comparison of Figure 5 with Figure 6(a) shows that the maximum membership cluster #1 corresponds to the veriscolor and virginica species

TABLE I.

FCM VALIDITY MEASURES FOR THE ORIGINAL IRIS DATA:
USING THE EUCLIDEAN NORM,∈ = 0.01 AND
m = 1.25, m = 1.5 AND m = 2.0

c	F	G	ΔJ
2	0.987	–0.183	1.716
3	0.971	–0.775	1.002
4	0.954	–0.642	0.783
5	0.960	–0.747	0.569
6	0.942	–1.234	0.655

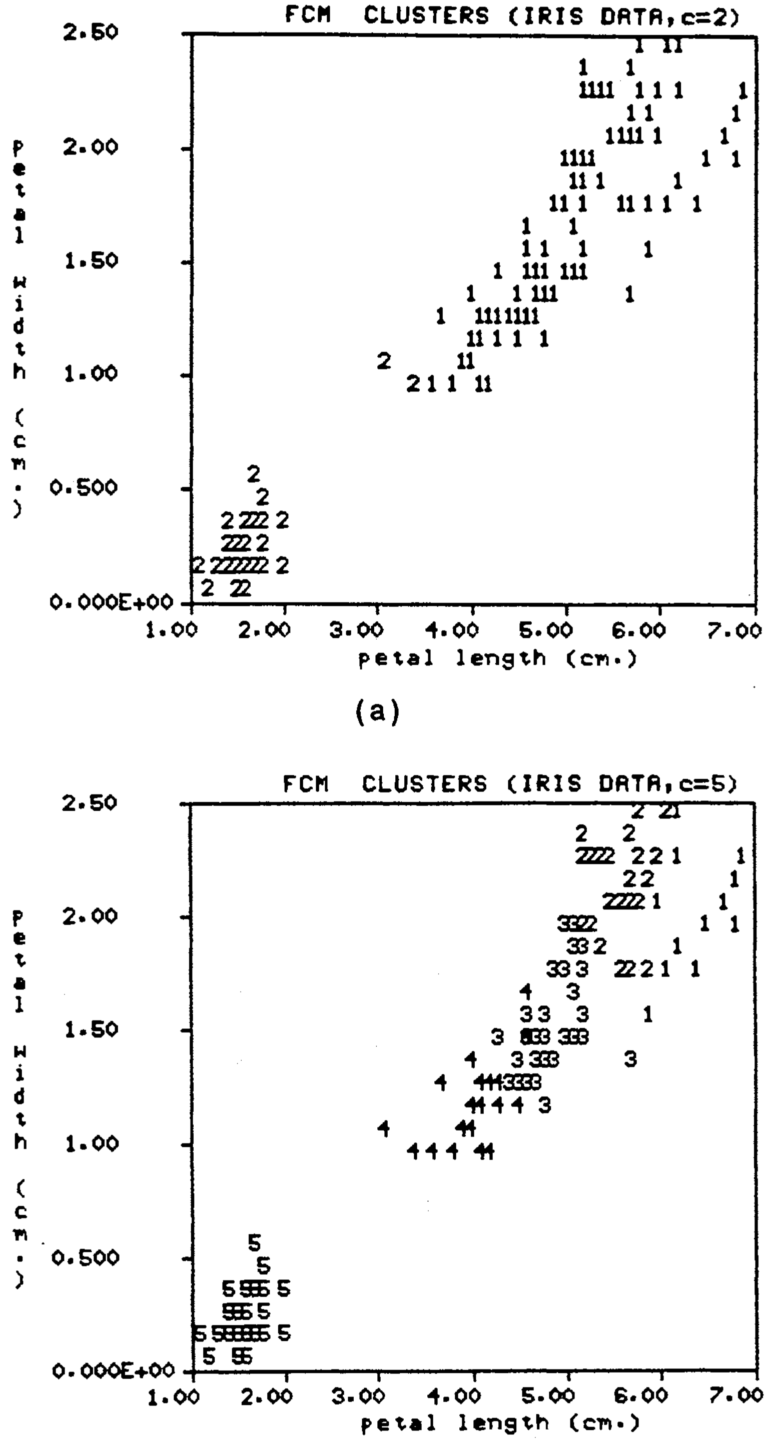
(a) m = 1.25

c	F	G	ΔJ
2	0.968	–0.183	6.049
3	0.919	–0.788	4.473
4	0.888	–0.602	3.543
5	0.881	–0.747	2.936
6	0.638	–1.249	3.396

(b) m = 1.5

c	F	G	ΔJ
2	0.892	–0.183	23.452
3	0.783	–0.804	18.518
4	0.707	–0.652	15.849
5	0.664	–0.795	13.906
6	0.594	–1.249	14.347

(c) m = 2.0



(a)

(b)

Figure 6. FCM Clusters Using Original Iris Data
(a) c = 2 and (b) c=5

and cluster #2 corresponds to the setosa species, with the exception of three veriscolor samples that are grouped in cluster #2. A similar comparison of Figure 5 with Figure 6(b) shows that FCM clusters #1 and #2 correspond to virginica, clusters #3 and #4 correspond to veriscolor and cluster #5 corresponds to setosa. This correspondence between FCM clusters in Figure 6(b) and the Iris data in Figure 5 results in 14 of the virginica samples being grouped incorrectly with the veriscolor samples.

This simple introductory example illustrates how cluster validity measures F, G and ΔJ can be used to choose an appropriate value of c, based upon the FCM clustering results. Each validity measure takes into consideration a different aspect of the FCM clustering results. F directly evaluates unshared membership of the fuzzy partition, G measures spatial separation of hard clusters derived from the fuzzy partition and ΔJ measures how close the maximum membership hard partition and the fuzzy partition are in terms of within-group-sum-of-square-error. The above example illustrates that these measures of cluster validity need not agree upon the best value of c, but each validity measure has its own merit. Also, the best value or values of c as determined by the validity measures do not necessarily agree with another grouping of the data based upon an examination of the physical process generating the data. In the above example, the biologist chooses three to be the best number of subpopulations for the Iris data, whereas interpretation of the validity measures points to either two or five subpopulations. It is always prudent to evaluate clustering results in light of any additional information known about the physical process represented by the data. One difficulty associated with the FCM algorithm is choosing an appropriate value for the weighting exponent m. The weighting exponent directly effects the values and interpretation of the cluster validity measures. For the above example, $m = 1.5$ was chosen empirically as a good nominal value for m, but this is certainly subject to question.

2.2 Principal Components and Data Scaling

The development of the method of principal components is usually credited to Hotelling [35]. The method of principal components transforms discrete variables that are correlated into a set of uncorrelated principal components; this is done by accounting for the variance of the original variables. The analogous transformation for continuous data was discovered by Karhunen and Loeve and is called the Karhunen-Loeve Transformation [26,58]. Principal component analysis and factor analysis are related, yet different, statistical procedures. As mentioned, principal component analysis transforms the original variables into new variables by accounting for the variance of the original variables, whereas factor analysis transforms the original variables into new variables by accounting for the correlation among the original variables. The terms "components" and "factors" are sometimes used interchangeably in the literature, when in fact there is a distinction between them. Typically, the use of principal components is motivated by the need to reduce the dimensionality of the data in a clustering problem and still retain the necessary discriminatory information to distinguish between different classes of data [20]. Such is not the case for the clustering of wireline log data. A standard suite of logs may consist of seven to ten wireline logs and present day computational systems can easily handle problems of this dimensionality. The use of principal components is motivated on two counts. First, the use of principal components gives insight into the structure of the wireline log data by representing the data in terms of uncorrelated principal component features that are linear combinations of the original data features. Second, if the clustering is being done in a high dimension space, then principal components provide a better means for the visual display of clustering results in a lower dimension space.

Miesch [40] reviews a variety of principal component methods which have become popular in the geological sciences. These methods range from the R-mode method, based on the correlation or covariance matrix, to Q-mode methods, which are based on coefficients that express the relations among observations rather than variables. Miesch [40] points out that the dominant feature distinguishing one method of principal components from another is the manner in which the original data are scaled prior to the other computations. A second distinction between principal component methods is whether the eigenvectors of the inner product-moment of the scaled data matrix are taken directly as the Q-mode scores or normalized and called the R-mode loadings. Most often the inner product-moment is a correlation, covariance or a scaled psuedo cos θ matrix; however, the inner product-moment need not be one of these to form the basis for a valid principal component solution [40]. Section 2.2.1 outlines how principal component logs are generated using the discrete Karhunen-Loeve Transform(KLT), which is an R-mode method based on the covariance matrix of the scaled wireline logs. Section 2.2.2 explains why data scaling is sometimes necessary and illustrates the impact which different types of scaling can have on the resulting principal components. For continuity of presentation the effects of scaling are illustrated using the Iris data shown in Figure 5.

2.2.1 The Discrete Karhunen-Loeve Transformation(KLT)

The theory and properties of the KLT are discussed in the literature [16,20,21, 26,45,58]. This section gives the development of the discrete KLT as applied to a set of wireline logs. Let data set **X** be an M x N matrix of the original digitized wireline logs.

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1N} \\ x_{21} & x_{22} & \cdots & x_{2N} \\ \vdots & \vdots & & \vdots \\ x_{M1} & x_{M2} & \cdots & x_{MN} \end{bmatrix} \tag{18}$$

Each row of **X** corresponds to a different wireline log. Data matrix **Z** is formed by a linear scaling of the rows of **X**. The linear scaling considered in this discussion will have the form:

$$z_{ij} = (x_{ij} - a_i)/b_i \tag{19}$$

where z_{ij} and x_{ij} are elements of the i-th row and j-th column of **Z** and **X** respectively and a_i and b_i are scaling constants for the i-th row of **X**. Now an M element column vector, $\underline{z}$, may be formed by considering any column of **Z**. The i-th column of **Z** is given by:

$$\underline{z}_i = [z_{1i}, z_{2i}, \dots , z_{ji}, \dots , z_{Mi}]^T \tag{20}$$

with the T indicating transposition. The elements of a $\underline{z}$ vector represent M scaled log readings which correspond to a particular depth in the borehole. The covariance matrix of the $\underline{z}$ vectors is defined as:

$$C_z = E\,[(\underline{z} - \underline{m}_z)(\underline{z} - \underline{m}_z)^T] \tag{21}$$

where

$$\underline{m}_z = E\,[\,\underline{z}\,] \tag{22}$$

is the mean vector and E is the expectation operator. Equations (21) and (22) may be approximated from the scaled log measurements by replacing the expectation operator with the sample average.

$$\underline{m}_z = \frac{1}{N} \sum_{i=1}^{N} \underline{z}_i \tag{23}$$

$$C_z = \frac{1}{N} \sum_{i=1}^{N} \underline{z}_i \underline{z}_i^T - \underline{m}_z \underline{m}_z^T \tag{24}$$

The mean vector, $\underline{m}_z$, is of dimensionality M and $\mathbf{C_z}$ is an M x M matrix. Since $\mathbf{C_z}$ is a real symmetric positive semidefinite matrix, it is always possible to find a set of orthonormal eigenvectors [26]. Let $\underline{e}_i$ and λ_i, $i = 1,2, \dots, M$, be the normalized eigenvectors and eigenvalues respectively of $\mathbf{C_z}$. For convenience, assume the eigenvalues are arranged in decreasing order such that $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_M$. The transformation matrix for the discrete KLT is then formed using the normalized eigenvectors of $\mathbf{C_z}$ to form the rows of the transformation matrix, **B**. The discrete KLT

$$\mathbf{B} = \begin{bmatrix} e_{11} & e_{12} & \cdots & e_{1N} \\ e_{21} & e_{22} & \cdots & e_{2N} \\ \vdots & \vdots & & \vdots \\ e_{M1} & e_{M2} & \cdots & e_{MN} \end{bmatrix} \tag{23}$$

then consists of multiplying a centralized vector $(\underline{z} - \underline{m}_z)$ by **B** to obtain a new vector $\underline{Y}$.

$$\underline{Y} = \mathbf{B}(\underline{z} - \underline{m}_z) \tag{24}$$

It has been shown that the covariance matrix of the $\underline{Y}$ vectors is a diagonal matrix with elements equal to the eigenvalues of $\mathbf{C_z}$ [26]. This implies, the elements of $\underline{Y}$ are uncorrelated and each eigenvalue, λ_i, is equal to the variance of the i-th element of $\underline{Y}$

$$\mathbf{C}_Y = \begin{bmatrix} \lambda_1 & 0 & \cdot & \cdot & \cdot & 0 \\ 0 & \lambda_2 & & & & 0 \\ \cdot & & \cdot & & & \cdot \\ \cdot & & & \cdot & & \cdot \\ \cdot & & & & \cdot & 0 \\ 0 & 0 & \cdot & \cdot & 0 & \lambda_M \end{bmatrix} \qquad (25)$$

along eigenvector, $\underline{e}_i$. In the present application an M-dimensional vector, $\underline{z}$, consisting of M scaled log measurements is centralized about the mean vector, $\underline{m}_Z$, which has been calculated over some interval of interest. The resulting vector is multiplied by the KLT matrix, **B**, and yields a new vector $\underline{Y}$. In matrix form, let **Z'** equal the scaled data matrix, **Z**, with the mean vector, $\underline{m}_Z$, subtracted from each column, then the KLT can be implemented by the matrix multiplication in Equation (26). The rows of matrix **Y** represent the uncorrelated principal component(PC) logs. Since the eigenvalues are in decreasing order the PC logs are ordered such that PC log #1(i.e., row #1 of matrix **Y**)

$$\mathbf{Y} = \mathbf{BZ'} = \begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1N} \\ y_{21} & y_{22} & \cdots & y_{2N} \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ y_{M1} & y_{M2} & \cdots & y_{MN} \end{bmatrix}. \qquad (26)$$

has the largest variance and subsequent PC logs have nonincreasing variances, with the last PC log having the least variance. The KLT compresses the essential variability of the log data into relatively few signals, which is nice for the visual display of clustering results in a lower dimensional space and allows the option of reducing the dimensionality of the clustering problem. Geometrically, the KLT consists of a translation and rotation of the scaled wireline log data such that the first PC axis lies along the line of maximum variance of the scaled data. The second PC axis is orthogonal to the first PC axis and in the direction of the next greatest variation of the scaled data and so on for the remaining

PC axes. The KLT is a fixed body transformation of the data and has no effect on the clustering performance of an algorithm such as FCM. The use of the KLT and the effects of scaling are illustrated with another example using Anderson's Iris data.

2.2.2 A Second Example Using Anderson's Iris Data

Of the unlimited number of ways to derive principal components, only three methods are contrasted in this example. For all three methods the raw data matrix, **X**, is a 4 x150 matrix representing the original Iris data. The rows of **X** contain the 150 measurements for sepal length, sepal width, petal length and petal width respectively. The only difference in the three methods is the manner in which the raw data matrix, **X**, is scaled to obtain the data matrix **Z**. Method 1 uses $a_i = 0$ and $b_i = 1$ in Equation (19) for scaling the rows of **X**. In method 1, **Z** equals **X**. Method 2 uses a_i equal to the row mean and b_i equal to the row standard deviation while method 3 uses a_i equal to the row mean and b_i equal to the magnitude of the largest excursion from the row mean. Table II shows the means and standard deviations of the Iris data after scaling by Equation (19) using the scaling constants for the three methods listed above. Method 1 yields the means and standard deviations for the original Iris data. Method 2 scales the original data to give zero mean and unit standard deviation signals while method 3 produces zero mean signals whose magnitude is bounded by plus and minus one. The discrete KLT transforms the scaled Iris data into principal components according to Equation (27). The principal components are linear combinations of the scaled Iris features which account for the variation in the scaled data. Table III lists the KLT matrix and corresponding eigenvalues for the three methods of scaling. Notice that without scaling, (method 1), the variables are effectively weighted according to their standard deviations without regard for the relative magnitude of the variables. A comparison of the first row of the KLT matrix for method 1 in Table III and the standard deviations for method 1 in Table II

$$\overset{4x150}{\begin{bmatrix} \text{principal component feature \#1} \\ \text{principal component feature \#2} \\ \text{principal component feature \#3} \\ \text{principal component feature \#4} \end{bmatrix}} = \overset{4x4}{\begin{bmatrix} \text{KLT} \\ \text{matrix} \end{bmatrix}} \overset{4,150}{\begin{bmatrix} \text{sepal length} \\ \text{sepal width} \\ \text{petal length} \\ \text{petal width} \end{bmatrix}} \qquad (27)$$

TABLE II.

MEANS AND STANDARD DEVIATIONS OF THE IRIS DATA
FOR THE THREE METHODS OF SECTION 2.2.2

Method	Means				Standard Deviations			
	sepal length	sepal width	petal length	petal width	sepal length	sepal width	petal length	petal width
1	5.843	3.057	3.758	1.199	0.825	0.434	1.759	0.760
2	0.000	0.000	0.000	0.000	1.000	1.000	1.000	1.000
3	0.000	0.000	0.000	0.000	0.401	0.324	0.560	0.584

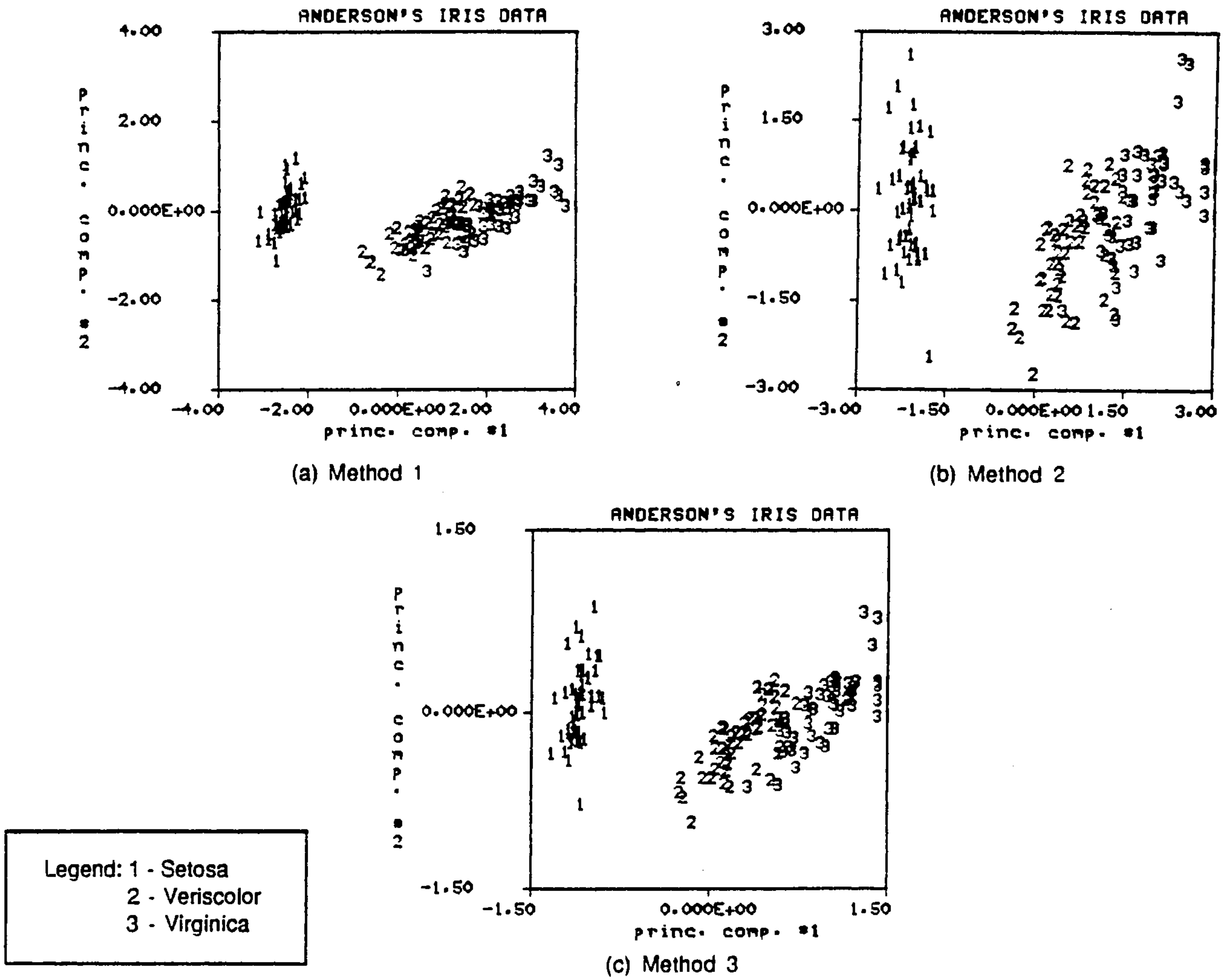
TABLE III.

EIGENVALUES AND KLT MATRICES FOR THE
THREE METHODS OF SECTION 2.2.2

Method	Eigenvalues	KLT Matrix			
1	4.200	0.361	-0.085	0.857	0.358
	0.241	0.657	0.730	-0.173	-0.075
	0.077	0.582	-0.598	-0.076	-0.546
	0.024	0.315	-0.320	-0.480	0.754
2	2.918	0.521	-0.269	0.580	0.565
	0.914	0.377	0.923	0.024	0.067
	0.147	0.720	-0.244	-0.142	-0.634
	0.021	0.261	-0.124	-0.801	0.524
3	0.782	0.403	-0.148	0.629	0.648
	0.102	0.417	0.907	-0.056	0.002
	0.031	0.716	-0.322	0.096	-0.611
	0.006	0.389	-0.227	-0.769	0.454

reveal that the Iris features with the larger standard deviations are more closely accounted for in the determination of principal component features. Specifically, the first principal component is dominated by the petal length. The purpose of the scaling used in methods 2 and 3 is to weight the respective Iris features more evenly by taking into account their standard deviations in method 2 and their largest excursion from the mean in method 3. The effects of the scaling procedures are reflected by the transformation matrices for methods 2 and 3 listed in Table III. The petal length no longer dominates the determination of the first principal component for method 2 or method 3. The first principal component for both methods 2 and 3 is most dependent on the petal features with method 3 giving slightly more weight to the petal features than method 2. This emphasis on the petal features is desirable since it is well known that the petal features are the most discriminating of the original Iris features. Typically, the eigenvalues of Table III are used as a measure of the amount of variance accounted for by a particular principal component. The sum of the k largest eigenvalues divided by the sum of all the eigenvalues gives an eigenvalue ratio that is commonly interpreted as the proportion of total variance in the original data that can be accounted for by the first k principal components. Such an interpretation is correct only when the scaling procedure of method 2 is used [40]. When the scaling is done in any other manner the eigenvalue ratio can only be used to determine the degree to which the first k principal components account for the variance of the scaled data, not the original data. In fact, regardless of the type of principal component analysis that is performed, the eigenvalues of method 2 give a better measure of the manner in which the principal components account for the variance of the original data [40]. This does not imply that the principal components derived using scaling procedures other than method 2 are less valid, just that one should not misinterpret the associated eigenvalues. For example, the eigenvalues of method 2 in Table III indicate that the first principal component accounts for approximately 73% of the variance in the original data. This percentage would also accurately reflect the proportion of total variance in the original data accounted for by the first principal components of methods 1 and 3 respectively.

Figure 7 shows a comparison of the Iris data plotted using the first two principal component features for the three different scaling methods. The separation coefficients, which measure the compactness and separation of the different Iris species, are given in Table IV for each scaling method. The calculation of the separation coefficients is presented as part of the discussion regarding cluster validity at the end of Section 2.1.2. In the present context, the separation coefficients give an indication of the effects of scaling on the compactness and separation of the labelled Iris data. A comparison of the



(a) Method 1

(b) Method 2

(c) Method 3

Figure 7. Anderson's Iris Data plotted in 2-dimensional Principal Component Space for the Three Methods of Section 2.2.2

TABLE IV.

SEPARATION COEFFICIENTS FOR THE LABELLED IRIS DATA OF FIGURE 7

cluster #	1	2	3
1	0.000	0.127	0.302
2		0.000	−1.236
3			0.000

(a) Method 1

cluster #	1	2	3
1	0.000	−0.707	−0.313
2		0.000	−2.156
3			0.000

(b) Method 2

cluster #	1	2	3
1	0.000	−0.216	0.098
2		0.000	−1.355
3			0.000

(c) Method 3

separation coefficients in Table IV indicates the best structure exists for method 1, the next best structure for method 3 and the poorest structure exists for method 2. One ill effect of the scaling procedures used in methods 2 and 3 is a loss of compactness of the labelled Iris subgroups, with method 2 being considerably worse than method 3 in this respect. The more important question is, how do the scaling procedures effect the ability of the FCM algorithm to detect the respective Iris subgroups?

Table V shows the FCM validity measures for the scaled Iris data shown in Figure 7. Notice the validity measures for method 1 are exactly the same as those recorded in Table 1(b). This is expected since the KLT is a fixed body transformation consisting of a rotation and translation of the original Iris data and does not effect the clustering performance of the FCM algorithm. The validity measures F and G of method 1 point to two as a good choice for c while ΔJ points to five as a good choice for c. Figure 6 shows the FCM clusters for c equals two and five in the domain of the original Iris data. For method 2, all three validity indicators point to two as the best choice for c. Figure 8(a) shows the FCM maximum membership clusters for method 2 and c = 2. Notice that the clusters in Figure 8(a) correspond exactly to the two visually obvious clusters with cluster #1 corresponding to the veriscolor and virginica species and cluster #2 corresponding the setosa species. This is an improvement over the performance of method 1 when c = 2. The validity measures for method 3 indicate either two or three as appropriate choices for c. For c = 2, method 3 clusters the Iris data in precisely the same manner as method 2 displayed in Figure 8(a). Method 3 is the only method that indicates that three is a reasonable choice for c and the resulting FCM maximum membership clusters are shown in Figure 8(b). The original setosa samples correspond exactly to cluster #3 in Figure 8(b), veriscolor corresponds to cluster #2 and virginica corresponds to cluster #1 with a total of 17 misclassifications in the latter two groupings. By comparison, when c = 3, method 1 yields 17 misclassifications and method 2 yields 25 misclassifications of the labelled Iris data. Method 3 is at least as good as the other methods in grouping the Iris data for c = 3 and has the advantage of having a validity measure, ΔJ, which indicates that three is a reasonable choice for c. Table VI compares the fuzzy and hard prototypes generated by the FCM algorithm for method 3 and c = 3 with the sample means of the original labelled Iris data. The fuzzy prototypes are derived via Equation (14b) and are the fuzzy cluster centers. The hard prototypes are the cluster centers for the maximum membership hard clusters derived from the fuzzy partition of the Iris data using Definition 3. Even though there are a significant number of misclassifications of the Iris data, the FCM algorithm does

TABLE V.

FCM VALIDITY MEASURES FOR SCALED IRIS DATA:
USING METHODS 1,2 AND 3 OF SECTION
2.2.2 WITH EUCLIDEAN NORM,
$\in = 0.01$ AND M = 1.5

c	F	G	ΔJ
2	0.968	–0.183	6.049
3	0.919	–0.788	4.473
4	0.888	–0.602	3.543
5	0.881	–0.747	2.936
6	0.838	–1.249	3.396

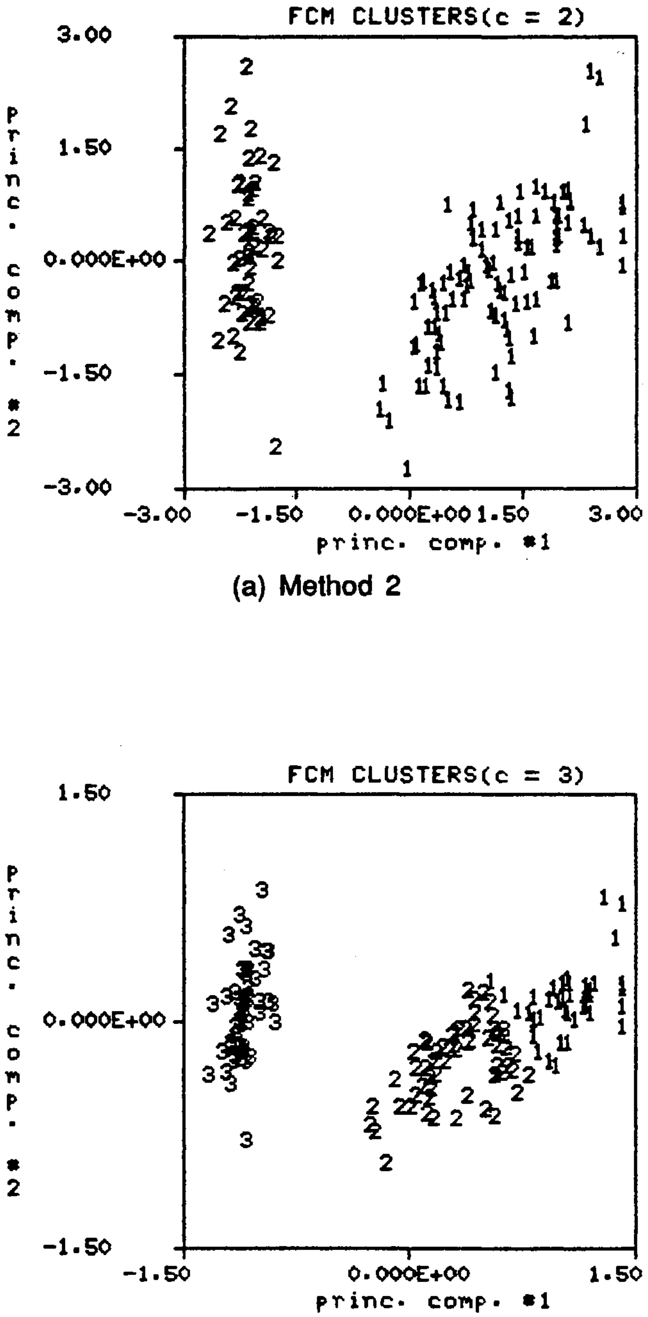
(a) Method 1

c	F	G	ΔJ
2	0.953	–0.705	8.803
3	0.888	–1.277	9.584
4	0.837	–1.370	9.705
5	0.802	–1.548	9.785
6	0.783	–1.641	11.181

(b) Method 2

c	F	G	ΔJ
2	0.960	–0.345	1.742
3	0.899	–0.889	1.290
4	0.866	–1.117	1.317
5	0.825	–1.493	1.392
6	0.807	–1.493	1.441

(c) Method 3



(a) Method 2

(b) Method 3

Figure 8. FCM Clusters for Scaling Methods 2 and 3

TABLE VI.

COMPARISON OF FCM FUZZY AND HARD PROTOTYPES WITH THE SAMPLE MEANS OF THE LABELLED IRIS DATA

Iris Species	Feature	Fuzzy Prototypes	Hard Prototypes	Sample Means
Setosa	sepal length	5.01	5.01	5.01
	sepal width	3.43	3.43	3.43
	petal length	1.47	1.46	1.46
	petal width	0.248	0.246	0.246
Veriscolor	sepal length	5.87	5.89	5.94
	sepal width	2.72	2.74	2.77
	petal length	4.36	4.40	4.26
	petal width	1.38	1.42	1.33
Virginica	sepal length	6.79	6.85	6.59
	sepal width	3.07	3.08	2.97
	petal length	5.65	5.70	5.55
	petal width	2.07	2.08	2.03

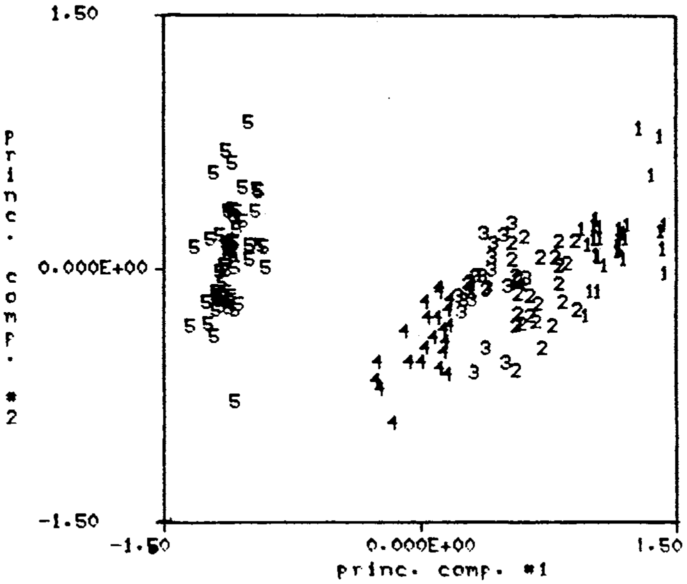


Figure 9. FCM Clusters for c=5 and Covariance Norm

generate a set of prototypes which are good representatives of the respective labelled Iris groupings with the fuzzy protoypes being slightly closer to the sample means than the hard prototypes. This example illustrates how a given scaling procedure impacts not only the determination of principal components but also impacts the FCM clustering performance. The intended purpose of scaling is to weight the respective data features more evenly so that all features of the original data with significant variation are accounted for in the derivation of principal component features. The KLT determines the principal components by forming linear combinations of the scaled data features and inspection of the KLT matrix gives some insight into which features are mainly responsible for the variability in the original data. Perhaps a more important aspect of scaling is its impact on the geometrical structure of the data, for it is precisely this type of structure that influences FCM clustering performance. In this example, for c = 2, the scaling procedures of methods 2 and 3 account for the fact that the FCM algorithm grouped the data from these methods into the two visually obvious clusters and the algorithm failed to do so for the unscaled data of method 1. In addition, a comparison of the scaling procedures of methods 2 and 3 shows that the scaling procedure of method 3 has the desirable property of maintaining better structural integrity of the labelled Iris subgroups as measured by the separation coefficients listed in Table IV.

One final tool which may be easily utilized is a different inner product norm for measuring dissimilarity between the sample points. Up to this point, the Euclidean norm has been used exclusively. The Euclidean norm 'seeks' spherically or hyperspherically shaped clusters. Changing **A** in Equation (12) to a positive definite, symmetric matrix other than the identity matrix will cause the clusters determined by the FCM algorithm to have a different shape. For example, consider the principal components data of method 3 above and let **A** equal the covariance matrix of the data. According to the discussion of Section 2.2.1 **A** is a diagonal matrix with the diagonal elements equal to the eigenvalues of the covariance matrix of the scaled Iris data. These eigenvalues are displayed in Table III. The FCM algorithm is run with the 'covariance' norm, $\varepsilon = 0.01$ and $m = 1.5$. The cluster validity measures F, G and ΔJ indicated that two and five were good choices for c and Figure 9 shows the resulting FCM clusters for $c = 5$. The clusters displayed in the two dimensional principal component space of Figure 9 tend to be elliptically shaped with the major axis of each ellipse lying perpendicular to the first principal component axis. Part of the motivation for trying the 'covariance' norm is the assumption that the first principal component has more physical significance than the other principal components. Any significant movement along the first principal component axis is assumed to correspond to a significant physical change. Use of the 'covariance' norm results in six misclassifications of the Iris data when the following correspondence is made between the clusters of Figure 9 and the original Iris samples.

Cluster #'s	Iris Species
1,2	virginica
3,4	veriscolor
5	setosa

There are many other candidates for the matrix **A** in Equation (12). For example, setting **A** equal to the inverse covariance matrix of the scaled data induces the Mahalonobis norm and the resulting clusters are again elliptically shaped, but oriented with the major axis parallel to the first principal component axis. The Mahalonobis norm is popular in many pattern recognition applications, but is not appropriate for the Iris data example. The point to be made is that if the shape of the clusters being sought is known then an appropriate norm can easily be added to the FCM clustering algorithm.

2.3 Chapter Summary

This chapter has introduced and illustrated the major tools which will be used in a borehole segmentation algorithm in Chapter III. The discussion and examples of this chapter help illustrate how the FCM algorithm, cluster validity measures and principal component analysis can be used to objectively evaluate the structure of a given data set. The use of cluster validity measures in conjunction with the FCM algorithm makes an excellent tool for exploratory data analysis because the FCM algorithm makes no explicit assumptions concerning the distribution of the data and the cluster validity measures eliminate the need to know a priori the number of subgroups in the data. This is very important since the intent of the analysis is to discover naturally occurring structure in a given data set rather than impose structure on the data set by making unwarranted or unnecessary assumptions. As was evidenced in the example of Section 2.1.3, the number of subgroups determined algorithmically do not necessarily agree with the number of subgroups determined physically. It was observed in the example of Section 2.2.2 that data scaling improved the correspondence between the FCM clusters and the labelled Iris subgroups. The choice of a particular scaling process is important because the clustering results of the FCM algorithm are influenced by the effect the scaling process has on the spatial distribution of the data. Additional insight is gained into the structure of the data by expressing the original data in terms of uncorrelated principal component data. Principal component analysis indicates which of the original data features account for the majority of the variation in the original data. Additionally, the use of principal components provides a better means to display FCM clustering results in two dimensions even though the clustering is being done in a higher dimension.

CHAPTER III

SEGMENTATION ALGORITHM

3.0 Introduction and Basic Methodology

The general segmentation problem involves dividing one or more time-varying signals into segments that are in some sense homogeneous. In the present study, the wireline logs are a function of depth rather than a function of time but similar analysis methods still apply. In the multivariate case, the segmentation process is performed on multiple signals that correspond to the same physical process over a given time interval. In problems of high dimesionality the KLT is often an effective transformation to reduce the dimensionality of the problem while still retaining the majority of the statistical information contained in the original signals. The segmentation problem is easily formulated as a pattern recognition problem and can be approached in a number of ways as outlined in Section 1.1. The following paragraphs give an overview of some possible approaches to the segmentation problem.

The signature recognition problem is a relatively narrow approach to the segmentation problem. In the signature recognition problem, it is desired to locate a particular segment of a signal rather than perform a general segmentation of the signal. A 'signature' signal corresponding to something of physical interest is known and the problem is to search other signals for the signature of interest. This approach is similar to pattern recognition by template matching but is complicated by the warping phenomena that the signature signal undergoes due to variations in the physical process. Cartinhour [14] addresses the signature recognition problem and makes specific application to the well log signature recognition problem.

A more general approach to the segmentation problem invloves the segmentation of a signal into k homogeneous segments where k is known a priori. Section 1.4 reviews a variety of segmentation techniques that operate under the assumptions that k is known and that homogenity is defined with respect to a normal distribution of the data. These assumptions reduce the segmentation problem to a parametric estimation problem.

Perhaps the most general approach, and the approach taken in this study, is when the segmentation problem is interpreted as an unsupervised pattern recognition problem and no assumptions are made with regard to the distribution of the data or the number of segments to be determined. The intent is to search for naturally occuring data structure rather than impose structure on the data with unnecessary assumptions.

The segmentation of a borehole into relatively distinct intervals based upon wireline log responses is accomplished by successive application of the Karhunen-Loeve Transform(KLT) and the Fuzzy-c-Means(FCM) clustering algorithm. The KLT gives insight into the structure of the data by expressing the original logs in terms of uncorrelated principal component logs. The FCM clustering algorithm clusters the principal component data into a specified number of clusters and then the cluster validity measures are used to objectively evaluate the clustering results. The basic segmentation algorithm methodology is outlined by the following steps.

1. Make sure that all wireline logs are on depth relative to each other and that any obviously errant or missing data has been corrected.

2. Pick an appropriate set of input logs to be analyzed.

3. Choose an appropriate scaling procedure for the analysis being performed.

4. Transform the scaled wireline logs into principal components(PC) logs using the discrete Karhunen-Loeve Transform(KLT).

5. Inspect the KLT matrix to help determine which logs account for the majority of the variance in the log data.

6. Cluster the PC log data using an FCM algorithm.

7. Use the cluster validity measures: F, G and ΔJ, to determine the number of clusters which best fits the data.

8. Plot the maximum membership cluster information as a function of depth, thus segmenting the borehole into distinct intervals.

9. If geological analyses of the same interval exist, then they can be used to evaluate the geological meaning of the intervals determined by the segmentation algorithm.

Certain variations from this basic methodology are taken in Chapter III as different segmentation strategies are evaluated, but steps 1-9 outline the key steps used in segmenting a borehole into distinct intervals based upon wireline log responses.

3.1 Data Base

The data base for the analysis presented in this chapter was supplied by AMOCO of Tulsa, Oklahoma and consists of over 1600 feet of log and core information [33]. A test interval of approximately 360 feet was chosen from this data base for detailed analysis. Figure 10 shows the gamma ray(GR), spontaneous potential(SP), short normal(SN), deep induction(ILD), neutron porosity(NPHI) and bulk density(RHOB) logs for the test interval. A simplified core description for the test interval is given in track #1 of Figure 10 and the same description is repeated in track #2. The numbers in tracks #1 and #2 represent different lithology types: 0-undefined, 1-shale, 2-sandstone, 3-limestone and 4-coal. Although there exists significant variation within each lithology, the description given in Figure 10 is consistent with the detailed core description given in the Appendix. The major points of the detailed core description given in the Appendix are summarized in the following paragraph.

The 360 ft. test interval consists of 113 cumulative feet of sandstone, 47.5 feet of limestone, 181.5 feet of shale, 3 feet of coal and 15 feet of unknown lithology. Three sandstones with apparently good reservoir potential occur at depths of 143-157.5 ft., 227-241 ft., and 267-311 ft. respectively. The sandstone units at 143-157.5 ft. and 267-311 ft. are the better developed of the three sandstones. There are essentially three limestone units within the interval. These occur at 6-42 ft., 101.5-111.5 ft., and 322-327 ft. respectively. The first limestone unit is interrupted by a shale bed approximately 6 ft. thick. All three limestone units exhibit apparently good reservoir potential. The shales in the interval are considered relatively poor quality reservoir rocks. Three thin coals occur at 46 ft., 142 ft. and 224 ft. and a very thin coal marks the end of the test section. Finally, there are three oil shows which occur in the test interval. The first two oil shows occur in the first limestone unit at 6-42 ft. and the third oil show is in the sandstone unit at 143-157.5 ft. More specifically, the first oil show has 1.5 net feet of oil in the interval from 17.3-24.3 ft., the second oil show has 4.0 net feet of oil in the interval from 34.3-38.3 ft. and the third oil show has 5.4 net feet of oil showing in the interval from 146.4-155.7. Of these oil shows, the latter two may be of slight economic importance. This core description information will be used to evaluate the geological nature of the borehole segments determined by the segmentation algorithm. This test interval was chosen from the original data base because it possesses a fairly comprehensive set of geological conditions. First, the test interval has a

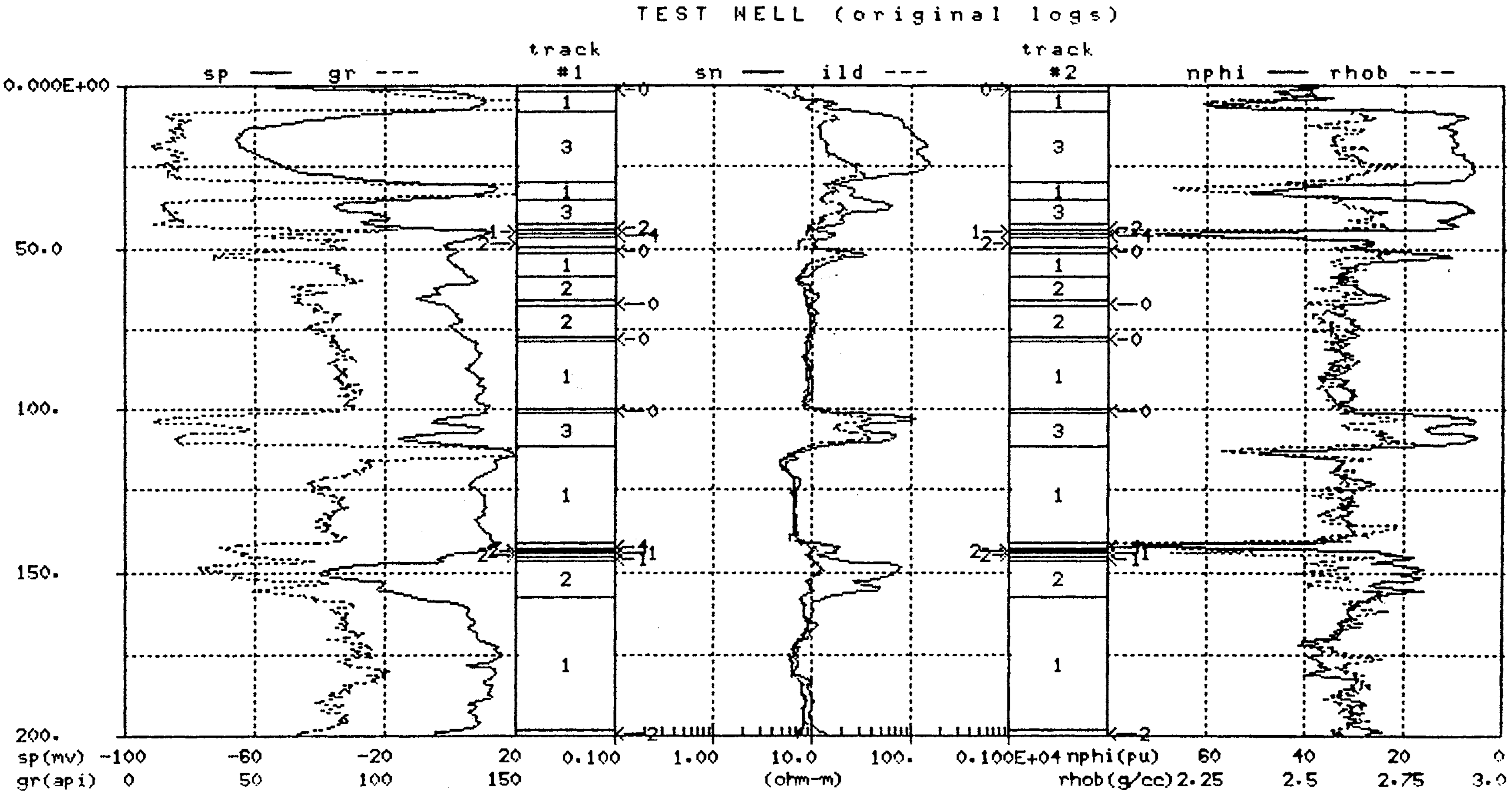


Figure 10. Original Wireline Logs with Simplified Core Description in Tracks #1 and #2 (0-undefined, 1-shale, 2-sandstone, 3-limestone, 4-coal)

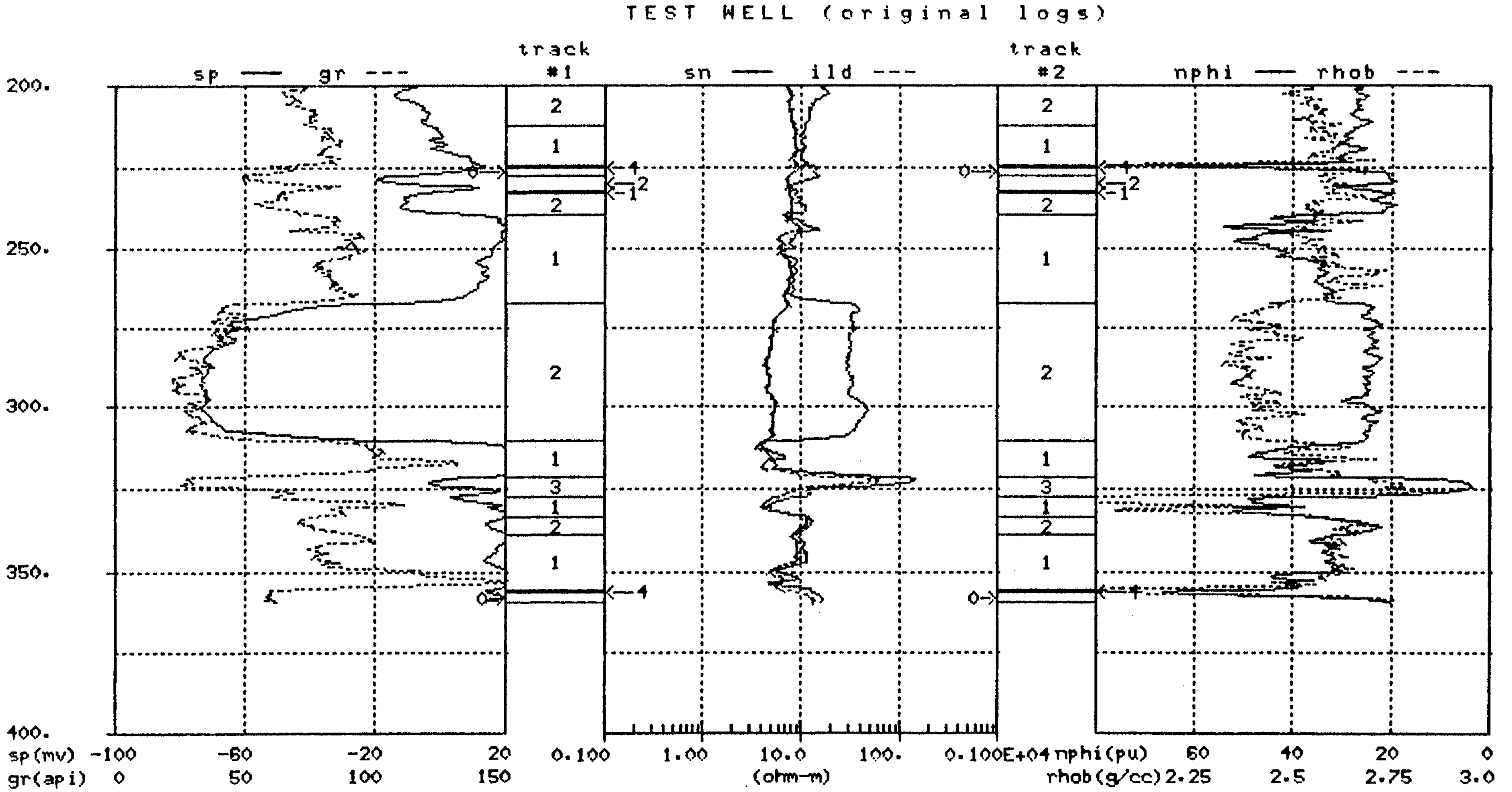


Figure 10. (continued)

relatively diverse lithology with four main lithological units: shale, sandstone, limestone and coal. Second, these lithological units occur as both thick well developed units and thin bed units. The thin beds encountered in the interval range from some relatively obscure sand/shale sequences to the very distinctive thin coal beds. Third, the test interval has both water bearing and hydrocarbon bearing zones. The most prominent water bearing zone occurs in the sandstone unit from 267-311 ft. and the hydrocarbon bearing zones are enumerated in the previous paragraph.

3.2 Segmentation Algorithm Examples

This section uses a series of four examples to demonstrate the performance of the borehole segmentation procedure on the test interval described in Section 3.1. In each example the physical significance of the segments is determined by comparing the borehole segmentation results with the core description information of the same interval.

The examples presented in this section are taken from a multitude of examples which were run on the test interval. Steps 3 and 6 of the basic methodology listed in Section 3.0 were given special consideration in the development of a specific segmentation algorithm.

Step 3 deals with the procedure used to scale the input logs. Data scaling is necessary on two counts. First, it is desired to account for the essential variability in the original log data and without some type of scaling process, those logs with the largest original variance would dominate the subsequent analysis. Without scaling, a log such as the bulk density log, which has a relatively small variance but is still very useful in porosity estimation and lithology determination, would not be properly accounted for in the determination of principal components. Second, if any type of metric is used as a measure of dissimilarity between log samples in the multidimensional data space, then it is necessary to have the respective axes, which form the data space, measured in similar units. The choice of a particular scaling method is important because data scaling affects the spatial distribution of the log data which in turn impacts the clustering performance of an FCM algorithm. Three different scaling methods were used on the wireline log data of the test interval. Two of these scaling methods are discussed in Section 2.2.2 and are referred to as scaling methods 2 and 3. The third scaling method consists of mapping the maximum log value within a given interval to plus one and the minimum log value to minus one and all other log values are scaled proportionately between the extreme values of plus and minus one. Of these three scaling methods, the one described as method 3 in Section 2.2.2 is used exclusively in the first three examples of this section and a slight variation of this scaling method is used in the fourth example. This chosen scaling method changes the original input logs to zero mean signals and then scales the logs using the maximum excursion from the mean. The resulting signals are zero mean and bounded by plus and minus one. The decision was made to use this scaling method over the other two methods because the resulting borehole segments had a more consistent physical interpretation. It should be noted that this physical interpretation was made using core description information which is primarily mineralogical and lithological in nature. In addition to the chosen scaling process, the analyst may choose to assign weights to the respective input logs depending upon the object of the analysis. However, this type of secondary weighting is not performed for the examples in this section.

Step 6 of the segmentation algorithm deals with the use of a particular FCM algorithm. The basic algorithm is outlined in Section 2.1.2 and requires a number of parameters to be set prior to application of the algorithm. Two FCM parameters of particular interest are the measure of dissimilarity, d, and the weighting exponent, m. The measure of dissimilarity directly affects the shape of the FCM clusters and the weighting exponent affects the amount of shared membership between fuzzy clusters. The Euclidean metric norm is used exclusively as a measure of dissimilarity in the four examples of this section. Another inner product norm, which sets the matrix $\mathbf{A}$ in Equation (12) equal to the covariance matrix of the principal components data, was investigated but this proved to be a very poor choice for the given application. Choosing an appropriate value for the weighting exponent was done empirically and a nominal value of $m=1.5$ is preferred for the log data in the test interval. Before proceeding with the segmentation examples, a brief overview of each example is given.

Example 1 uses all six logs shown in Figure 10 as inputs to the segmentation algorithm, varies c from 2 to 10 and uses the validity measures, F, G, and ΔJ to determine the value(s) of c which best fits the data. Example 2 removes the resistivity logs as inputs to the segmentation algorithm and uses the same clustering strategy as Example 1. Example 1 and Example 2 results are compared. Example 2 also demonstrates the impact of varying the weighing exponent, m, within the FCM algorithm. Example 3 uses a sequential clustering strategy. This strategy uses validity measures F and G to determine the number of clusters which best fits the data and then the FCM algorithm is applied to each of these clusters to further subdivide them. This sequential process is continued as long as there is evidence to warrant its continuation. Example 3 results are contrasted with previous results. Finally, Example 4 is similar to Example

1 but modifies the scaling process to take the common logarithm of the resistivity logs prior to applying the linear scaling method used in the first three examples. This modification lessens the influence of the exceedingly large values that are sometimes encountered in the resistivity logs.

3.2.1 Example #1

The first example uses the six logs shown in Figure 10 as inputs to the segmentation algorithm. These logs have been depth shifted and the log depths have been adjusted to match the core depths given in the Appendix. Each input log is scaled using the linear scaling procedure of Method 3 outlined in Section 2.2.2. The resulting scaled logs are zero mean with magnitude between plus and minus one. Principal components(PC) logs are calculated from the scaled wireline logs using Equation (28) and are displayed in Figure 11 along with the simplified core description in tracks #1 and #2. The KLT matrix in Equation (28) is derived from the covariance matrix of the scaled wireline logs as described in Section 2.2.1. First consider PC logs #1 and #2 since they account for the majority of variation in the original data and they provide a useful way to display clustering results. Inspection of the KLT matrix shows that PC log #1 is dominated by

$$
\begin{bmatrix} pc\#1 \\ pc\#2 \\ pc\#3 \\ pc\#4 \\ pc\#5 \\ pc\#6 \end{bmatrix} = \overset{\underline{\text{KLT Matrix}}}{\begin{bmatrix} 0.199 & 0.851 & -0.321 & -0.057 & 0.360 & 0.007 \\ 0.113 & -0.419 & -0.295 & -0.310 & 0.628 & -0.484 \\ 0.054 & 0.158 & 0.745 & 0.388 & 0.332 & -0.396 \\ 0.527 & -0.192 & 0.265 & -0.121 & 0.367 & 0.682 \\ 0.816 & -0.036 & -0.102 & 0.110 & -0.433 & -0.349 \\ 0.001 & 0.191 & 0.418 & -0.850 & -0.210 & -0.148 \end{bmatrix}} \begin{bmatrix} gr \\ sp \\ sn \\ ild \\ nphi \\ rhob \end{bmatrix} \qquad (28)
$$

the SP log and PC log #2 is mainly dependent on the NPHI, RHOB and SP logs. At this point the analyst may wish to go back and reweight the original input logs if it seems the first few PC logs are unduly influenced by particular input logs. Adjusting the weights of the input logs is a subjective modification and is not performed in this or any subsequent examples in this chapter.

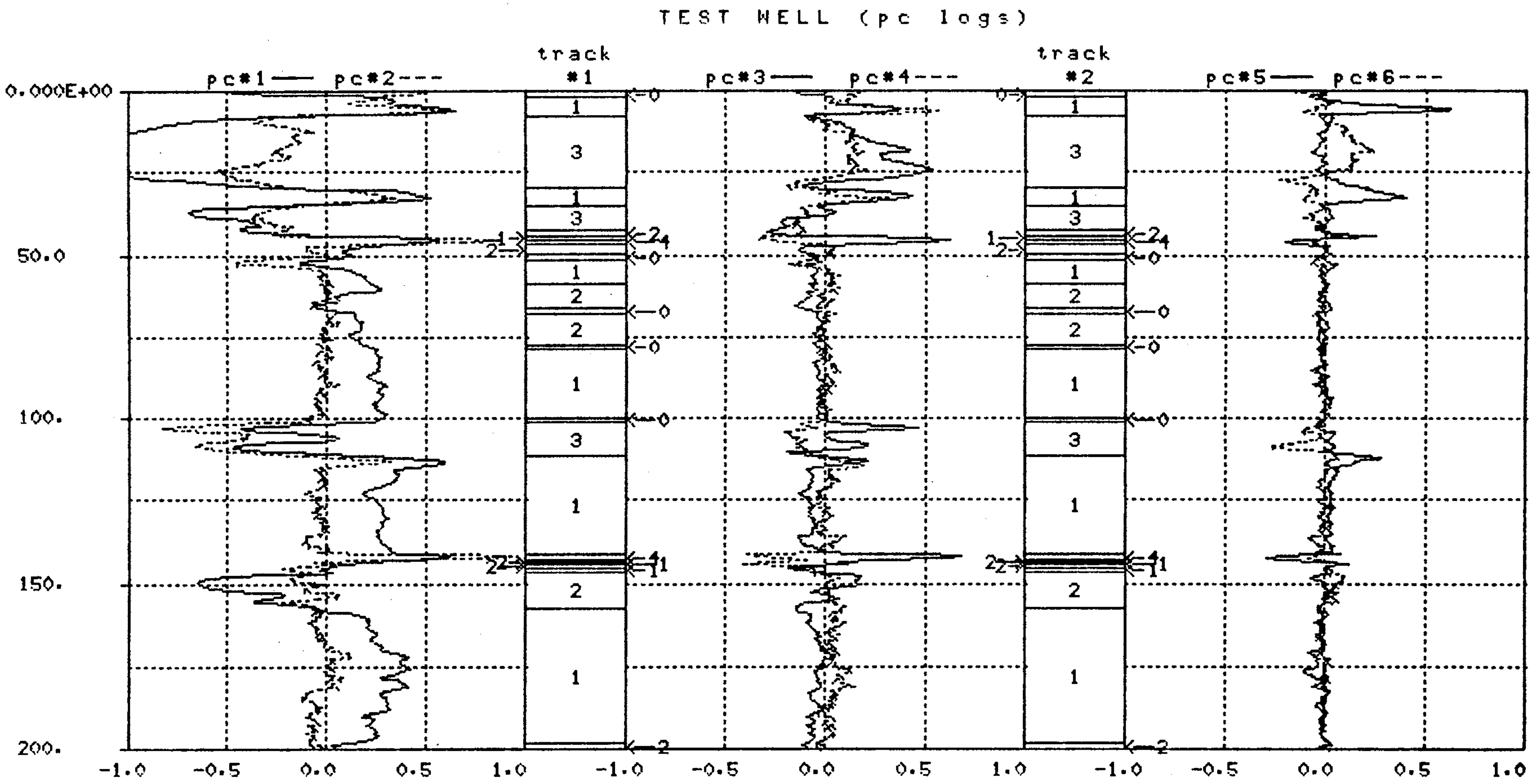


Figure 11. Principal Component Logs with Simplified Core Description in Tracks #1 and #2
(0-undefined, 1-shale, 2-sandstone, 3-limestone, 4-coal)

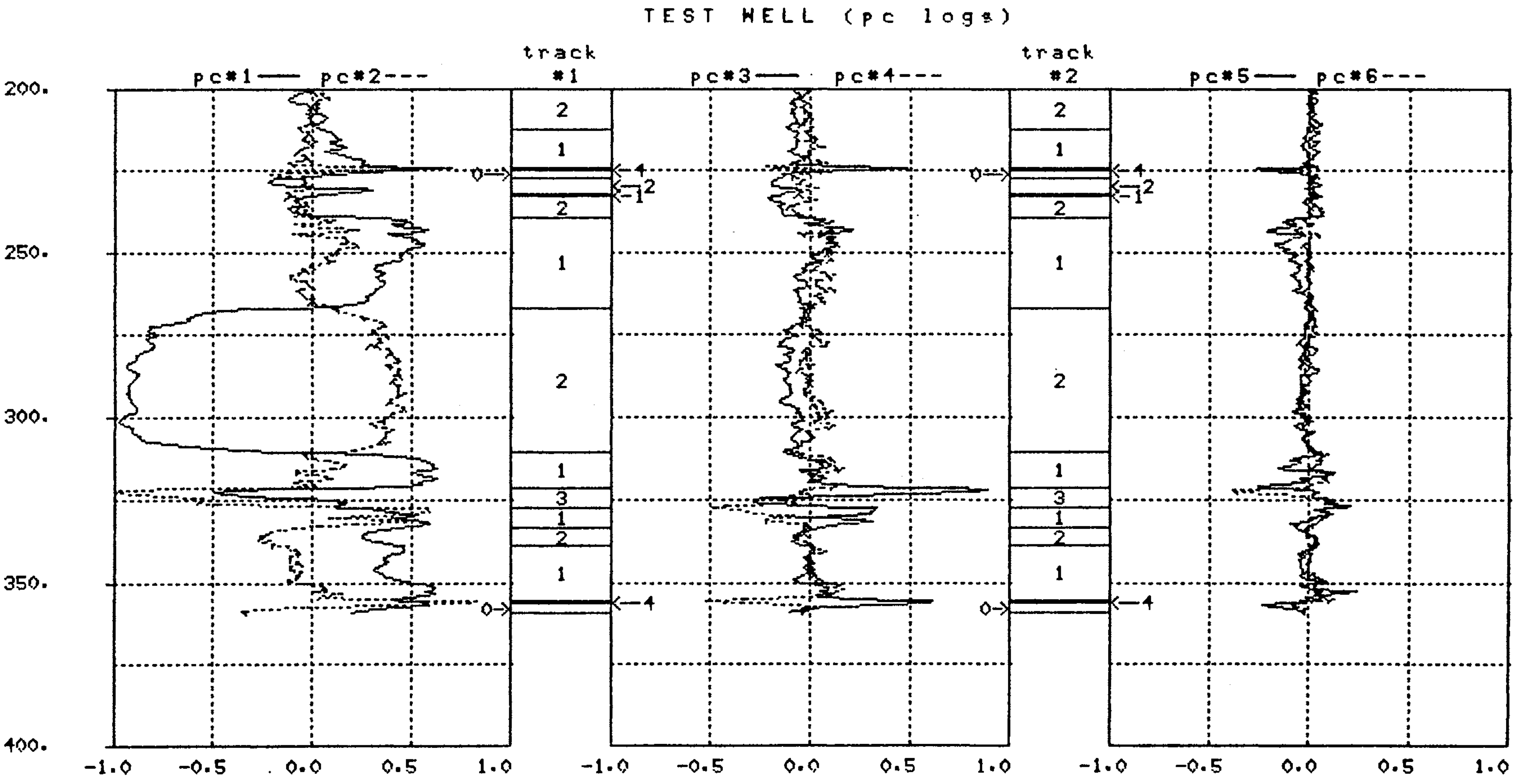


Figure 11. (continued)

Before applying the segmentation algorithm to the PC log data it will be beneficial to view the core description information in the two dimensional PC space created by PC logs #1 and #2 as shown in Figure 12. Notice a general clustering of the respective lithology types regardless of where they occur in the borehole. The coal samples tend to cluster in the upper right hand corner of Figure 12 while the shale samples are concentrated in the right center portion of the display. The sandstone samples have significant variation along the first PC axis ranging from middle center to the upper left hand portion of Figure 12 and finally, there are two visually distinct groupings of limestone samples at the left center and lower center portions of the figure. Those samples where oil shows are present are circled in Figure 12. It is informative to consider each lithology type in more detail.

The shale samples from Figure 12 are shown separately in Figures 13(a) and 13(b). The display in Figure 13(a) helps establish the exact boundary of the shale samples since there is considerable overlap between the shale and sandstone samples in Figure 12. The display scale is expanded and the shale samples are shown in more detail in Figure 13(b). The numbers in Figure 13(b) correspond the the identifying numbers given to each shale segment described in the Appendix. Most of the numbers are obscured due to the high concentration of samples in the center of Figure 13(b) but it is interesting to note the character of some of the outlying shale samples. For example, the samples from the top portion of shale segment #11 and the samples from the top portion of shale segment #16 separate out just to the left of the main concentration of points in the center of Figure 13(b). Shale segment #11 is part of a sandstone to shale transition and is described as dark gray and interlaminated with a very minor amount of ripple laminated light gray, micaceous sandstone. Shale segment #16 is also part of a sandstone to shale transition and is described as approximately 50% sandstone at the top grading downward to 100% shale at the bottom. In contrast to the sand/shale transitions are the black, organic rich shale segments which also tend to separate out from the main cluster of shale samples. Included in this group are shale segments: #1, #2, #7, #21 and #28. Other outlying shale segments include segments #18 and #19 which are dark gray to black carbonaceous shales. It is also of interest to note the character of the shale samples which adjoin the coal samples shown in Figure 12. These shale samples are shown at the very top of Figure 13(b) and come from shale segments #1, #23, #24 and #25. The attribute common to these shale samples is that one or both of the porosity logs(NPHI and RHOB) in Figure 10 indicate a very high porosity. Shale segment #1 is described as a black, organic rich shale, shale segments #23 and #25 are black

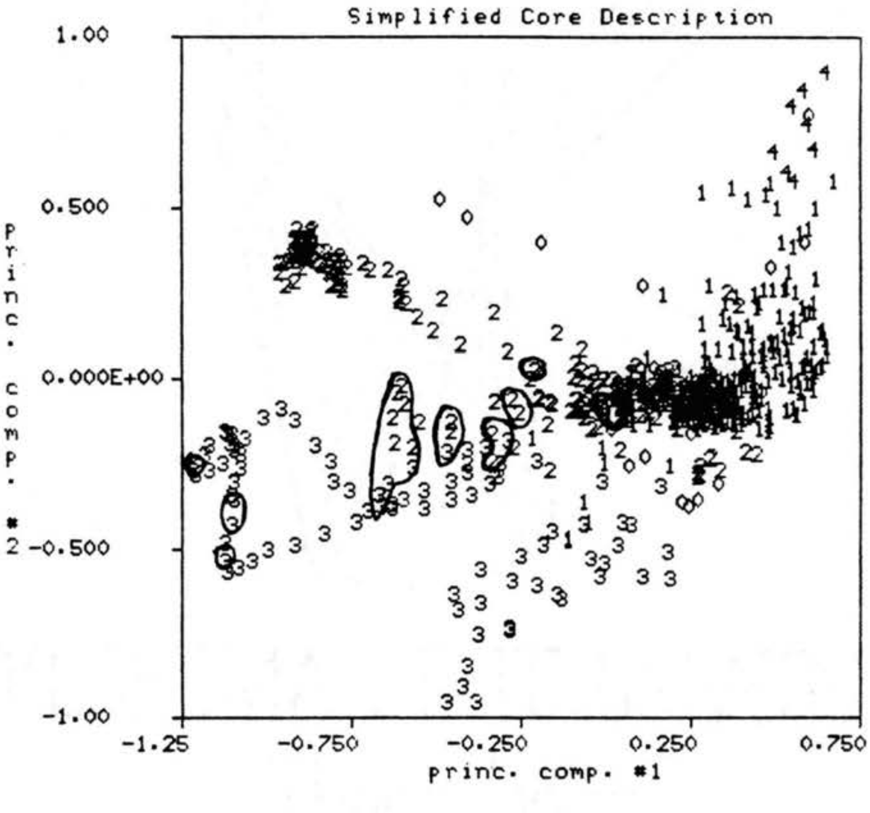
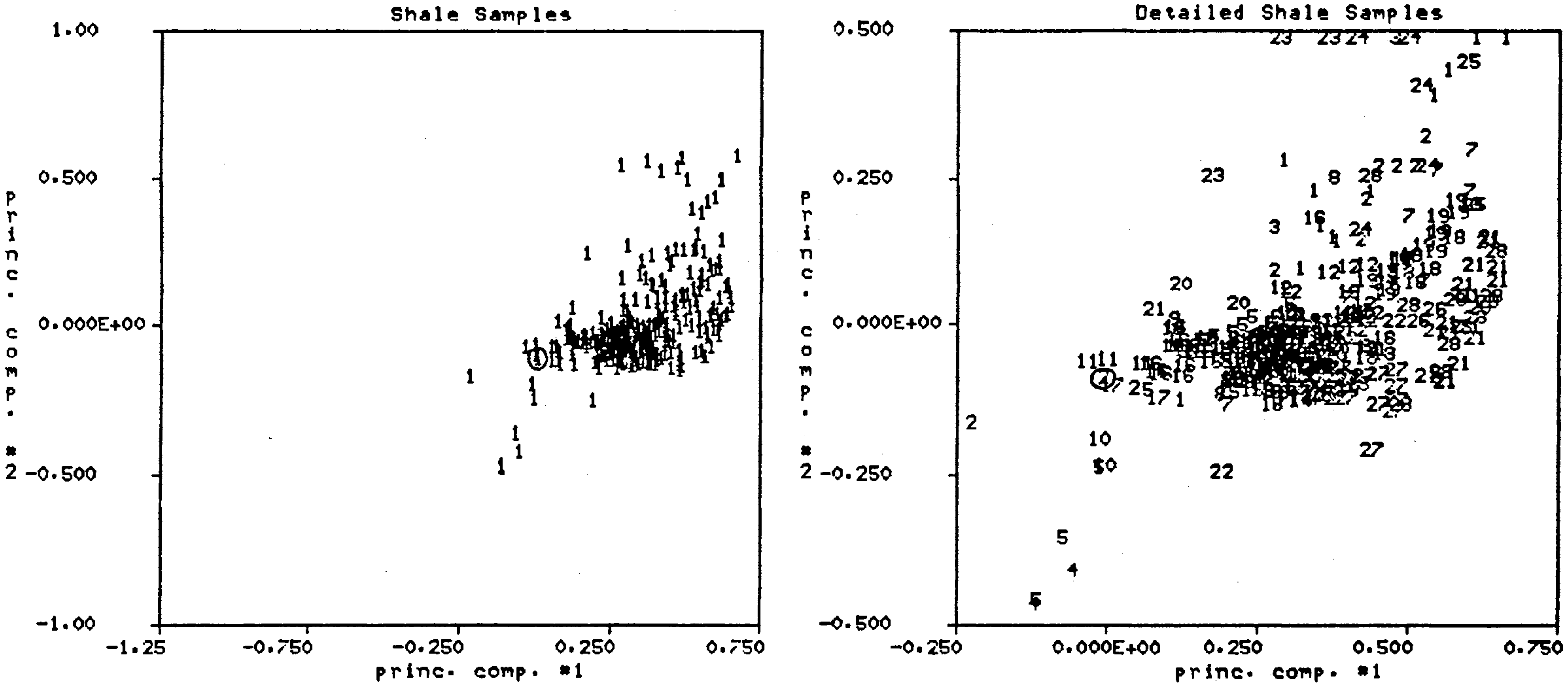


Figure 12. Simplified Core Information Displayed in Two Dimensional PC Space with Oil Shows Circled



(a)

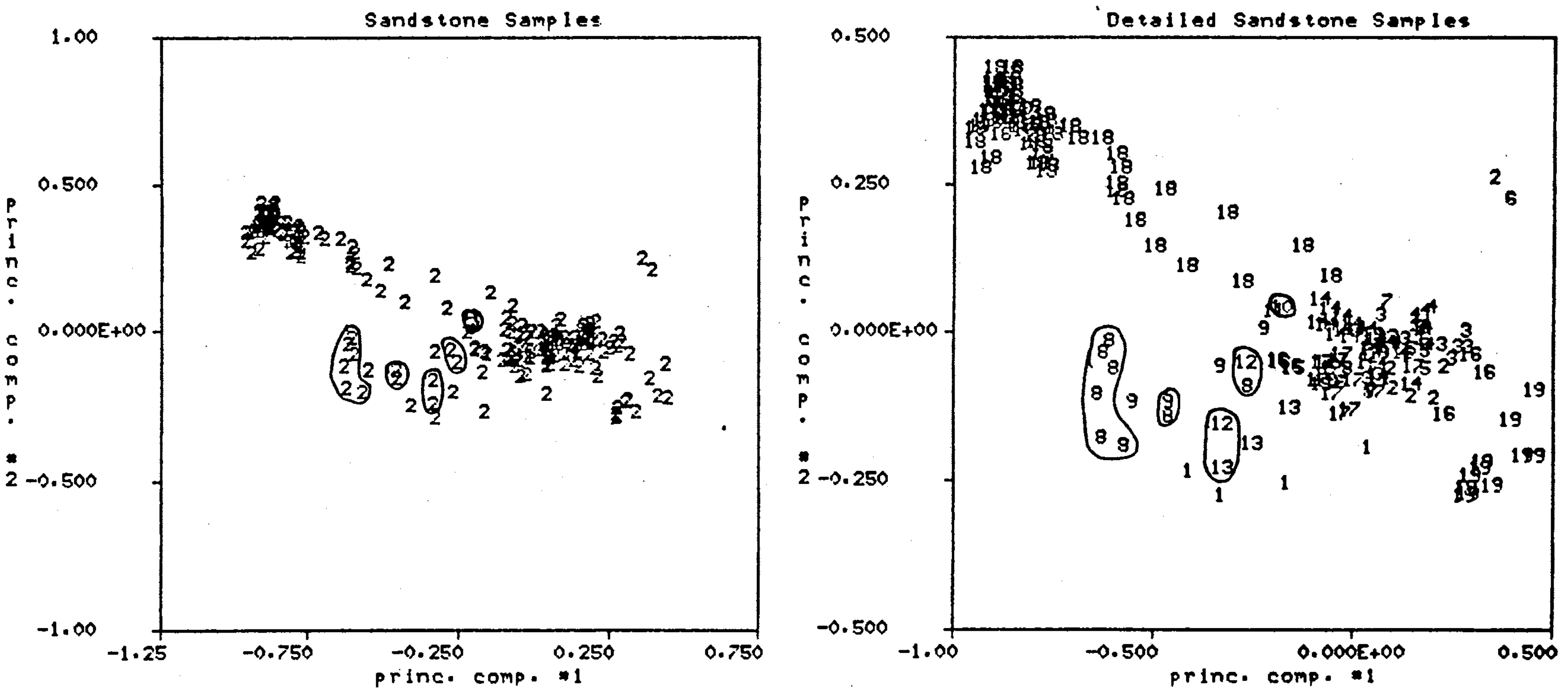
(b)

Figure 13. Core Shale Samples, (a) from Figure 12 and (b) Detailed Description from Appendix

fossiliferous shales and segment #24 is a gray to green, calcareous shale with high clay content.

At the other extreme are the shale samples from segment #4 and the first part of segment #5 shown in the lower left hand corner of Figure 13(b). The distinguishing characteristics for these shale samples are a relatively low porosity as indicated by the NPHI and RHOB logs as well as an uncharacteristicly high resistivity reading. It is possible that the information on the core and log is slightly off depth at this point since the log signature at 51.5-53.5 ft. is characteristic of a thin limestone and this interval is adjacent to a segment where the core was missing. Shale segment #4 is described as a gray, very calcareous, fossiliferous, arenaceous, pyritic shale and segment # 5 is a gray, pyritic shale that is arenaceous in the top one ft. In general, the distribution of the shale samples as viewed in the two dimensional PC space of Figure 13 is a function of permeability and porosity. Recall that the first PC log is dominated by the SP log which gives a relative indication of permeability and the second PC log in most heavily dependent on the porosity logs as well as the SP log.

In a manner similar to the shale samples, the sandstone samples of Figure 12 are separated out and displayed in Figures 14(a) and 14(b). As in Figure 12 all circled samples indicate the presence of hydrocarbons. There are two main groupings of sandstone samples. The first grouping is in the upper left hand corner of Figure 14(b) and corresponds to sandstone segment #18 which is a well developed water-bearing sandstone unit between 267-311 ft. The second grouping is in the right center portion of Figure 14(b) and represents several sandstone segments all of which contain varying amounts of shale or clay and are not as well developed as sandstone unit #18. For example, sandstone segment #14 is described as a light gray, coarse grained sandstone interbedded with dark gray shale and the sandstone content is estimated at over 75% at the top and grading gradually downward to approximately 50% at the bottom. Sandstone segments #8-#12 represent a hydrocarbon bearing sandstone unit between 146-157 ft. and these samples are sparsely distributed to the left of the second main grouping of sandstone samples. The samples from segments #8-#12 are described as having abundant shale partings or interlaminated with dark gray shale but generally speaking the sandstone content exceeds 70% through this interval. The samples from sandstone segment #19 separate out from the rest of the samples due to their particularly low SP values. This is seen in Figure 10 between 334-340 ft. Finally the two stray samples in the upper right portion of Figure 14(b) tend to separate out due to the influence of nearby coal beds.



(a)

(b)

Figure 14. Core Sandstone Samples, (a) from Figure 12 and (b) Detailed Description from Appendix

Figure 15 shows both the limestone and coal samples from Figure 12. The coal samples from three thin coal beds all cluster in the upper right hand corner of Figure 15(b). The thin coals are characterized by a very distinctive spike on the porosity logs as seen in Figure 10. The limestone samples cluster in two, rather sparse, clusters in the lower left hand portion of Figure 15(a). The limestone samples are described as five separate segments in the detailed description in the Appendix and are shown in Figure 15(b). Limestone segments #1 and #2 come from what was described in Section 3.1 as a single limestone unit between 6-42 ft. which is interrupted by a shale bed. Limestone segment #2 is interbedded with dark gray shale and clusters just to the right of the majority of segment #1 samples. Limestone segments #3 and #4 comprise the limestone unit between 101.5-111.5 ft. and limestone segment # 5 is the limestone unit between 322-327 ft. All five limestone segments are described as argillaceous and segments #3 and #4 are described as very argillaceous. The circled samples from segments #1 and #2 indicate oil shows and denote one obvious difference between the samples in segments #1 and #2 and the samples in segments #3, #4 and #5. Another difference between the two limestone groupings is that segments #3, #4 and #5 contain fossils. As with the other lithology types, most of the variation for the limestone samples along the first PC axis can be accounted for by observing the relative magnitude of the SP log opposite the respective limestone segments in Figure 10.

The somewhat detailed description of the test interval given in the preceding paragraphs and in Figures 13, 14 and 15 will prove to be a valuable tool in a visual evaluation of the clusters determined by the segmentation algorithm. The remainder of Example #1 will apply the basic segmentation algorithm described in Section 3.0 to the log data and relate the resulting segments back to the available core information.

Example #1 continues with step 6 of the the methodology listed in Section 3.0. The PC log data is clustered using an FCM algorithm. The Euclidean norm is used exclusively in this and subsequent examples since there is no apparent reason for using a different inner product norm. Other algorithmic parameters for the FCM algorithm are set as follows: c is allowed to vary from 2 to 10, m is set to a nominal value of 1.5 and ε = 0.01. The FCM algorithm is initiated by selecting c equally spaced points along each of the PC axes for the initial cluster centers and proceeding iteratively starting at step 3 of the FCM algorithm described in Section 2.1.2. The dimensionality of the clustering problem is varied from six to three to two in an effort to determine the differences in clustering performance using all six PC logs, the first three PC logs and the first two PC logs.

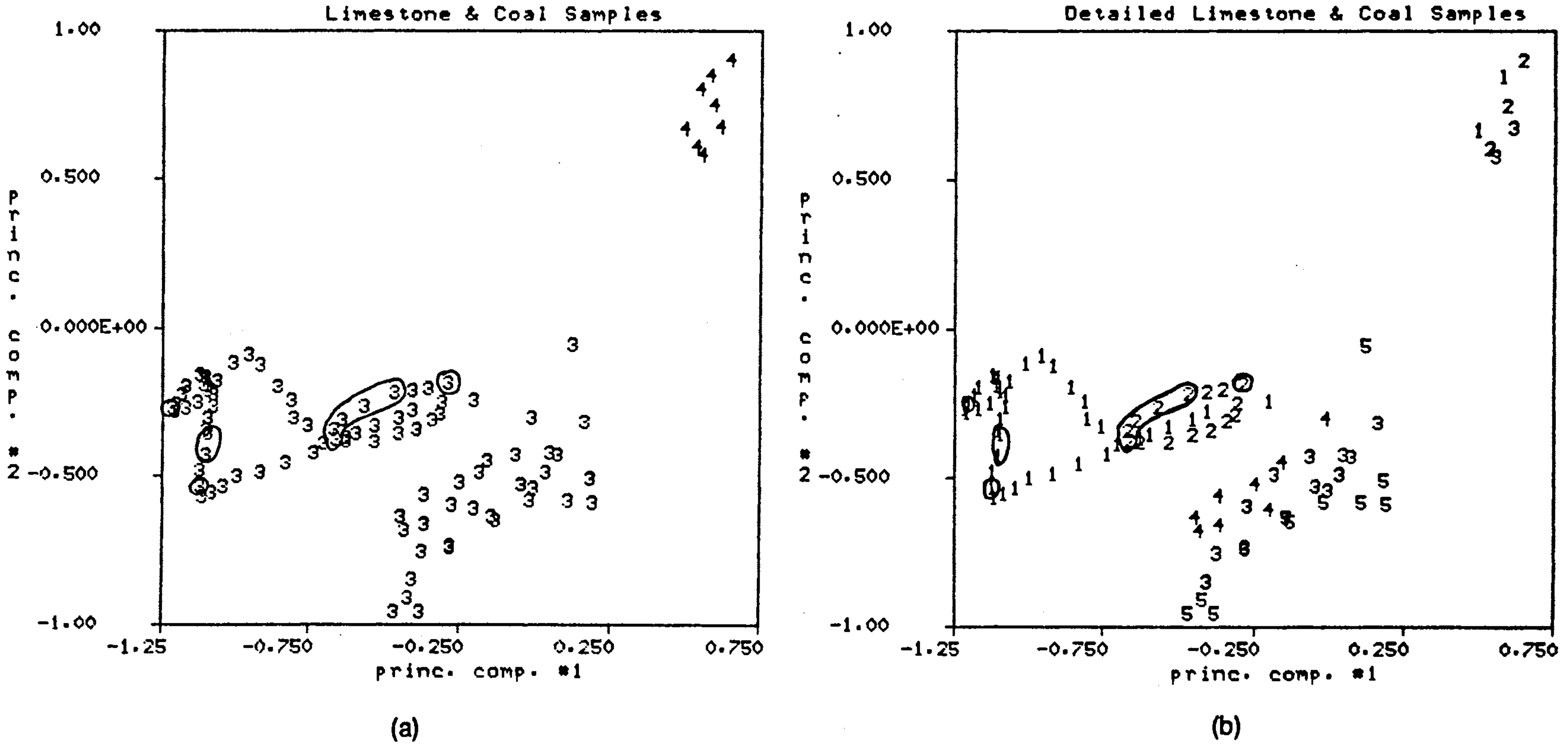


Figure 15. Core Limestone and Coal Samples, (a) from Figure 12 and (b) Detailed Description from Appendix

Table VII shows the FCM cluster validity measures for the six dimensional, three dimensional and two dimensional clustering cases. These validity measures are discussed at the end of Section 2.1.2 and interpretation of the validity measures is not always clear cut as was illustrated in the examples of Chapter II. Recall that cluster validity measure, F, indicates the amount of unshared membership between fuzzy clusters and larger F values indicate better structure in the data. The separation coefficient, G, is an indication of the spatial separation of the maximum membership hard clusters derived using Definition 3 and a fuzzy partition of the data. The values of G in Table VII indicate the worst case separation between a pair of hard clusters and large values of G indicate better spatial separation. The objective function coefficient, ΔJ, is a measure of how close the fuzzy partition is to the maximum membership hard partition in terms of within-group-sum-of-square-error and small values of ΔJ indicate better structure in the data. For this example, the 'best' value of c is determined by looking for some type of consensus among the three validity indicators.

In Table VII, maximum F and maximum G are achieved when c=2 for all three cases. However, minimum ΔJ is achieved when c = 10 in the six and three dimensional cases and when c=9 in the two dimensional case. It should be pointed out that 10 was judged to be an appropriate terminal value for c based upon the difference between successive values of the objective function, J_m. The FCM algorithm attempts to minimize J_m and when the objective function is not reduced significantly by incrementing c, then there is little reason to continue the process. Clearly, 2 is a reasonable choice for c based upon the validity measures F and G and Example #3 in this section pursues a sequential clustering procedure which applies the FCM clustering algorithm to each of the two clusters separately. The present example is concerned with trying to pick an appropriate value of c based upon the validity measures shown in Table VII and for the moment c=2 is excluded.

Now consider the validity measures of Table VII. One guideline used in the interpretation of the validity measures is to look for values of c where the partition coefficient, F, goes contrary to its decreasing tendency and increases or decreases only slightly from the F value for the previous value of c. The F values for c=3, 4 and 5 in Table VII(a) are all of comparable magnitude but c=5 is judged to be the better choice for c based upon the values of G and ΔJ. In a similar fashion, 7 and 10 are possible choices for c, but G indicates there is poorer spatial separation among the maximum membership clusters for these values of c and this fact detracts from choosing c=7 or c=10. Interpretation of the validity measures in Table VII(b) indicates that 4, 7 and 10 are reasonable choices for c but, here again, spatial separation among the clusters

TABLE VII.

FCM VALIDITY MEASURES FOR TEST WELL DATA: WITH EUCLIDEAN NORM, ε = 0.01 AND m =1.5

c	F	G	ΔJ
2	0.942	−1.652	7.751
3	0.858	−4.285	9.998
4	0.849	−2.697	7.946
5	0.842	−2.581	7.214
6	0.816	−3.255	8.118
7	0.822	−3.802	6.812
8	0.771	−5.231	7.339
9	0.759	−6.399	6.246
10	0.772	−5.870	5.223

(a) Six Dimensional PC Space

c	F	G	ΔJ
2	0.947	−1.533	6.640
3	0.873	−3.934	7.507
4	0.883	−2.189	5.385
5	0.865	−2.553	4.930
6	0.824	−2.506	4.804
7	0.850	−3.008	4.327
8	0.840	−3.391	3.780
9	0.798	−4.586	3.369
10	0.817	−4.114	3.159

(b) Three Dimensional PC Space

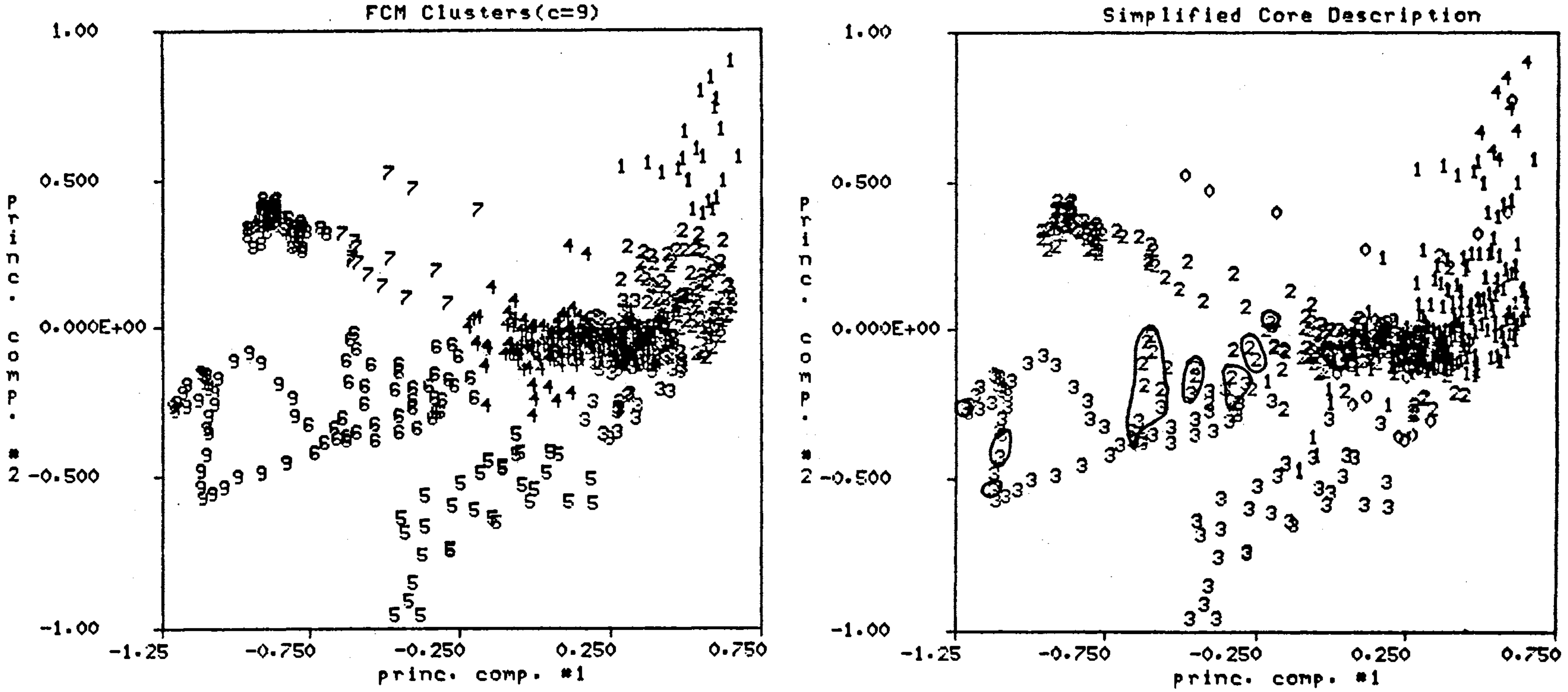
c	F	G	ΔJ
2	0.953	−1.045	4.975
3	0.902	−1.915	4.467
4	0.868	−2.548	4.571
5	0.870	−1.721	3.729
6	0.865	−2.626	2.886
7	0.862	−2.609	2.528
8	0.871	−2.100	1.664
9	0.870	−1.611	1.422
10	0.824	−3.325	1.529

(c) Two Dimensional PC Space

deteriorates for the higher values of c. Unlike the six and three dimensional clustering cases, there is better agreement among the validity measures for the two dimensional case. The validity measures for the two dimensional case are shown in Table VII(c) and 9 is judged to be the best value for c. When c=9, ΔJ is minimum, G is maximum(excluding c=2) and F indicates relatively good structure exists in the fuzzy partition of the log data. In the present example the best data structure seems to exist for the two dimensional case and c=9.

Figure 16 compares the FCM maximum membership clusters for the two dimensional case and c=9 with the simplified core description information from Figure 12. The information in Figure 16 is displayed in the two dimensional space formed by PC logs #1 and #2. Figures 13, 14 and 15 aid in making the following general correspondence between the FCM clusters and the core description information. The same FCM cluster information and core information is shown in Figure 17 along with the original wireline logs. The core information is shown in track #1 of Figure 17 and the FCM cluster information is shown in track #2. There are 22 samples in FCM cluster #1, 8 coal samples, 12 shale samples and 2 samples which were undefined by the core description information. It is disappointing that the coal samples are not more distinct in the clustering process since they have such a distinctive signature on the porosity logs but, as will be seen, this result occurs consistently in subsequent examples. It is observed in Figure 17 at 5-7 ft. and again at 328-332 ft. that the shale samples included in cluster #1 also have a relatively high porosity indicated on the porosity logs. FCM clusters #2 and #3 do a reasonable job of encompassing the remainder of the shale samples. One notable exception occurs at 334-340 ft. where a sandstone segment is included with the samples in cluster #3. FCM cluster #4 corresponds mainly to the sandstone segments which are argillaceous or interbedded with shale. There are also several very thin segments labelled 4 in track #2 of Figure 17 that occur in transition

FCM Cluster #	Description
1	coal/shale
2,3	shale
4	shaley sandstone
5,9	limestone
6	limestone/sandstone/oil shows
7,8	sandstone



(a)

(b)

Figure 16. Comparison of (a)FCM Maximum Membership Clusters with (b)the Simplified Core Description from Figure 12

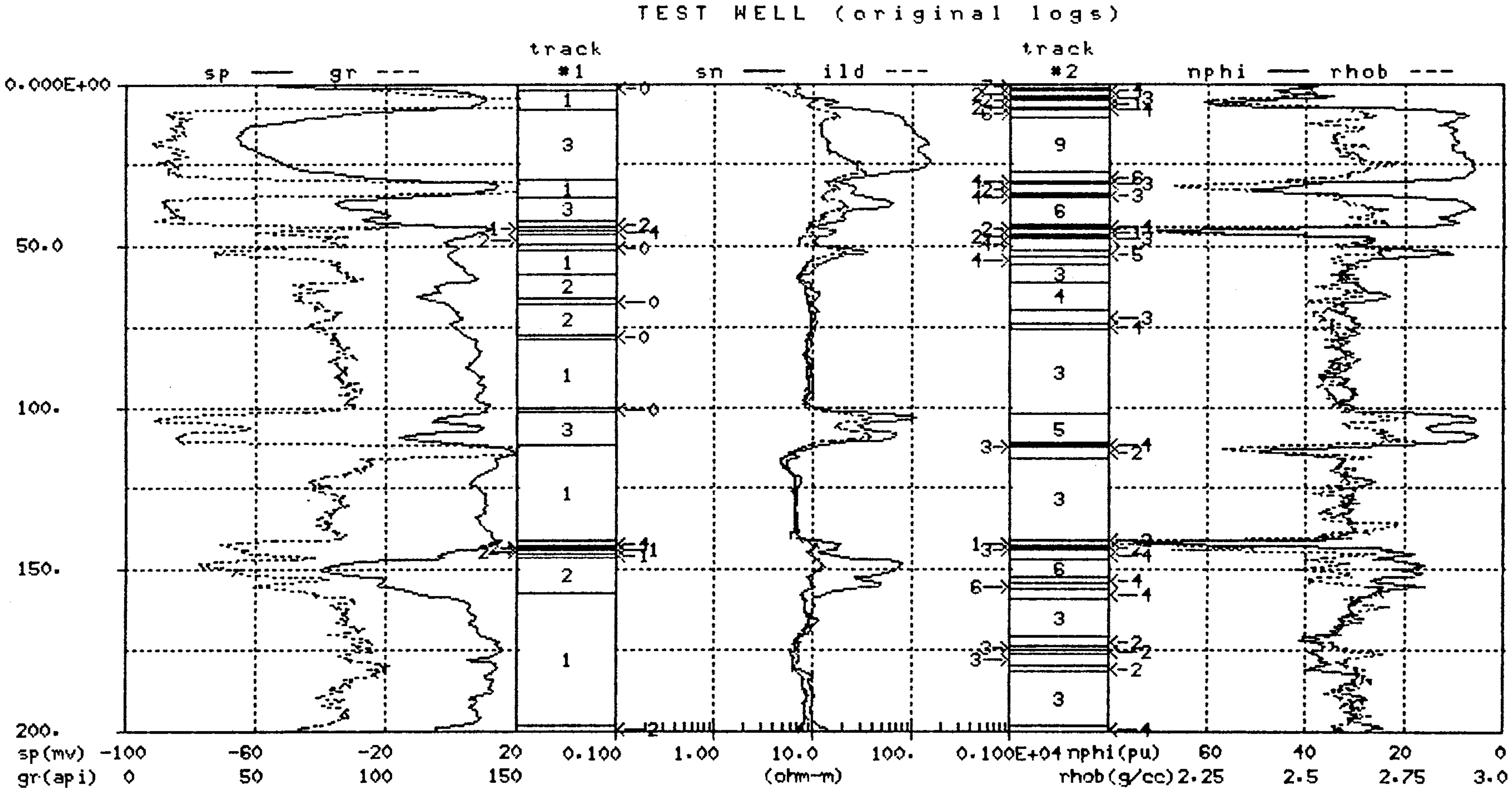


Figure 17. Comparison of Simplified Core Description Information (track #1) with the Two Dimensional Segmentation Algorithm Output, c=9 (track #2)

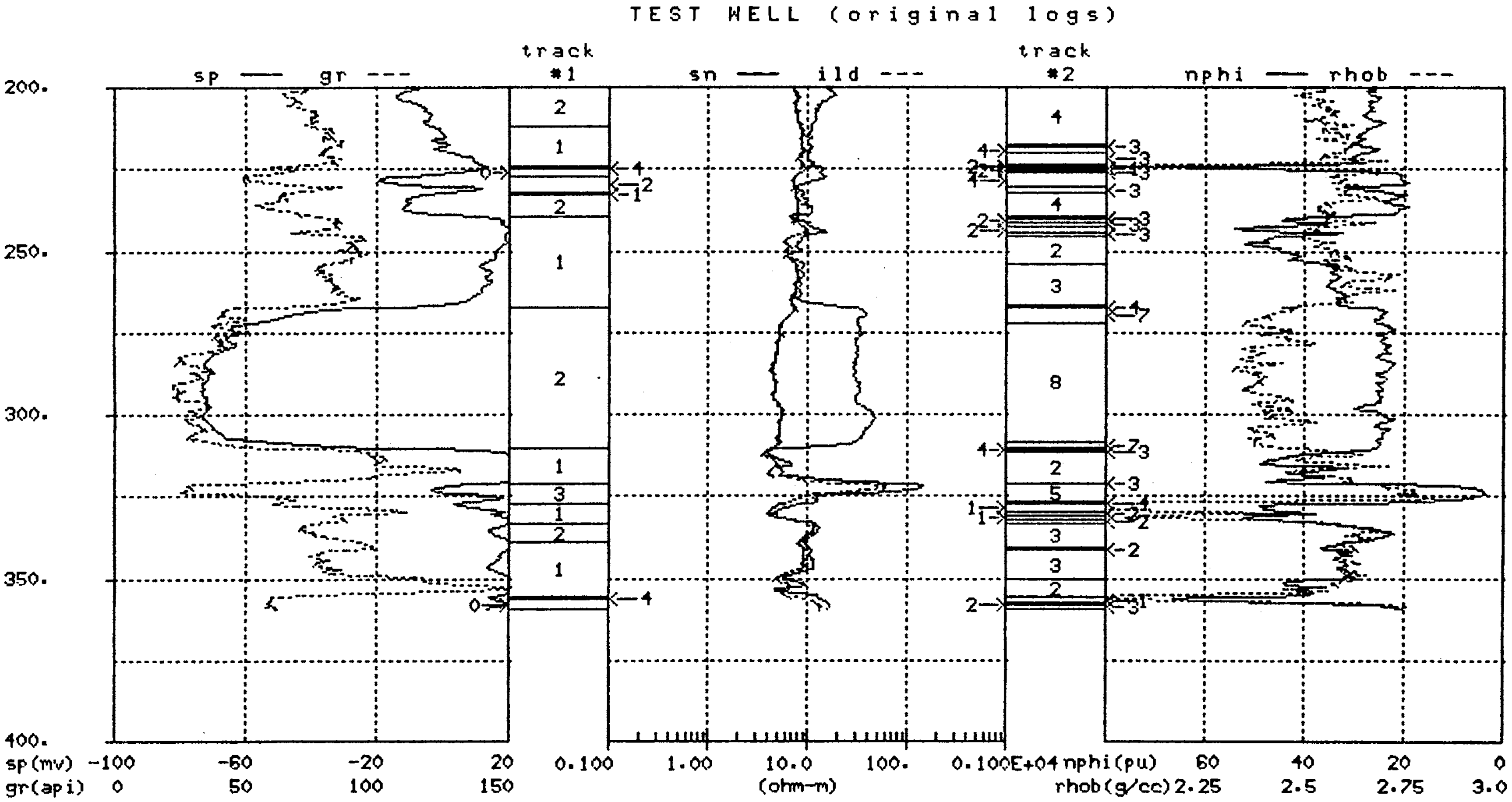


Figure 17. (continued)

regions between limestone and shale segments. FCM cluster #5 includes the limestone units at 101.5-111.5 ft. and 322-327 ft. as well as a small segment at 51.5-53.5 ft. which is described as a shale in the core information but has an uncharacteristic log signature for a shale. FCM cluster #6 has both limestone and sandstone samples. The majority of the samples from the limestone unit at 8-28 ft. are included in FCM cluster #9 but cluster #6 has those samples which correspond to the transition between the overlying and underlying shales. FCM cluster #6 includes all the samples from the limestone unit at 33-42 ft. as well as the majority of the samples from the hydrocarbon bearing sandstone at 146-156 ft. This particular cluster is not as distinctive in terms of lithology as the other FCM clusters and illustrates how different lithologic units can have similar log signatures. The common characteristics of the limestone at 33-42 ft. and the sandstone at 146-156 ft. are shaliness and significant hydrocarbon content. Finally, FCM clusters #7 and #8 correspond to the water bearing sandstone unit at 267-311 ft. To facilitate the comparison between the FCM clusters and the core information, certain FCM clusters are consolidated using the same general correspondence noted earlier on page 75. Figure 18 displays this simplified description of the FCM clustering results in track #2 along with the core information in Track #1.

New Cluster #	FCM Cluster #	Description
1	1	coal/shale
2	2,3	shale
3	4	shaley sandstone
4	5,9	limestone
5	6	limestone/sandstone/oil shows
6	7,8	sandstone

Displays similar to Figure 18 will be used in subsequent examples to provide a basis of comparison for different clustering results.

The physical interpretation of the FCM clustering results is most consistent for the two dimensional case displayed in Figures 16 and 17 but, to complete this example, let's consider the six and three dimensional clustering results. There is no clearly best choice for c in either the six or three dimensional case but, c=7 and c=10 are viewed as reasonable choices for both cases and are considered here. Figures 19 and 20 show the FCM clusters, when c=7 and c=10, for the six and three dimensional cases respectively. The FCM clusters in Figure 19(a) may be evaluated visually using the core information which is displayed in Figures 12, 13, 14 and 15. Cluster #1 contains all the coal

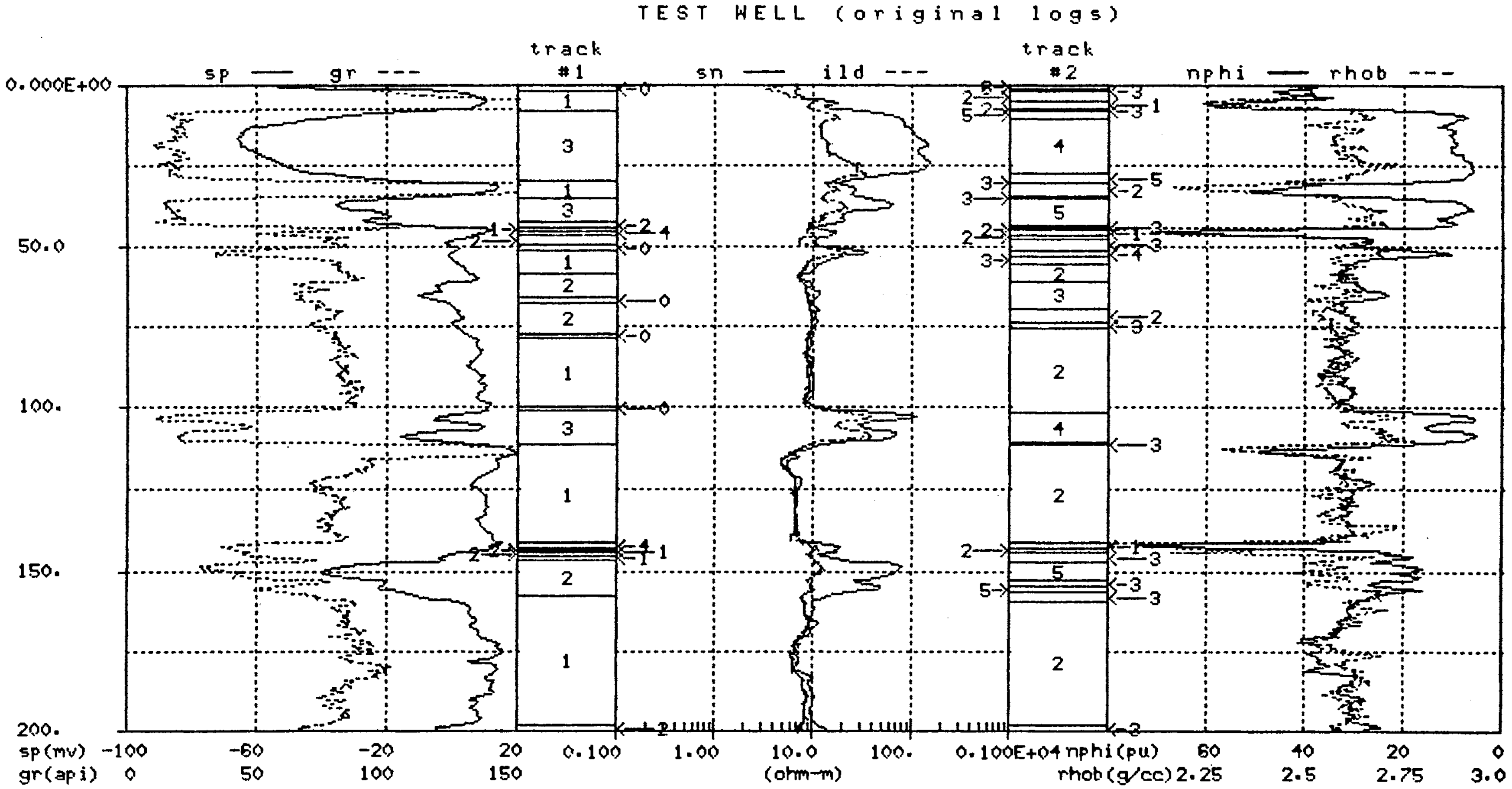


Figure 18. Comparison of Simplified Core Description Information (track #1) with the Simplified Segmentation Algorithm Output (track #2)

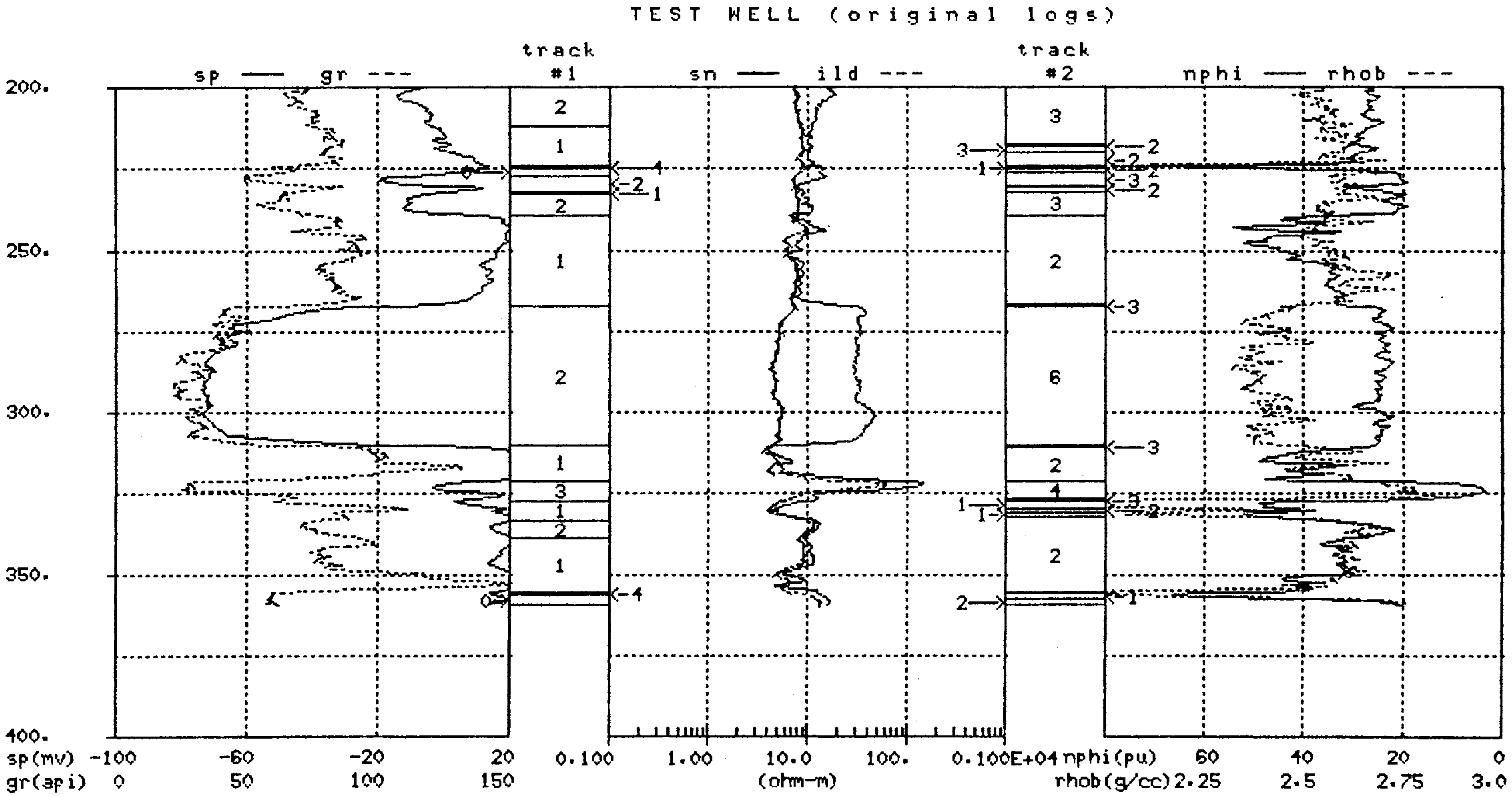
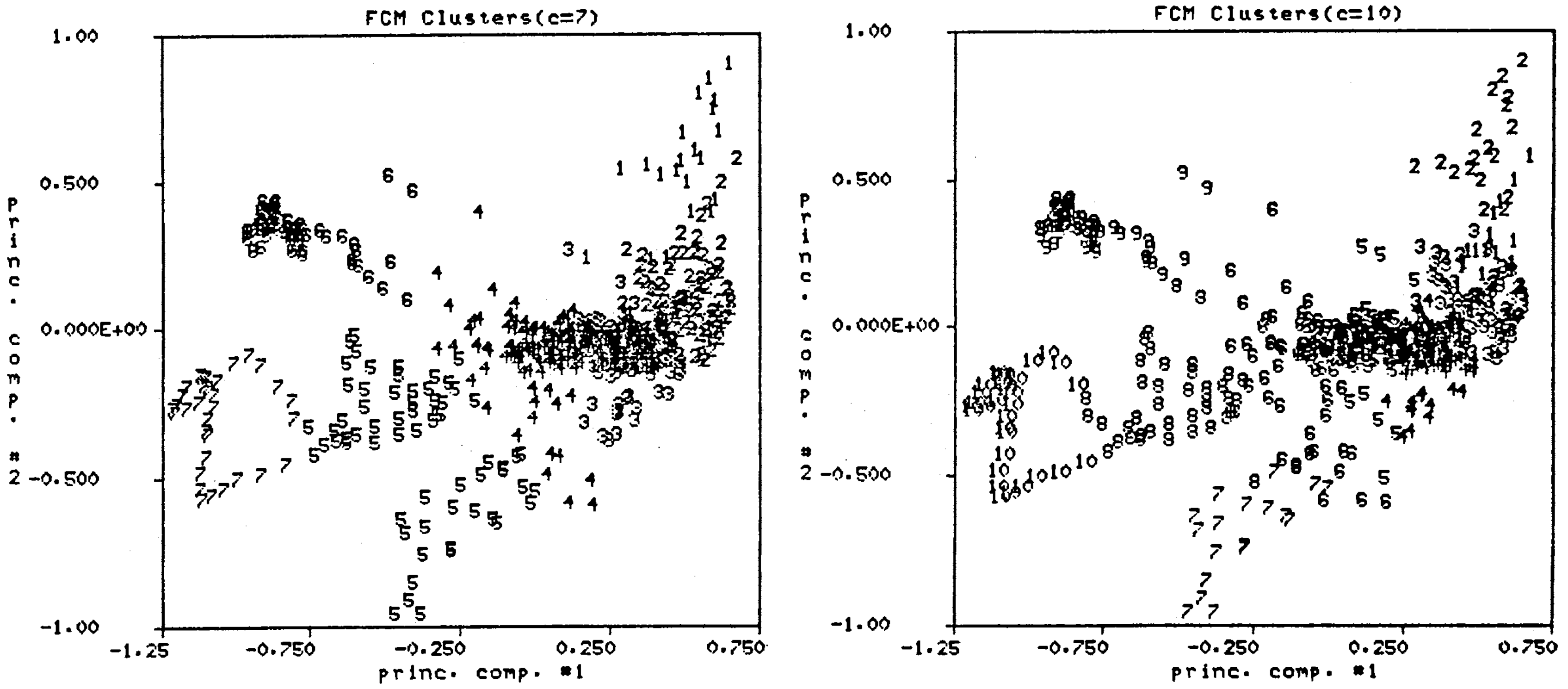


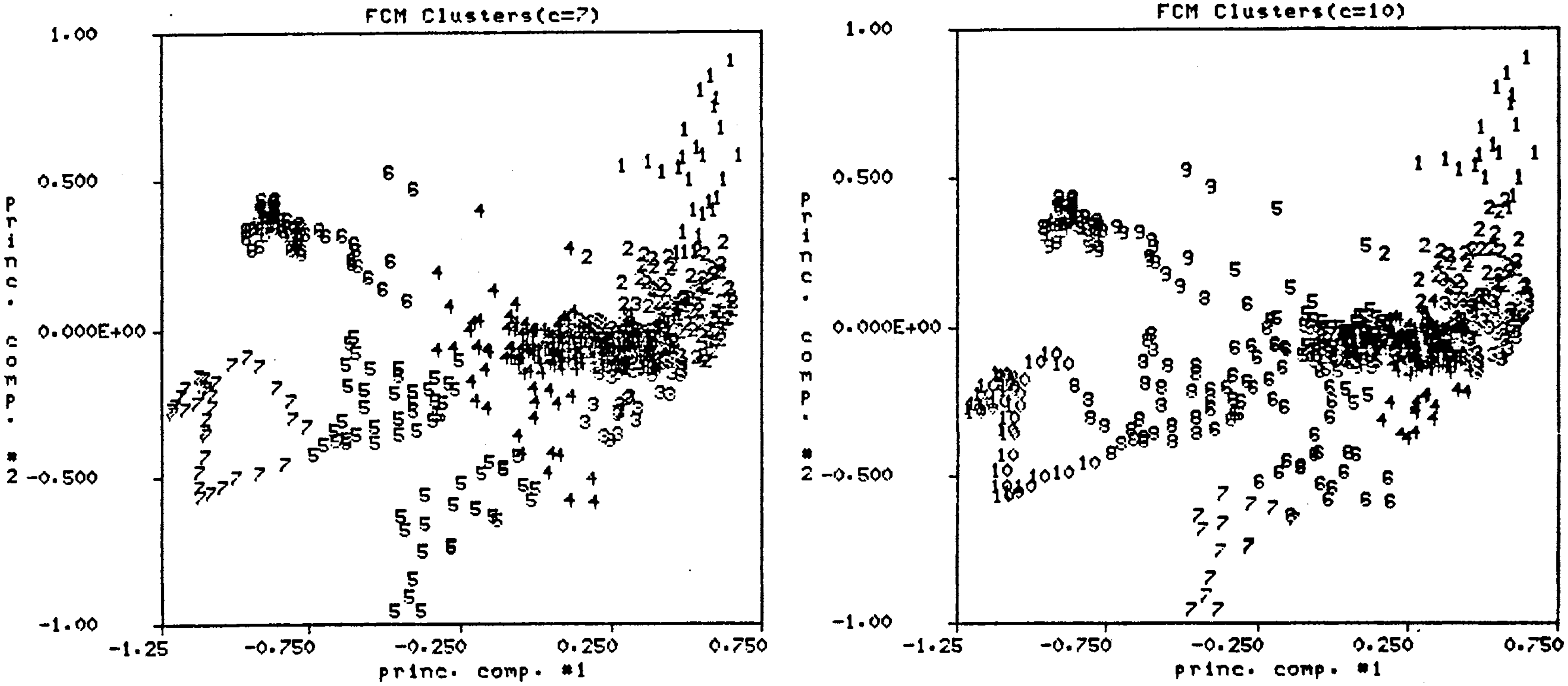
Figure 18. (continued)



(a)

(b)

Figure 19. Six Dimensional Clustering Results for (a) c=7 and (b) c=10



(a)

(b)

Figure 20. Three Dimensional Clustering Results for (a) c=7 and (b) c=10

samples plus a significant number of shale samples and cluster #1 does little, if any, better at isolating the coal samples than the two dimensional clustering results shown in Figure 16. Clusters #2 and #3 correspond primarily to some type of shale environment. Cluster #4 is very diverse and includes limestone, shaley sandstone and some relatively clean sandstone samples. Cluster #5 is also very diverse and contains limestone samples from three different limestone units as well as the majority of samples from a hydrocarbon bearing sandstone unit. Cluster #6 corresponds mainly to a relatively clean water bearing sandstone at 267-311 ft. and cluster #7 contains the majority of samples from the limestone unit at 8-28 ft. The physical meaning of the clusters in Figure 19(a) is not as straightforward as that of the clusters in Figure 16(a) due mainly to the diverse nature of clusters #4 and #5 in Figure 19(a). The FCM clustering results for the six dimensional case and c=10 are displayed in Figure 19(b). A visual inspection of Figures 19(a) and 19(b) indicates that the three additional clusters in Figure 19(b) are essentially the result of the FCM algorithm splitting clusters #2, #4 and #5 in Figure 19(a). The evaluation of the clusters in Figure 19(b) suffers from the same problem as the clusters in Figure 19(a) due to the diverse nature of clusters #6 and #8 in Figure 19(b).

It suffices to say that an explanation similar to the one for Figure 19 may also be given for the FCM clusters displayed for the three dimensional case in Figure 20. The clustering results for the three dimensional case have an obvious similarity to the six dimensional clustering results. The most notable difference is the grouping of the coal and shale samples located in the upper right and right center portions of Figure 20.

In summary, it should be pointed out that merely because good structure was not found in the six and three dimensional cases does not imply that none exists but, merely that none was found by the chosen algorithm. The fact that the best correspondence between core information and segmentation results occurs for the two dimensional case is not that surprising since PC logs #1 and #2 are primarily dependent on the SP, NPHI and RHOB logs, all of which are good lithology indicators. However, it is interesting that the ability of the FCM algorithm to find good structure in the data, as measured by the cluster validity measures, deteriorates with the inclusion of additional PC logs. Inspection of the KLT matrix in Equation (28) shows that PC log #3 is primarily a function of the resistivity and porosity logs with the largest dependence being on the SN log. Is it possible that the inclusion of PC log #3 tends to degrade the structure of the data set due to the influence of the resistivity logs? Example #2 tests this idea by excluding the resistivity logs as inputs to the segmentation algorithm.

3.2.2 Example #2

The procedure for Example #2 is identical to the procedure of Example #1 using the GR, SP, NPHI and RHOB logs shown in Figure 10 as inputs to the segmentation algorithm. The scaled input logs are transformed into PC logs according to Equation (29). Examination of the KLT matrix shows that the SP and NPHI logs are the main contributors to the first two PC logs. In this example, the FCM algorithm is applied to two different cases.

$$\begin{bmatrix} pc\ \#1 \\ pc\#2 \\ pc\#3 \\ pc\#4 \end{bmatrix} = \overset{\text{KLT Matrix}}{\begin{bmatrix} 0.206 & 0.909 & 0.361 & 0.018 \\ 0.149 & -0.319 & 0.745 & -0.567 \\ 0.521 & -0.268 & 0.341 & 0.735 \\ 0.815 & -.0002 & -0.445 & -0.371 \end{bmatrix}} \begin{bmatrix} gr \\ sp \\ nphi \\ rhob \end{bmatrix} \tag{29}$$

The first case applies the clustering algorithm to the four dimensional PC log data formed using all four PC logs and the second case applies the FCM algorithm to the two dimensional PC log data formed from PC logs #1 and #2. The dimensionality of the clustering problem is varied from four to two in an effort to determine how the detected data structure is altered by the higher numbered PC logs. Table VIII shows the FCM validity measures for the four and two dimensional clustering cases. As with Example #1, consideration of 2 as a reasonable choice for c is deferred until Example #3. The interpretation of the validity measures in Table VIII is fairly straightforward with c=4 and 8, and c=4,7 and 9 being the best choices for the four and two dimensional cases respectively. The ensuing discussion will consider c=8 for the four dimensional case and c=7 and c=9 for the two dimensional case.

Figure 21 displays the FCM clusters for the four dimensional case and c=8 along side the core description information. Even though the clustering is done in four dimensions, it is beneficial to display the clusters in the two dimensional space formed by PC logs #1 and #2. Notice how the spatial distribution of the log data in the two dimensional PC space has been altered by the elimination of the resistivity logs from the inputs to the segmentation algorithm. A comparison of Figure 21(b) with Figure 12 indicates the most notable change is with respect to the limestone samples (denoted by #3), which have a more distinctive resistivity log signature than the other lithologic groups. The circled samples in Figure 21(b) indicate the oil shows. The cluster

TABLE VIII.

FCM VALIDITY MEASURES FOR EXAMPLE #2: WITH EUCLIDEAN NORM,$\in$ = 0.01 AND m =1.5

c	F	G	ΔJ
2	0.947	−1.021	4.766
3	0.854	−2.700	5.850
4	0.858	−1.701	4.771
5	0.835	−3.594	5.114
6	0.827	−3.509	4.758
7	0.830	−3.484	4.072
8	0.830	−3.013	3.260
9	0.802	−4.705	3.365
10	0.786	−4.956	3.192

(a) Four Dimensional PC Space

c	F	G	ΔJ
2	0.952	−0.918	3.841
3	0.893	−2.141	3.749
4	0.880	−1.325	3.058
5	0.853	−1.926	3.137
6	0.854	−1.673	2.443
7	0.859	−1.947	1.946
8	0.817	−2.683	2.018
9	0.845	−1.961	1.480
10	0.816	−3.109	1.601

(b) Two Dimensional PC Space

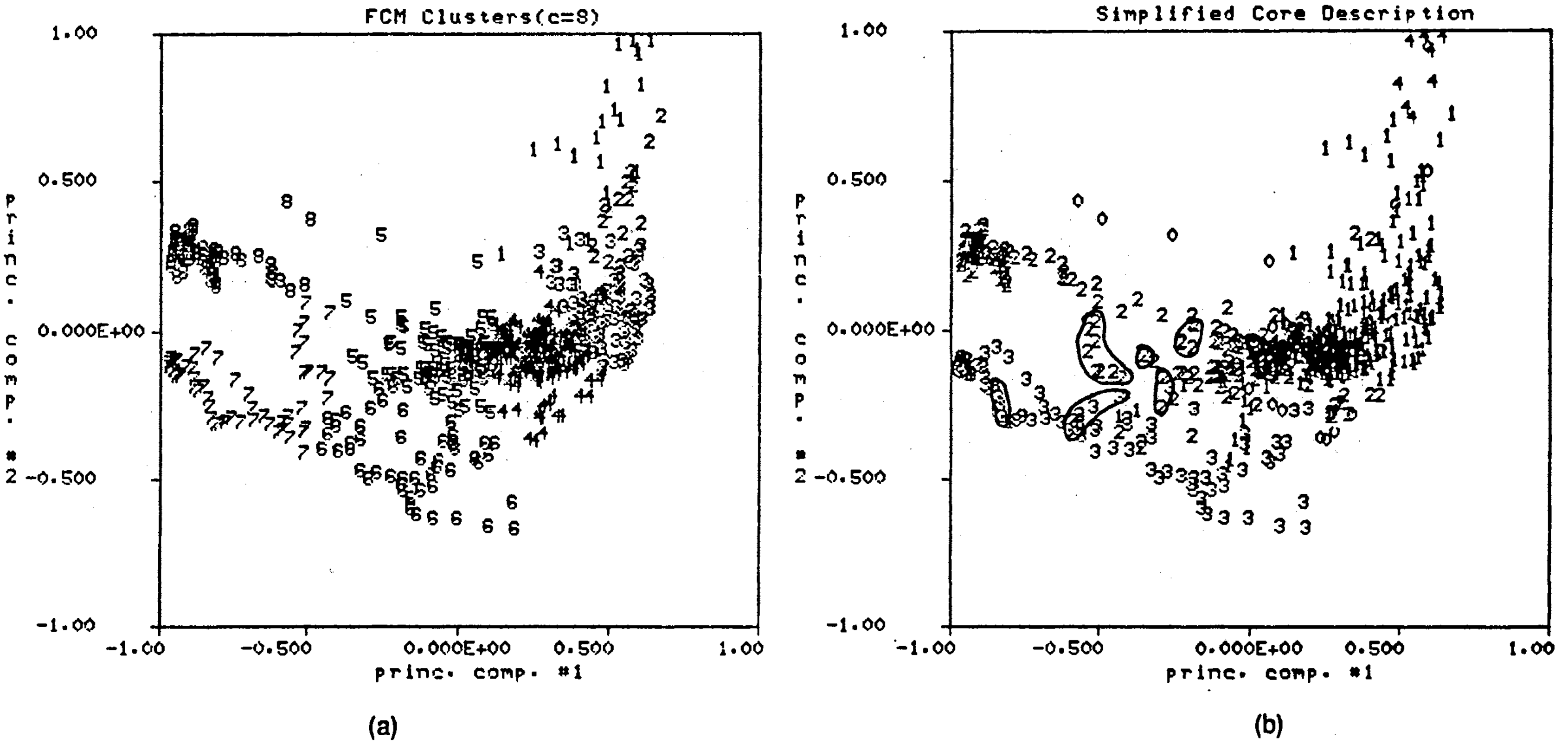


Figure 21. Comparison of (a) Four Dimensional Clustering Results (c=8) and (b) Simplified Core Description Clusters

information of Figure 21 is displayed along with the original input logs in Figure 22. Certain general observations can be made by examining the cluster information in Figure 21 and more detailed observations are possible by using the information in Figure 22. If even more detailed information is desired, then the core description information in the Appendix may be consulted. For example, Figure 21 shows that FCM cluster #1 includes all the coal samples as well as a significant number of shale samples. The segments labeled '1' in track #2 of Figure 22 (which correspond to FCM cluster #1) have a definite correspondence with the coal segments shown in track #1. The only exception to this relationship occurs between 328-332 ft. where a shale segment with very low bulk density(RHOB) is included in FCM cluster #1. From an algorithmic point of view, inclusion of these shale samples in FCM cluster #1 is reasonable. The detailed core description given in the Appendix indicates that two thin marine shales separated by a nonmarine shale occur in the interval 328-332 ft. This result for FCM cluster #1 is a slight improvement over a similar result from Example 1, displayed in Figure 17, where part of another shale segment at 5-7 ft. is also included in the FCM cluster containing all the coal samples.

In a similar fashion, the other FCM clusters for the four dimensional case may be evaluated with respect to the core description information. This evaluation process is summarized by the following description.

FCM Cluster #	Description
1	coal
2,3,4	shale
5	shaley sandstone
6,7	limestone
8	sandstone

A few observations before considering the borehole segmentation for the two dimensional case. FCM cluster #2 shown in Figure 21(a) corresponds to segments labeled '2' at 5 ft., 32 ft., and 113 ft. in Figure 22, track #2. Cluster #2 is of interest since a similar cluster does not exist for the two dimensional segmentation results. According to the core description, the segments labeled '2' correspond to shale intervals and these segments are characterized by relatively low indicated porosity on the NPHI and RHOB logs and relatively high GR readings. The point of this observation is that physically discriminating information is sometimes carried by the higher numbered PC logs. Clusters #3 and #4 correspond principally to shale environments with the members of

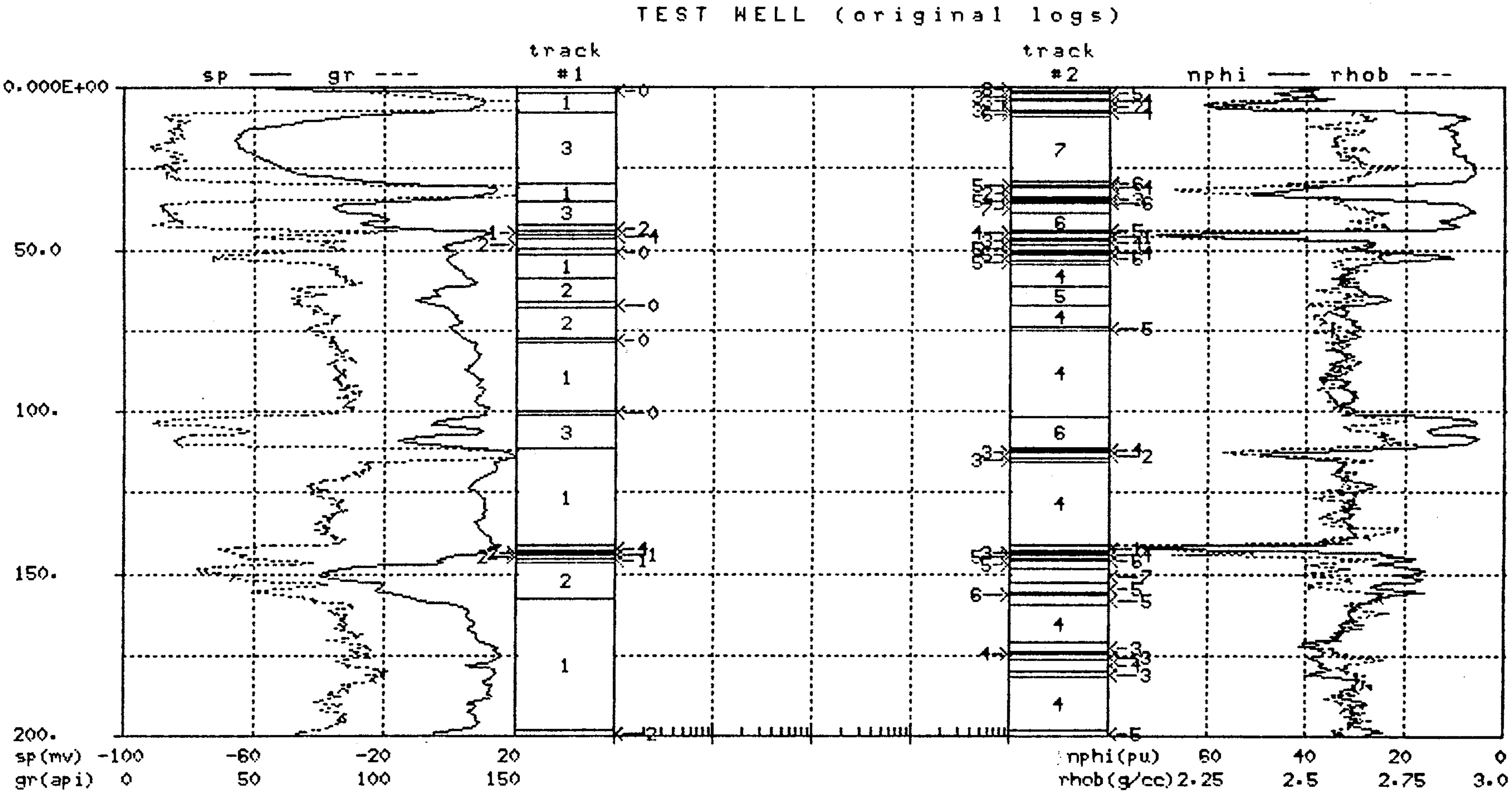


Figure 22. Comparison of Simplified Core Description Information (track #1) with the Four Dimensional Segmentation Algorithm Output, c=8 (track #2)

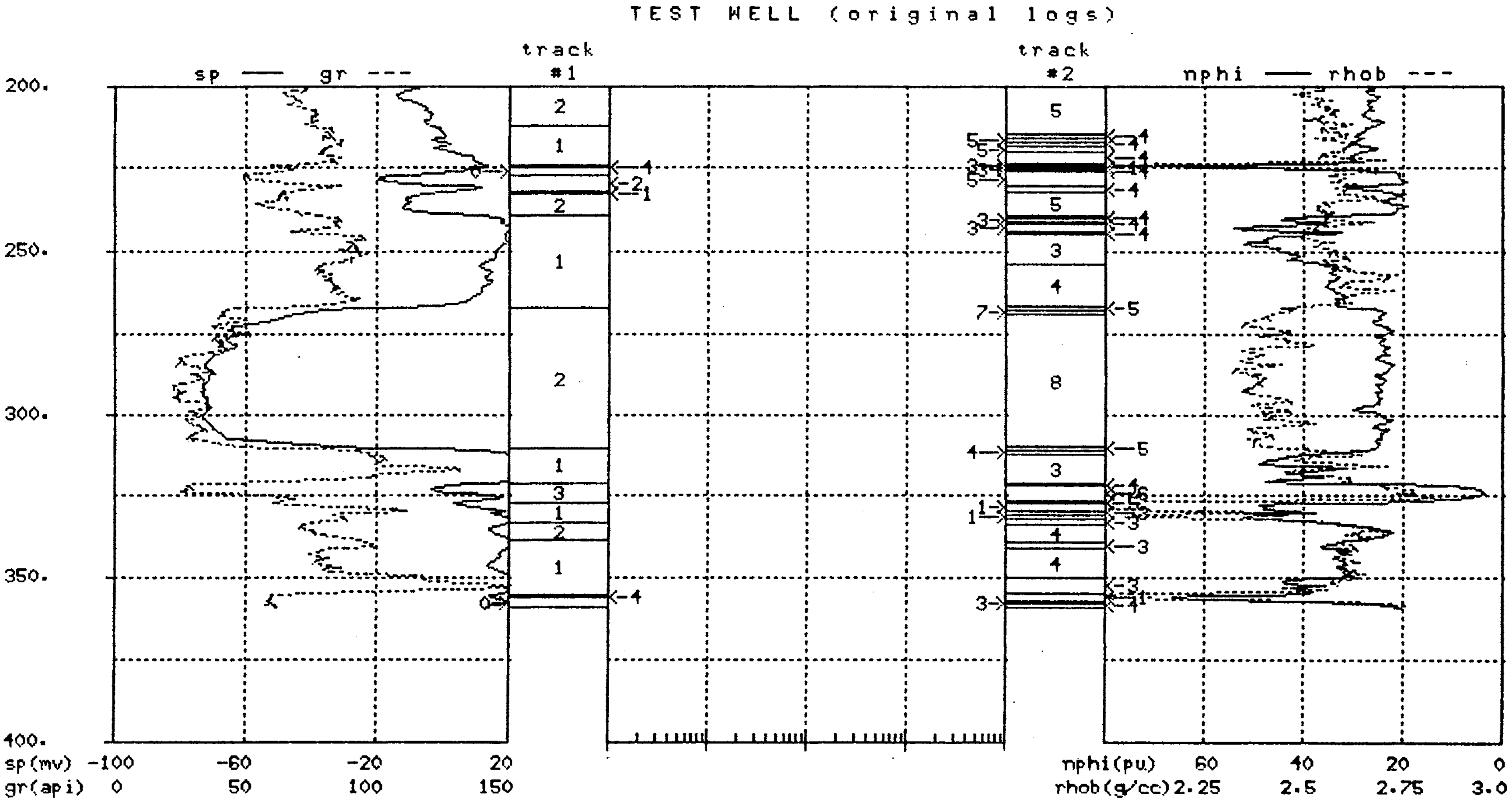


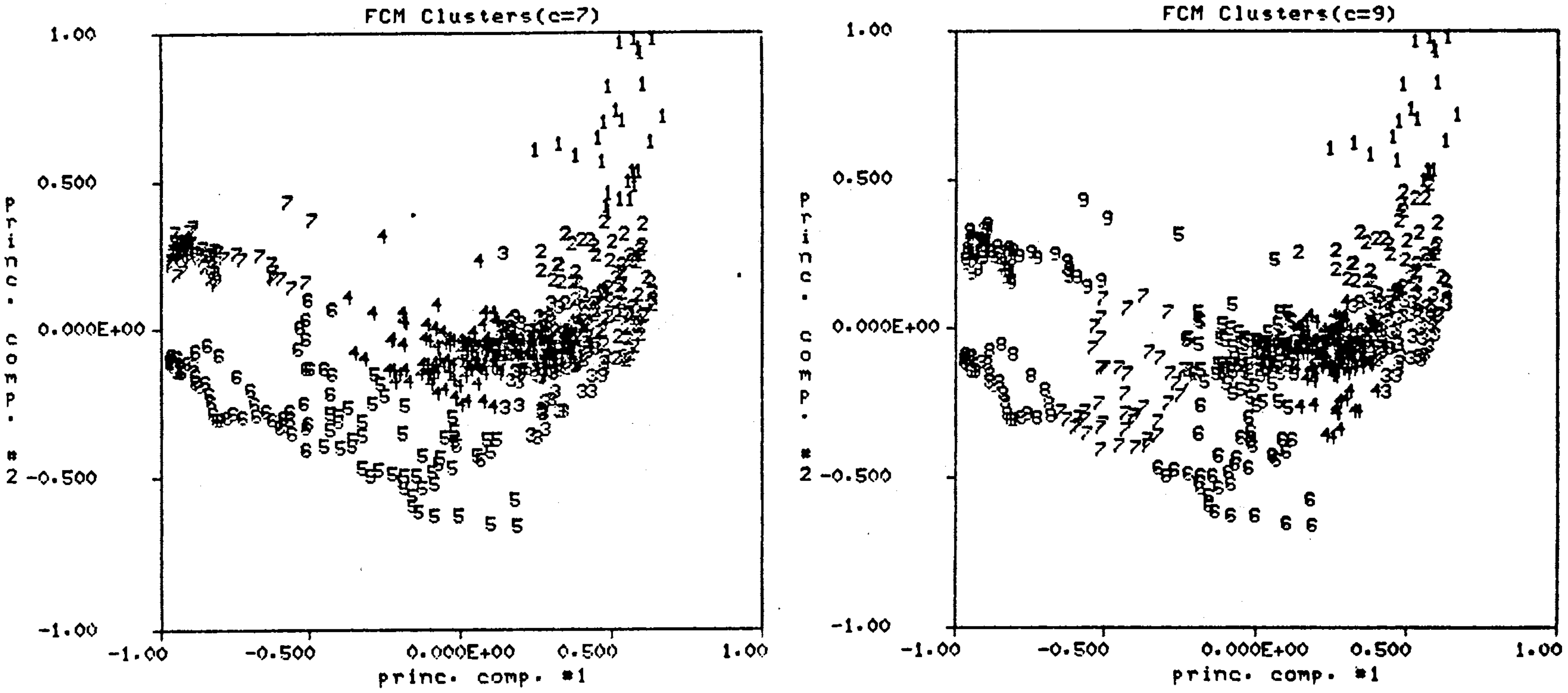
Figure 22. (continued)

cluster #4 generally being more sandy than the members of Cluster #3. There is a general correspondence between cluster #5 and a shaley sandstone environment. Clusters #6 and #7 do a good job of encompassing the respective limestone samples but also includes part of the hydrocarbon bearing sandstone unit at approximately 150 ft. Finally, cluster #8 corresponds to the well developed sandstone unit between 267-311 ft. The borehole segments determined for the four dimensional case have a reasonable physical interpretation

Example 2 continues by considering the borehole segmentation for the two dimensional case using PC logs #1 and #2. Figures 23(a) and 23(b) show the FCM clusters for the two dimensional case when c=7 and c=9 respectively. Visual inspection of the FCM clusters in Figure 23(a) and Figure 21(a) reveals an obvious similarity between the respective clusters. The most obvious difference involves the structure of the clusters in the upper right hand portion of the two figures. If in Figure 21(a), the members of cluster #2 are divided between clusters #1 and #3 then the resulting data structure would be nearly identical to the data structure shown in Figure 23(a). To facilitate the comparison between the respective FCM clusters the following definitions are made.

	Figure 21(a)	Figure 23(a)	
New Cluster #	Cluster #	Cluster #	Description
1	1	1	coal
2	2,3,4	2,3	shale
3	5	4	shaley sandstone
4	6,7	5,6	limestone
5	8	7	sandstone

The new cluster numbers shown above are used in Figure 24 to compare similar borehole segments for the four dimensional case with c=8 (track #1) and the two dimensional case with c=7 (track #2). The borehole segmentation shown in Figure 24 is nearly identical for the two cases with the exception of segment #1. Segment #1 in track #1 of Figure 24 includes fewer shale samples than segment #1 in track #2 and this difference gives the four dimensional segmentation results a slightly more consistent physical interpretation than the two dimensional results. The largest single inconsistency in the physical interpretation of the segmentation results shown in Figure 24 occurs for the hydrocarbon bearing sandstone unit at approximately 150 ft. In both cases, this sandstone unit is grouped with the limestone samples. From an algorithmic



(a)

(b)

Figure 23. FCM Clusters for the Two Dimensional Case with (a) c=7 and (b) c=9

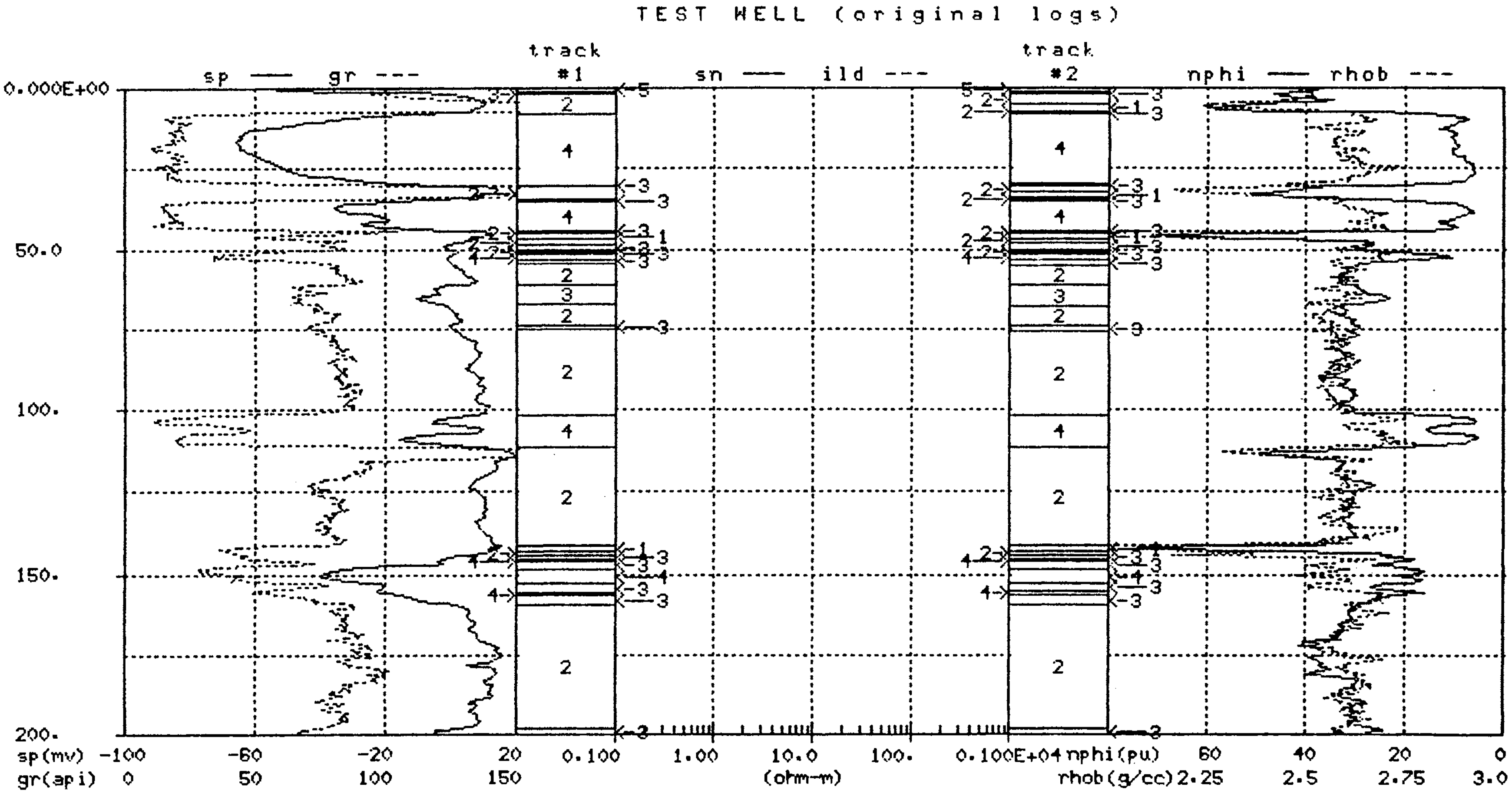


Figure 24. Comparison of the Segmentation Algorithm Output for the Four Dimensional Case, c=8 (track#1) and the Two Dimensional Case, c=7 (track #2)

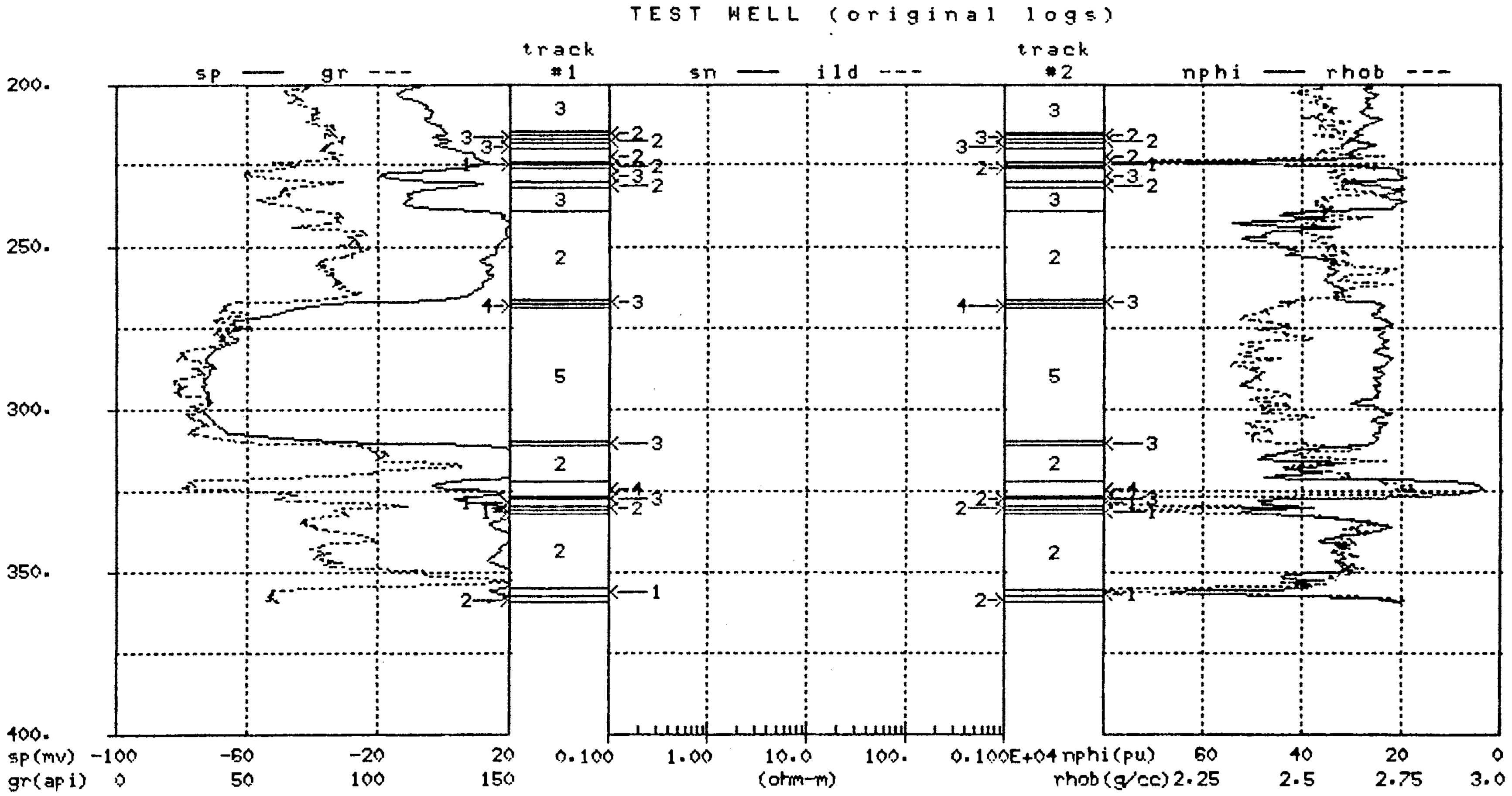


Figure 24. (continued)

point of view, there is an element of consistency in the Figure 24 segmentation results which was lacking in Example 1.

Example 2 continues by considering the segmentation algorithm output for the two dimensional case and c=9 (Figure 23(b)). In a manner similar to that used for the comparison shown in Figure 24, the borehole segments derived for the two dimensional case and c=9 are contrasted to the segmentation results of Example 1 shown in track #2 of Figure 18. The following descriptions are consistent with those used in Example 1.

New Cluster #	Figure 23(b) Cluster #	Description
1	1	coal/shale
2	2,3,4	shale
3	5	shaley sandstone
4	6,8	limestone
5	7	limestone/sandstone/oil
6	9	sandstone

Tracks #1 and #2 of Figure 25 display the borehole segmentation results from Example 1 and Example 2 respectively. Both results were obtained from the two dimensional case and c=9. In both cases, the detected data structure is very similar and lends itself to a reasonably consistent interpretation. The main inconsistencies for the segmentation results shown in Figure 25 involve segments labeled '1' and '5'. Segments labeled '1' include both coal and shale intervals and segments labeled '5' include hydrocarbon bearing limestone and sandstone intervals. Even though the segmentation results shown in tracks #1 and #2 of Figure 25 are very similar, the results from Example 1 are preferred over the borehole segmentation shown in track #2 due to the 'cleaner' transitions from one segment to the next segment. A good example of this 'cleaner' transition is the shale-sandstone-shale transition that occurs between 250-320 ft. in Figure 25.

All the results displayed in Examples 1 and 2 have used a value of m=1.5 for the weighting exponent in the FCM clustering algorithm. Example #2 concludes with a brief illustration of why m=1.5 is preferred over values such as m=1.25 or m=2.0 for the given test data.

For illustration purposes, let's consider the four dimensional clustering results shown in Figure 21(a). The FCM clusters shown in Figures 26(a) and 26(b) were

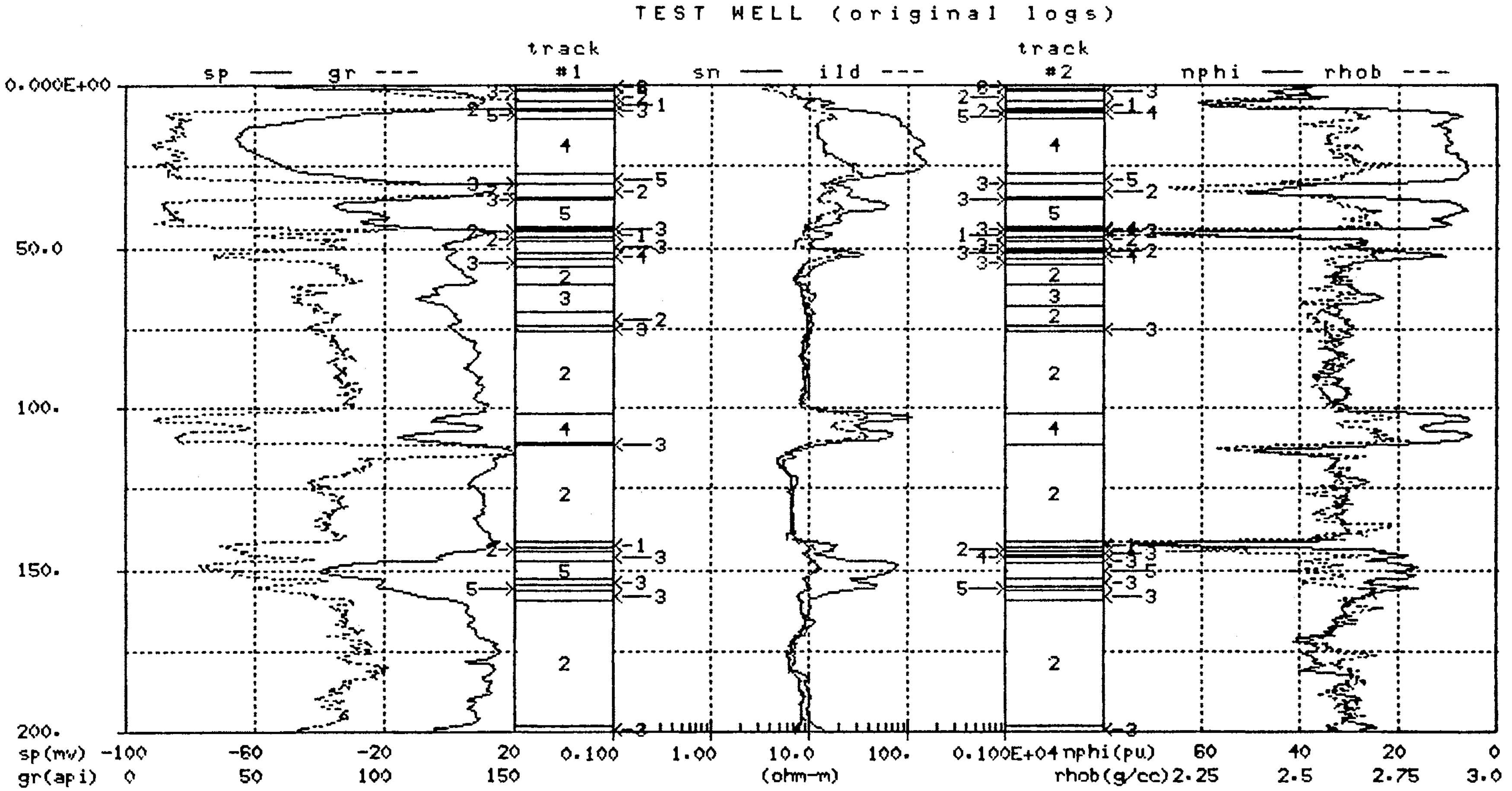


Figure 25. Comparison of the Segmentation Algorithm Output for the Two Dimensional Case, c=9 from Example 1 (track#1) and the Two Dimensional Case, c=9 from Example 2 (track #2)

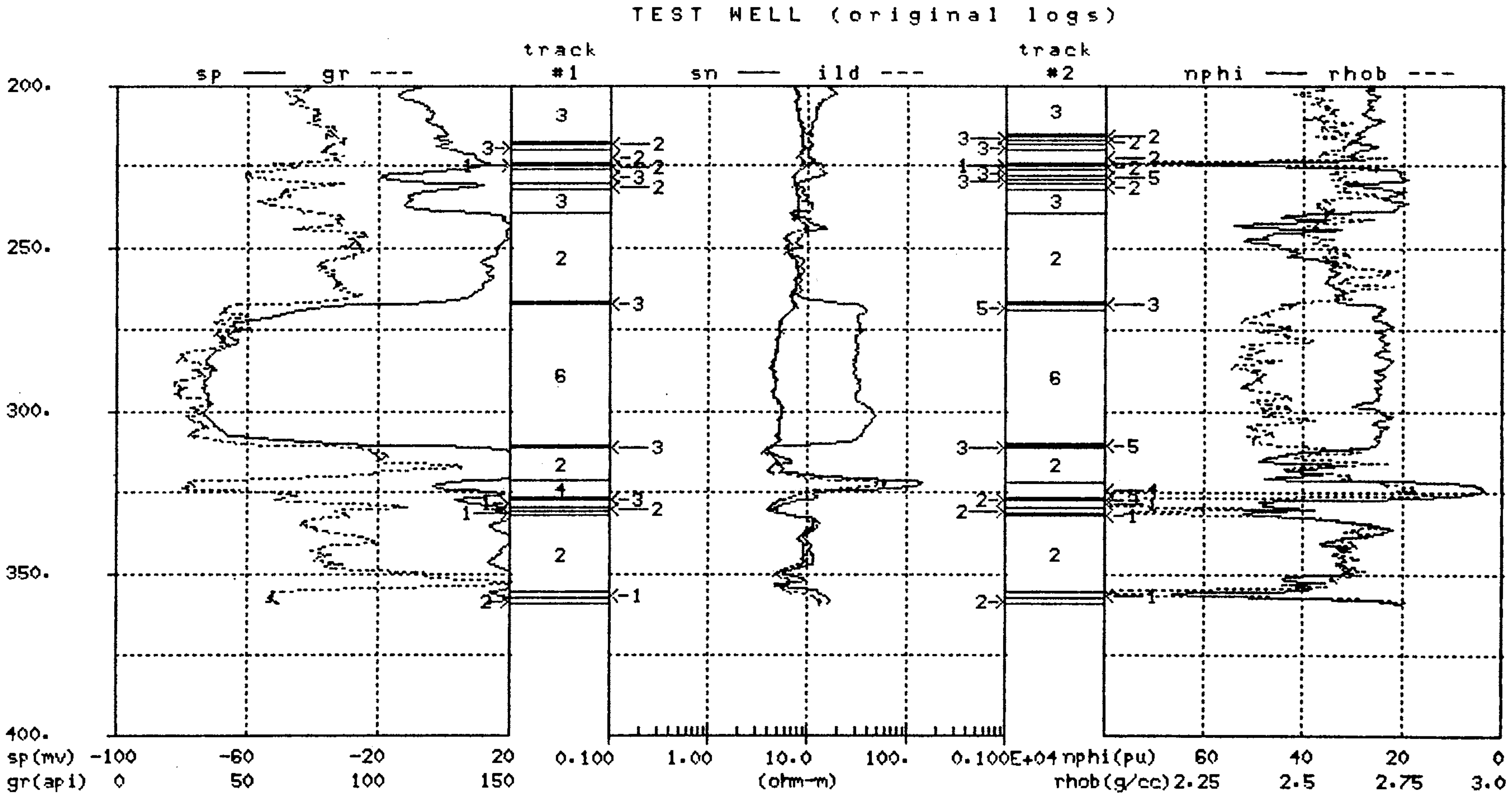
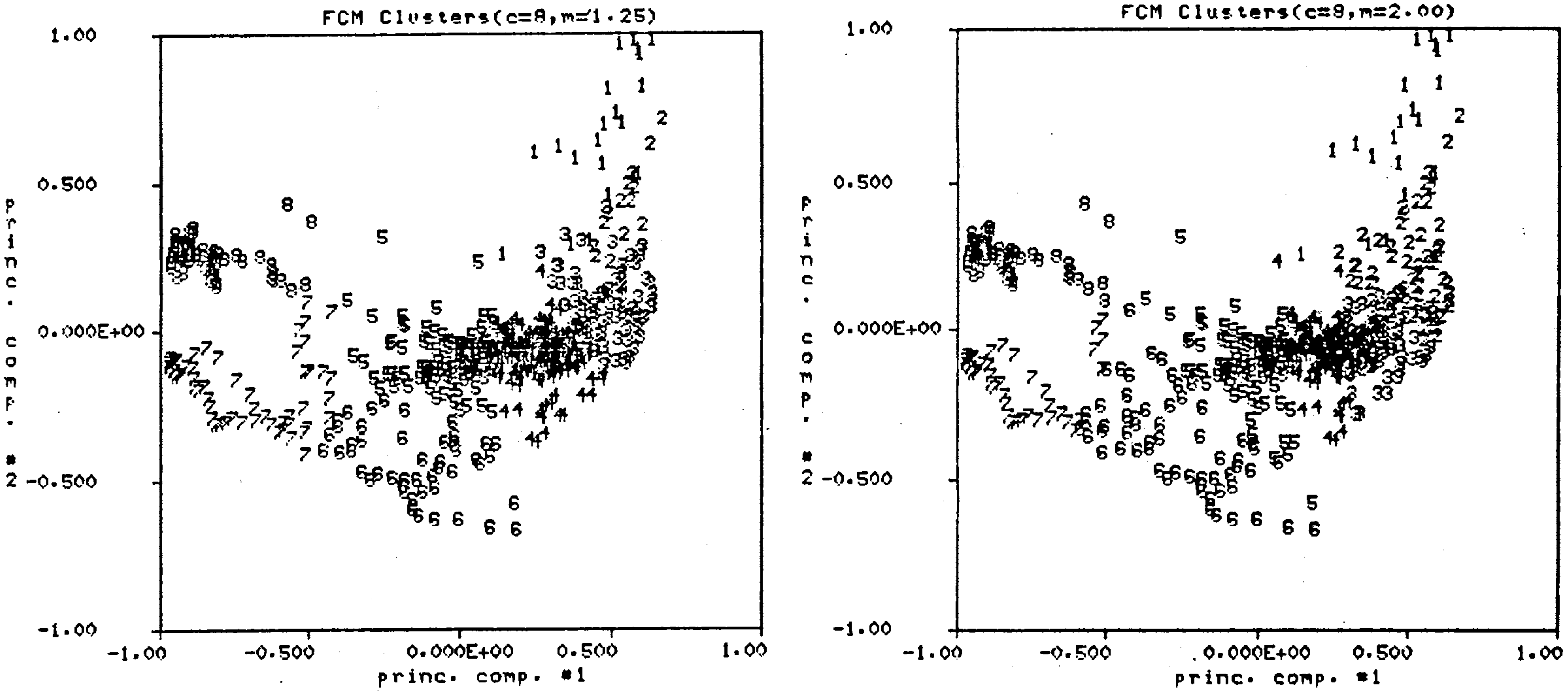


Figure 25. (continued)



(a)

(b)

Figure 26. Four Dimensional Clustering Results with c=8 and (a) m=1.25 and (b) m=2.00

obtained using exactly the same FCM parameters as the FCM clusters shown in Figure 21(a) except m=1.25 for Figure 26(a) and m=2.00 for Figure 26(b). The guideline used for choosing an appropriate value for the weighting exponent, m, was stated in Section 2.1.3. It is desirable to pick m large enough so that if the resulting FCM solution is relatively 'hard' then this is a good indication of substructure in the data. Choosing m too small results in solutions which are artificially hard, while choosing m too large essentially insures that the FCM solution will not have good structure as measured by the cluster validity measures. With this thought in mind, let's examine the the FCM clusters in Figures 26(a), 21(a) and 26(b) which represent weighting exponent values of 1.25, 1.50 and 2.00 respectively. A visual inspection of Figures 26(a) and 21(a) indicates that the maximum membership clusters for m=1.25 and m=1.5 appear to be identical. However, a similar comparison between Figures 21(a) and 26(a) shows a significant difference between the maximum membership clusters for m=1.5 and m=2.0. The samples comprising FCM cluster #2 in Figure 21(a) do not form a separate and distinct group for the case of m=2.00 in Figure 26(b). For larger m values there is a tendency for clusters with small populations to be grouped with other samples to form larger clusters. This is viewed as an undesirable tendency since important geological information might be obscured by this 'lumping' of clusters. Therefore, m=1.5 is preferred over m=1.25 or m=2.0 since a value of 1.5 seems to strike a balance between solutions which might appear artificially 'hard' in nature and solutions which can obscure potentially valuable information contained in relatively small populations.

In summary of Example 2, there exists a reasonable physical interpretation for the segmentation results in both the four and two dimensional cases. This type of consistent physical interpretation was lacking in Example 1 where only the two dimensional results yielded a reasonable physical interpretation. This fact supports the idea that undue emphasis on the resistivity logs tends to degrade the ability of the segmentation algorithm to detect data structure with a reasonable physical interpretation. There is a definite relationship between the four dimensional segmentation results for c=8 and the two dimensional segmentation results for c=7 with the four dimensional results yielding a slightly more consistent physical interpretation. Additionally, there is a good correspondence between the two dimensional results for c=9 in Example 1 and the two dimensional results for c=9 in Example 2. In this comparison, the Example 1 segmentation results are preferred because of the 'cleaner' transitions from one segment to the next segment. Example 2 concluded with an illustration motivating the use of m=1.5 as a nominal value for the weighting exponent in the FCM clustering algorithm.

3.2.3 Example #3

This example applies a sequential clustering strategy to the two dimensional principal components data of Example 1. The sequential clustering strategy takes advantage of the fact that the validity indicators, F and G, in Table VII indicated 2 as a good choice for c in all three cases and specifically for the two dimensional case in Table VII(c). A similar statement can be made for Example 2 but only the data from Example 1 will be used to illustrate the sequential clustering strategy.

Figure 27 shows the maximum membership clusters for the two dimensional PC log data of Example 1 and c=2. The sequential approach applies the FCM algorithm to each of the two clusters shown in Figure 27. All FCM parameters are identical to those used in Example 1. Table IX(a) shows the FCM validity measures when the FCM algorithm is applied to cluster #1 in Figure 27. When c=5, F and G are maximum and ΔJ is near its minimum value. Clearly, 5 seems to be the best choice for c and the corresponding FCM maximum membership clusters are shown in Figure 28(b). Similarly, the FCM algorithm is applied to cluster #2 in Figure 27. Table IX(b) displays the validity measures as c is varied from 2 to 8. A value of 5 is judged to be the best choice for c. At c=5, F is near its maximum, G is maximum and ΔJ is near its minimum. Figure 28(a) displays the FCM clusters for c=5.

At this point a decision is made not to further subdivide any of the clusters shown in Figures 28(a) or 28(b). This decision is based on the relatively good agreement among the validity measures and the fact that the performance of the FCM algorithm deteriorates for small populations. The FCM algorithm does continue to converge properly for small populations but, the performance deteriorates in the sense that there tends to be no clear interpretation for the cluster validity measures.

The Figure 28 clusters are renumbered and merged into a single display shown in Figure 29. Although it is not necessary, cluster #1 in Figure 28(a) is lumped with cluster #5 in Figure 28(b) to form cluster #5 in Figure 29. Such manual lumping of clusters is undesirable but this is done, in part, to allow an equitable comparison of the sequential clustering results with the results of Example 1. There is an obvious similarity between the Figure 29 clusters and the Example 1 clusters shown in Figure 16. A more detailed comparison of results is possible in Figure 30 where the Example 1 segments are displayed in track #1 and the Example 3 segments are shown in track #2. The description used for the clusters in Example 1 is duplicated here. In this specific case, the segmentation results using different clustering strategies are essentially the

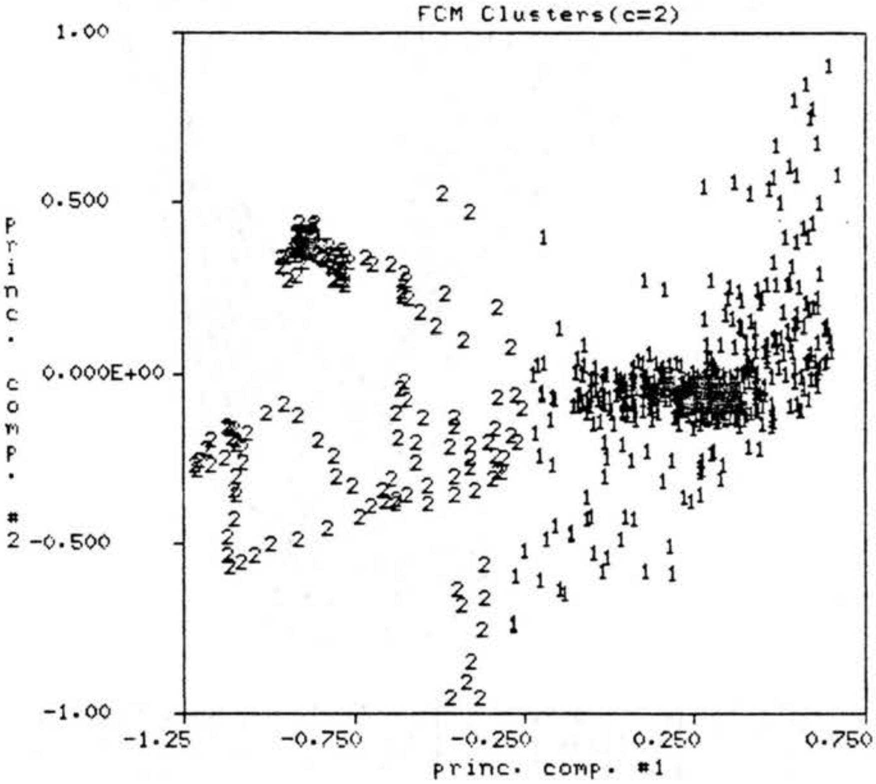


Figure 27. Maximum Membership Clusters for the Two Dimensional Case and c=2 from Example 1

TABLE IX.

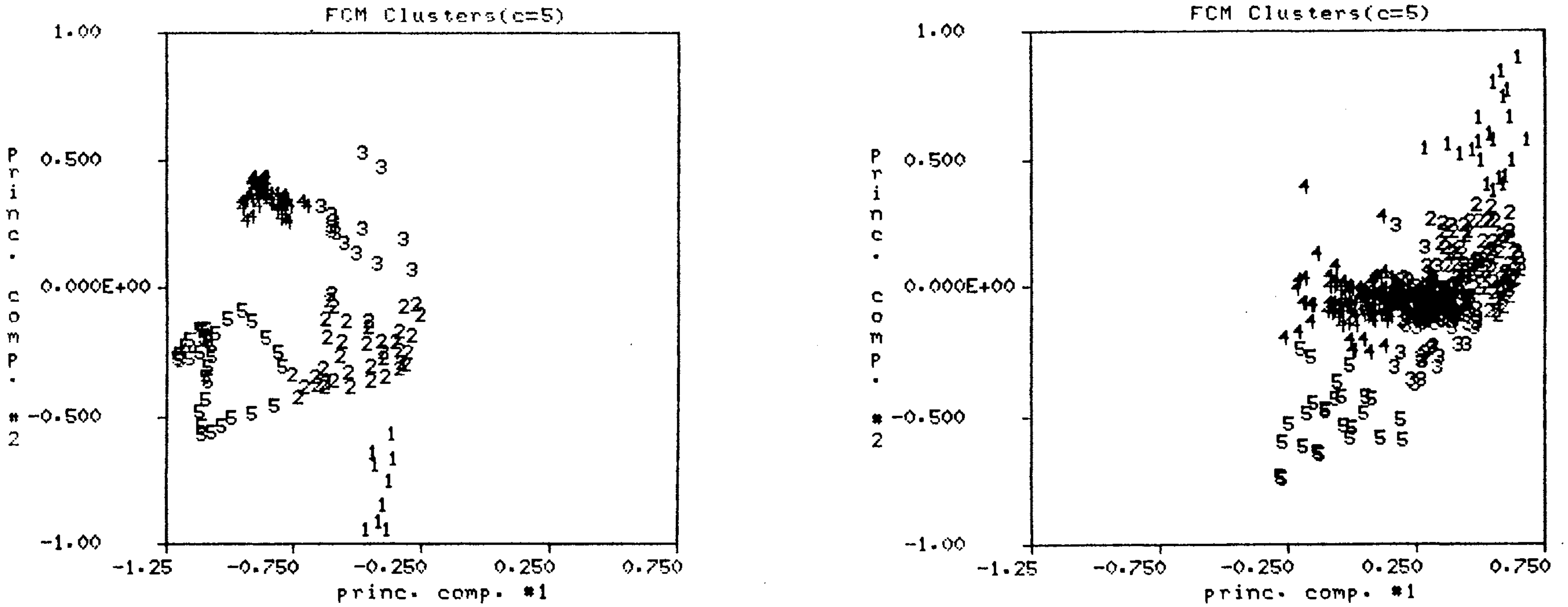
FCM VALIDITY MEASURES FOR EXAMPLE #3,
SEQUENTIAL CLUSTERING STRATEGY FOR
CLUSTER #1 AND CLUSTER #2
IN FIGURE 27

c	F	G	ΔJ
2	0.869	–3.233	1.905
3	0.868	–2.260	1.786
4	0.855	–2.210	1.497
5	0.871	–2.137	0.912
6	0.816	–2.927	0.994
7	0.812	–2.984	0.919
8	0.790	–2.816	0.811

(a) Cluster #1

c	F	G	ΔJ
2	0.912	–0.883	0.891
3	0.932	–0.692	0.596
4	0.934	–0.301	0.330
5	0.927	–0.213	0.225
6	0.916	–0.710	0.249
7	0.917	–0.765	0.220
8	0.895	–1.069	0.229

(b) Cluster #2



(a)

(b)

Figure 28. Maximum Membership Clusters Corresponding to (a) Cluster #2 and (b) Cluster #1 in Figure 27

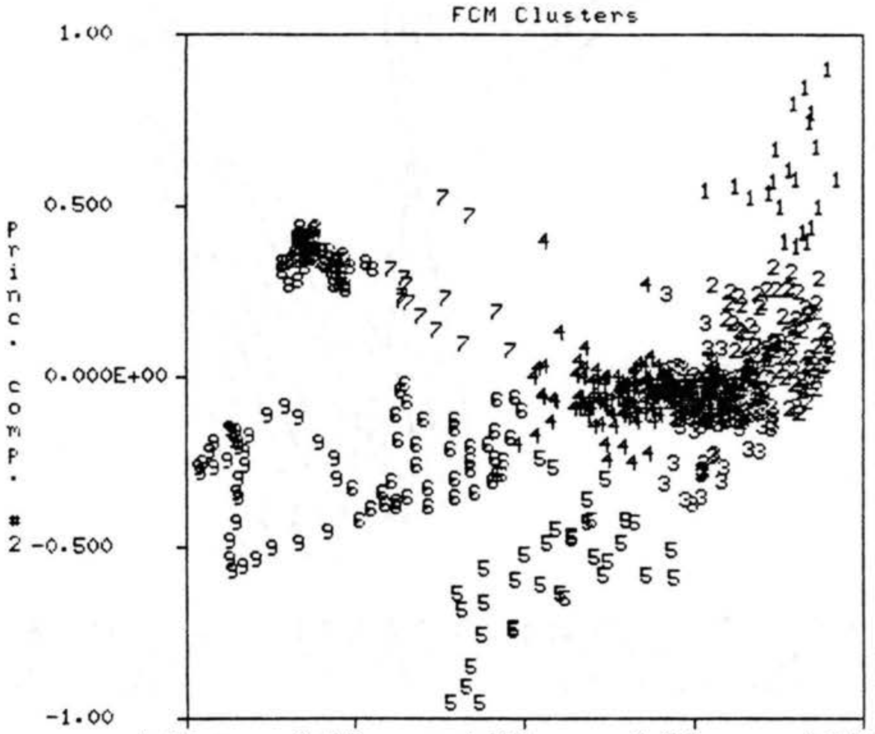


Figure 29. Merged Display of Clusters in Figure 28

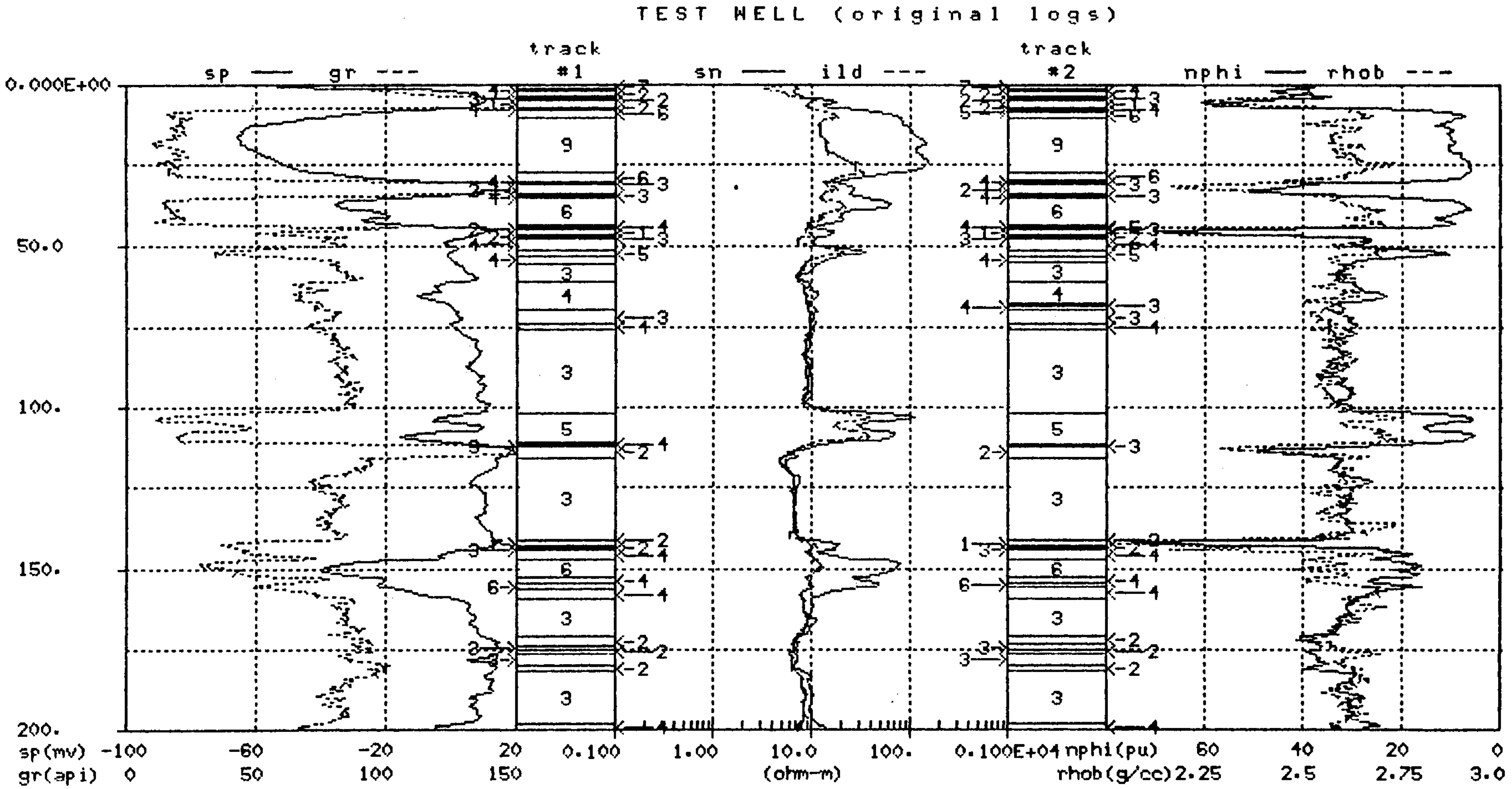


Figure 30. Comparison of Segmentation Results for: Example 1, Two Dimensionial Case, c=9 (track#1) and Example 3, Sequential Clustering Strategy (track #2)

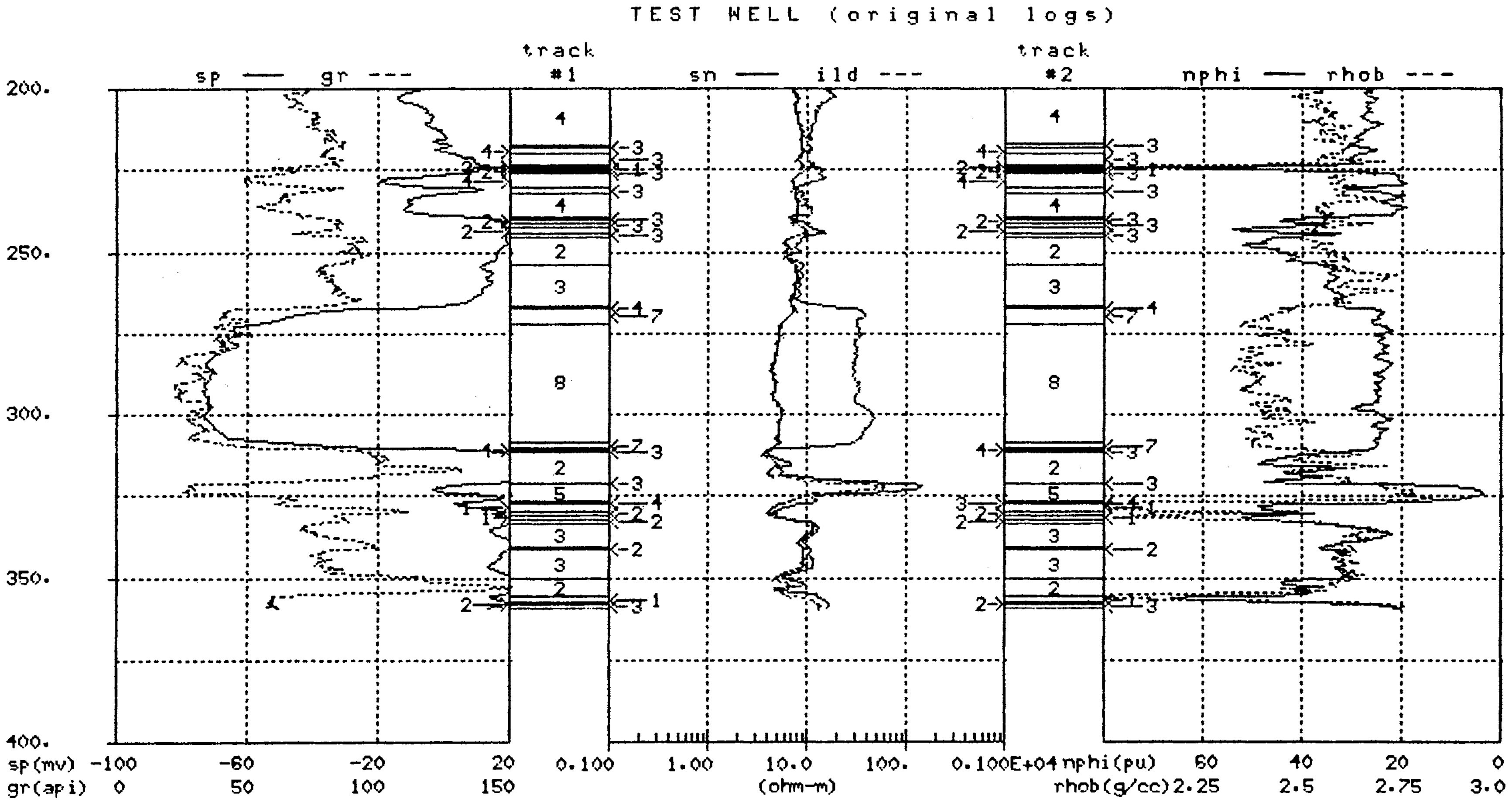


Figure 30. (continued)

FCM Cluster #	Description
1	coal/shale
2,3	shale
4	shaley sandstone
5,9	limestone
6	limestone/sandstone/oil shows
7,8	sandstone

same. This observation supports the contention that the discovered data structure occurs naturally, and the structure is not imposed by a particular process. Similar results were obtained when the sequential clustering strategy was used on the four dimensional PC log data of Example 2.

It should be noted that a variation of the sequential clustering strategy was applied to the log data of the test interval. The sequential clustering strategy as it has been applied consists essentially of looping back to step 6 from step 7 of the basic methodology given in Section 3.0. The variation consists of looping back to step 3 from step 7. For example, the original log values for cluster #1 in Figure 27 were scaled and transformed using the KLT to generate a new set of PC log values specific to cluster #1. This new PC log data was clustered and evaluated using the cluster validity measures. Of course this same process was applied to cluster #2. This variation of the sequential strategy yielded a much poorer physical interpretation for the resulting borehole segments.

3.2.4 Example #4

The last example of this section uses the six logs in Figure 10 as inputs to the segmentation algorithm. These same inputs were used in Example 1. Step 3 of the basic methodology used in Example 1 is modified to take the common logarithm of the resistivity logs prior to the linear scaling process. Taking the common logarithm of the resistivity logs lessens the influence that very large resistivity values have on the linear scaling process. After taking the logarithm of the resistivity logs, all six logs are scaled using the same linear scaling process of Example 1 and PC logs are calculated according to Equation (30). One result of this modified scaling process is to increase the

$$\begin{bmatrix} pc\#1 \\ pc\#2 \\ pc\#3 \\ pc\#4 \\ pc\#5 \\ pc\#6 \end{bmatrix} = \overset{\substack{\text{KLT} \\ \underline{\text{Matrix}}}}{\begin{bmatrix} 0.174 & 0.718 & -0.578 & -0.120 & 0.325 & 0.003 \\ 0.036 & -0.522 & -0.303 & -0.646 & 0.362 & -0.294 \\ 0.142 & 0.115 & 0.468 & 0.284 & 0.611 & -0.541 \\ 0.585 & -0.141 & 0.218 & -0.067 & 0.354 & 0.678 \\ 0.772 & -0.103 & -0.173 & 0.140 & -0.438 & -0.390 \\ 0.096 & 0.409 & 0.527 & -0.681 & -0.270 & -0.094 \end{bmatrix}} \begin{bmatrix} gr \\ sp \\ sn \\ ild \\ nphi \\ rhob \end{bmatrix} \tag{30}$$

weight which the resistivity logs have in the calculation of the principal components logs. This is evidenced by a comparison of the coefficients in the KLT matrices of Equations (28) and (30) for the resistivity values. For the first principal component, this comparison shows an increase in magnitude from 0.321 to 0.578 for the SN coefficient and an increase in magnitude from 0.057 to 0.120 for the ILD coefficient. The remaining coefficients for the first principal component show a decrease in magnitude. A similar comparison for the second principal component shows a significant increase in the magnitude of the ILD coefficient and a significant decrease for the NPHI coefficient. The PC logs for Example 4 are shown in Figure 31 along with the simplified core description information in tracks #1 and #2. One very noticeable difference between the PC logs for Example 4 and those of Example 1(Figure 11) is the prominent peaks in PC log #3 in Figure 31 which correspond to the coal intervals within the borehole. Similar peaks occur in PC log #2 in Example 1. A result of the increased weighting of the resistivity logs is poorer spatial separation of the coal samples from neighboring samples as displayed in Figure 32. Figures 32(a) and 32(b) display a crossplot of the core information using PC logs #1 and #2 in Figure 3(a) and PC logs #1 and #3 in Figure 32(b). Recall the definitions used in Figure 32 are: 0-undefined, 1-shale, 2-sandstone, 3-limestone and 4-coal. Notice that the coal samples are completely obscured in Figure 32(a) but do separate out in Figure 32(b). In Example 1, the separation of the respective lithologic groups was evident using only PC logs #1 and #2.

Even though there is little evidence to suggest that the segmentation results using the PC log data of Example 4 would have a better physical interpretation than results of previous examples, the segmentation algorithm is applied to the six and three dimensional PC log data of Example 4. All FCM parameters are set identically to those in

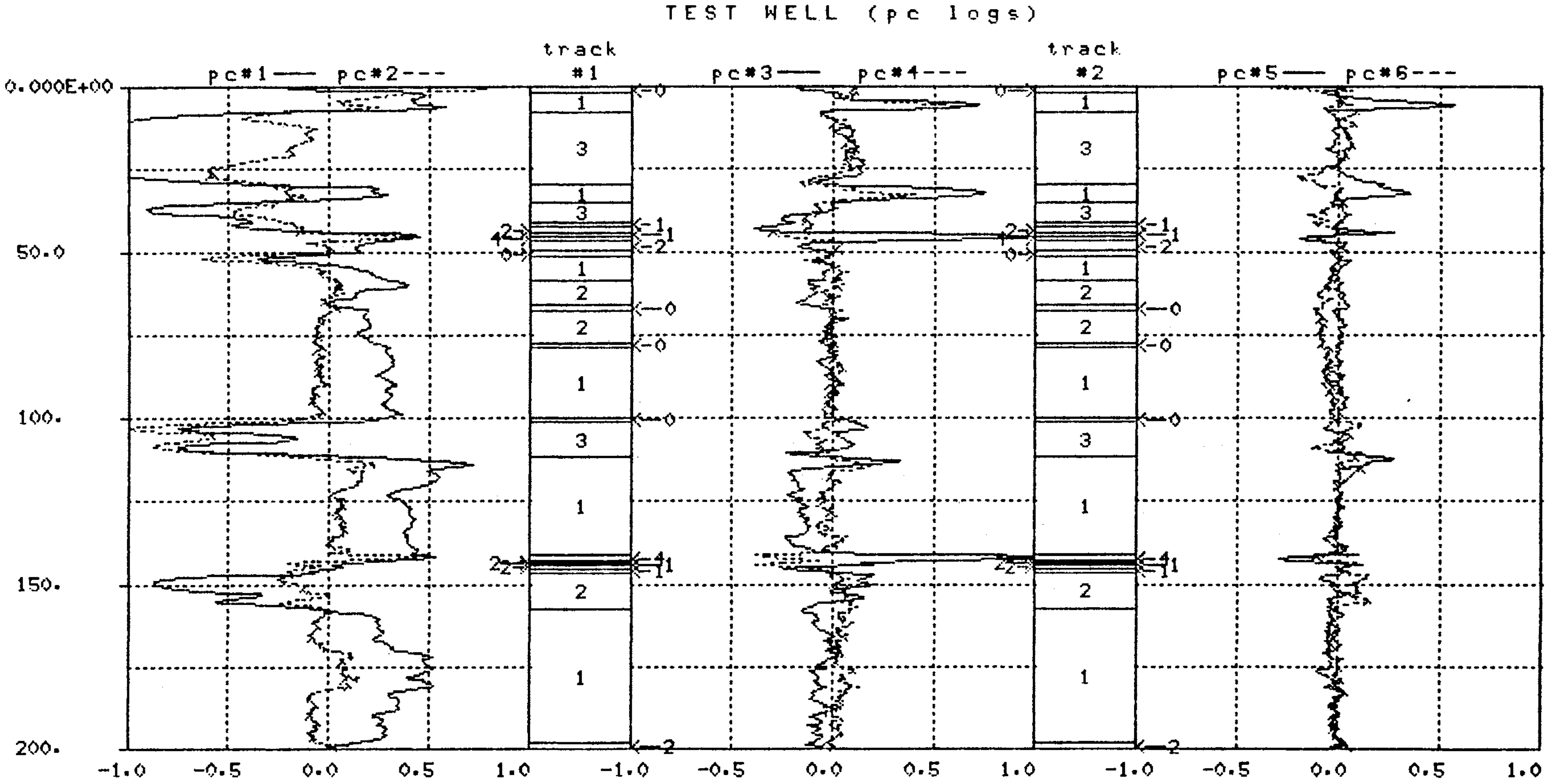


Figure 31. PC logs for Example 4 with Simplified Core Description(tracks #1 and #2)

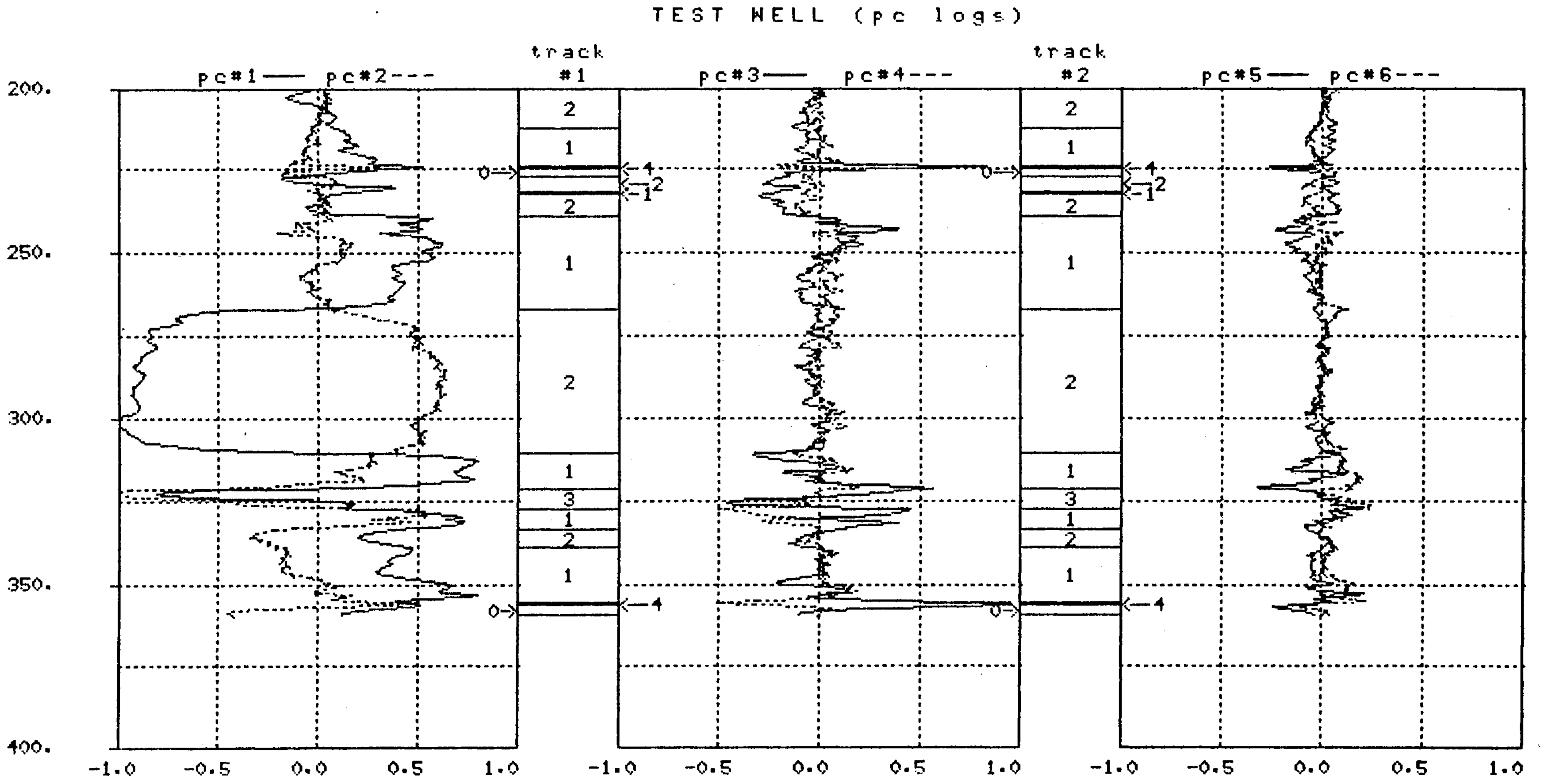
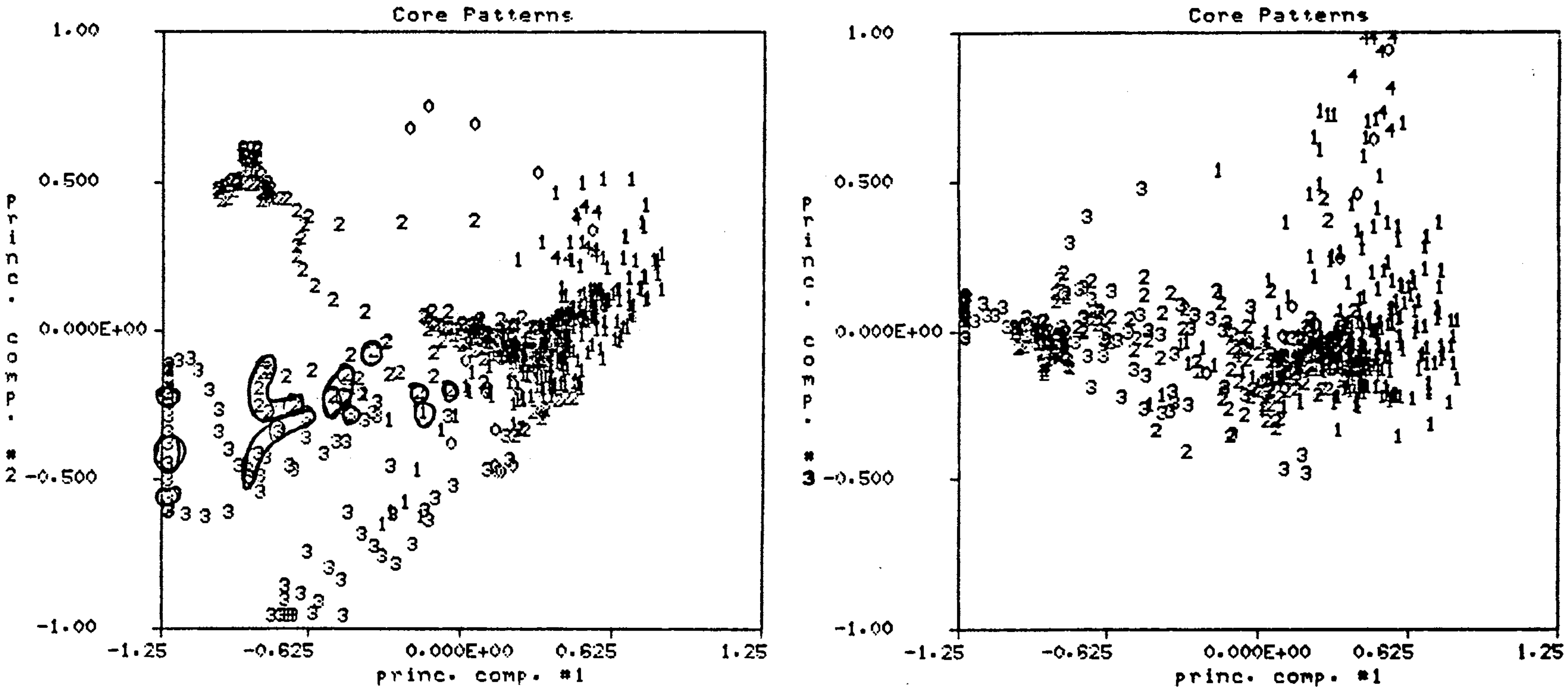


Figure 31. (continued)



(a)

(b)

Figure 32. Crossplot of Simplified Core Information with Oil Shows Circled
(a) PC logs #1 and #2 (b) PC logs #1 and #3

Example 1. The two dimensional case was not considered since Figure 32(a) suggests that the coal samples would be indistinguishable from the surrounding shale samples.

Tables X(a) and X(b) show the cluster validity measures for the six and three dimensional clustering results respectively. There is a basic problem interpreting the validity measures in Table X in the fact that there is no clear consensus among F, G and ΔJ. F is maximum for c=2 in both cases, G is maximum for c=3 and ΔJ is minimum for c=6 in the six dimensional case and for c=8 in the three dimensional case. A few general observations can be made to help in the interpretation process. First, let's restrict the interpretation of the validity measures to the three dimensional case in Table X(b). The gross structure of the data is carried by the first few principal components and if the detected data structure for the three dimensional case does not have a reasonable physical interpretation then it is doubtful that increasing the dimension of the clustering problem will improve the segmentation results. Second, F can be used to reduce the possible candidates for best c. In Table X(b), the F values for c=2 and 3 are of comparable magnitude, the F values for c=4, 5 and 6 are also of comparable magnitude as are the F values for c=7,8 and 9. Validity measures G and ΔJ can be used to pick a 'best' c from each of these three groups determined by F. Thus, c=3,6 and 8 are judged to be the candidates for 'best' c. Figure 33 shows the FCM maximum membership clusters for the three dimensional case and c=3. It is obvious by comparing Figures 32 and 33 that further subdivision of the clusters in Figure 33 is necessary if a reasonable physical interpretation is to be obtained. Rather than pursuing the sequential clustering strategy, it was decided to implement a strategy similar to the one used in Example 1 and then contrast the results with the segmentation results shown in Figure 18.

Figures 34 and 35 show the FCM maximum membership clusters for the three dimensional case and c=6 and c=8 respectively. The result for c=8 in Figure 35 is judged to have a better physical interpretation when compared to the core information in Figure 32. The following description is identical to the one used in Example 1 and is used to merge the clusters in Figure 35 so that the segmentation results for this example can be compared with those of Example 1 shown in Figure 18.

TABLE X.

FCM VALIDITY MEASURES FOR EXAMPLE #4: WITH EUCLIDEAN NORM,ε = 0.01 AND M =1.5

c	F	G	ΔJ
2	0.939	−1.270	9.146
3	0.899	−1.125	8.958
4	0.800	−4.540	13.37
5	0.780	−3.971	13.36
6	0.797	−2.918	8.760
7	0.773	−3.731	9.709
8	0.742	−4.505	10.95
9	0.746	−4.329	10.76
10	0.698	−7.945	12.16

(a) Six Dimensional PC Space

c	F	G	ΔJ
2	0.944	−1.210	8.010
3	0.929	−0.814	6.100
4	0.821	−3.892	9.670
5	0.825	−2.500	7.077
6	0.827	−2.927	5.751
7	0.806	−3.061	5.941
8	0.805	−3.464	5.471
9	0.793	−3.657	5.677
10	0.758	−5.635	6.328

(b) Three Dimensional PC Space

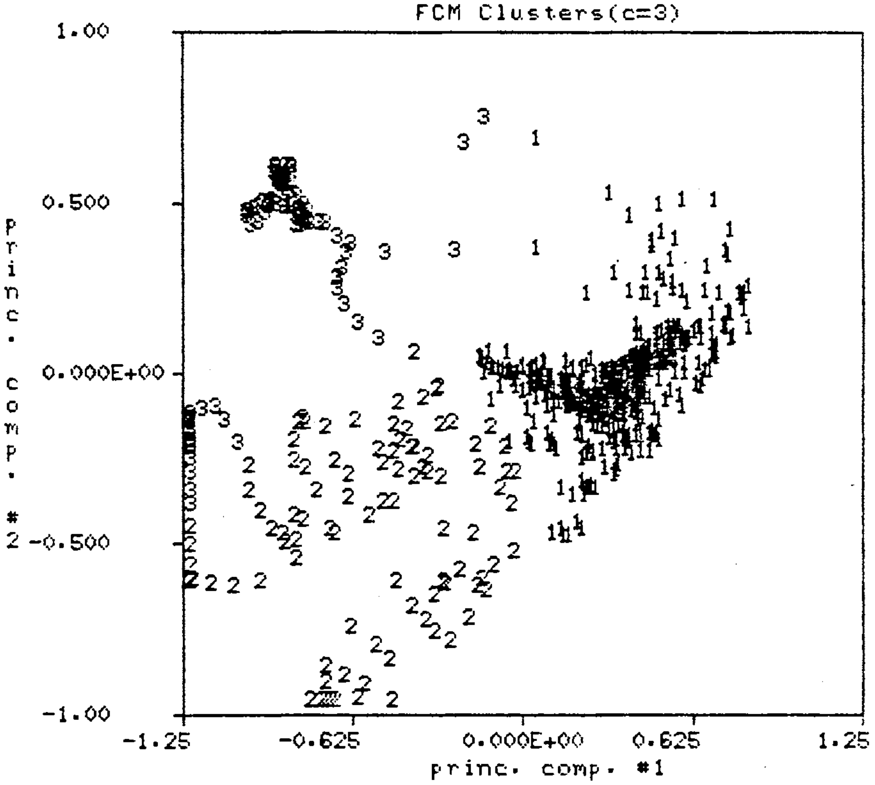
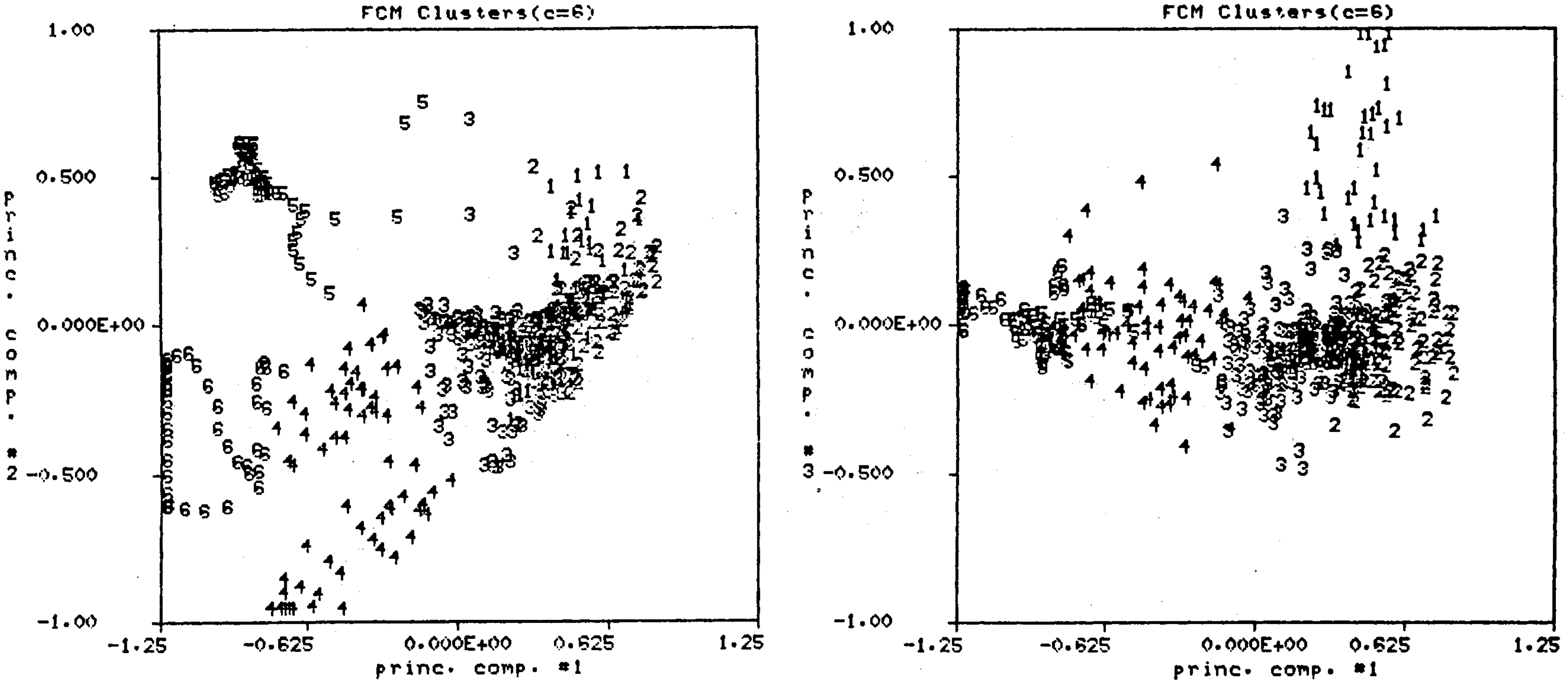


Figure 33. Maximum Membership Clusters for the Three Dimensional Case, c=3



(a)

(b)

Figure 34. Maximum Membership Clusters for the Three Dimensional Case, c=6
(a) PC logs #1 and #2 (b) PC logs #1 and #3

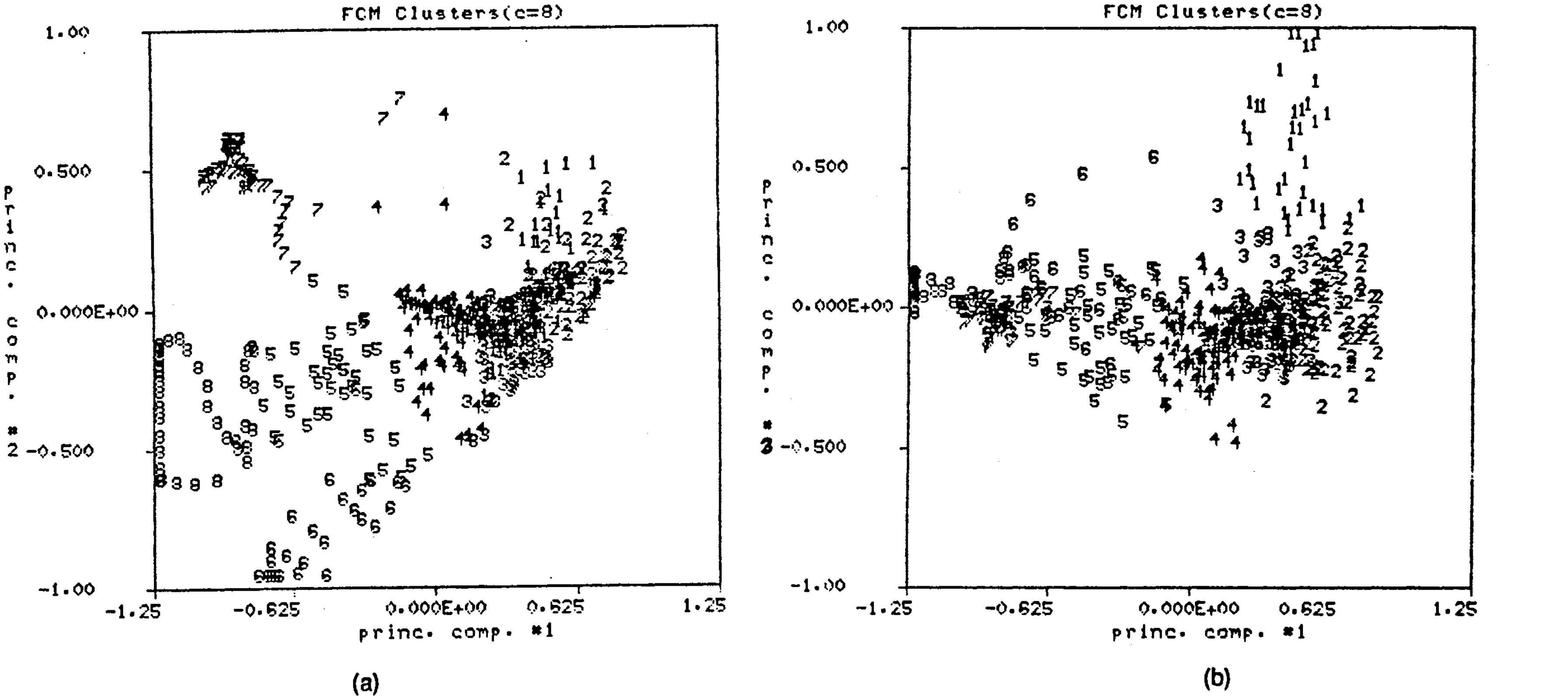


Figure 35. Maximum Membership Clusters for the Three Dimensional Case, c=8
(a) PC logs #1 and #2 (b) PC logs #1 and #3

New Cluster #	FCM Cluster # from Figure 35	Description
1	1	coal/shale
2	2,3	shale
3	4	shaley sandstone
4	6,8	limestone
5	5	limestone/sandstone/oil shows
6	7,8	sandstone

Figure 36 displays the original input logs along with the segmentation results from Example 1 in track #1 and the segmentation results from Example 4 in track #2. The numbers in both tracks comply with the description given in the previous paragraph. In general, there is good agreement between the segments 2,3 and 6 which represent shale, shaley sandstone and sandstone respectively. However, there are several notable differences between the segmentation results involving segments #1, #4 and #5. Consider the segments labeled '1' in track #2 at approximately 33 ft., 113 ft., 244 ft. and 321 ft. All four of these segments correspond to shale environments according the core description. The Example 1 segmentation results(track #1) label the same intervals '2' which are described as shale environments. In other words, the Example 4 segmentation results include more shale intervals in the segments labeled '1' and does a poorer job of isolating the thin coal intervals. Next, consider the segments labeled '4' in track #2 at approximately 37 ft. and 150 ft. Corresponding segments in track #1 are labeled '5'. The segments labeled '4' are described as limestone intervals. The segment in track #2 at 37 ft. agrees with the core information and might be considered an improvement over the Example 1 result. However, the segment at 150 ft. is a hydrocarbon bearing sandstone interval not a limestone interval. Perhaps more interesting is the fact that both the track #2 segments labeled '4' at 37 ft. and 150 ft. have significant hydrocarbon content. The final comparison between the segmentation results of Example 1 and Example 4 involves segments labeled '5'. In both examples these segments have the poorest correspondence to a particular lithology. There is a greater frequency of occurrence of segments labeled '5' in track #2 than in track #1. The thin segments labeled '5' in track #2 at approximately 52 ft., 53 ft., 101 ft., 105 ft., 111 ft., 268 ft. and 325 ft. have no similarly labeled counterparts in track #1. Except for the segment at 268 ft., these thin segments are associated with limestone intervals, which have distinctive resistivity log signatures, and the occurrence of these

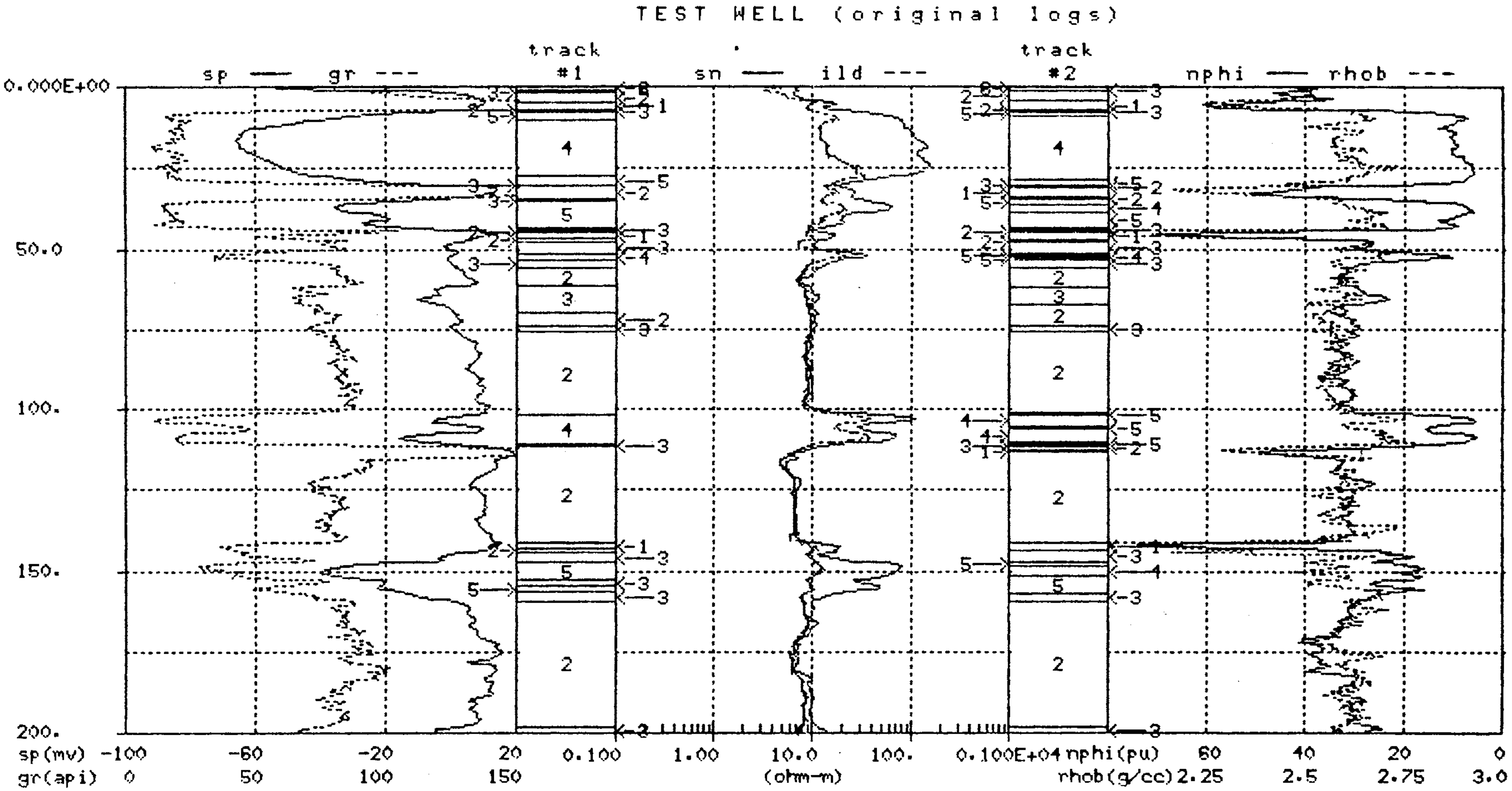


Figure 36. Comparison of Segmentation Results for: Example 1, (track#1) and Example 4, (track #2)

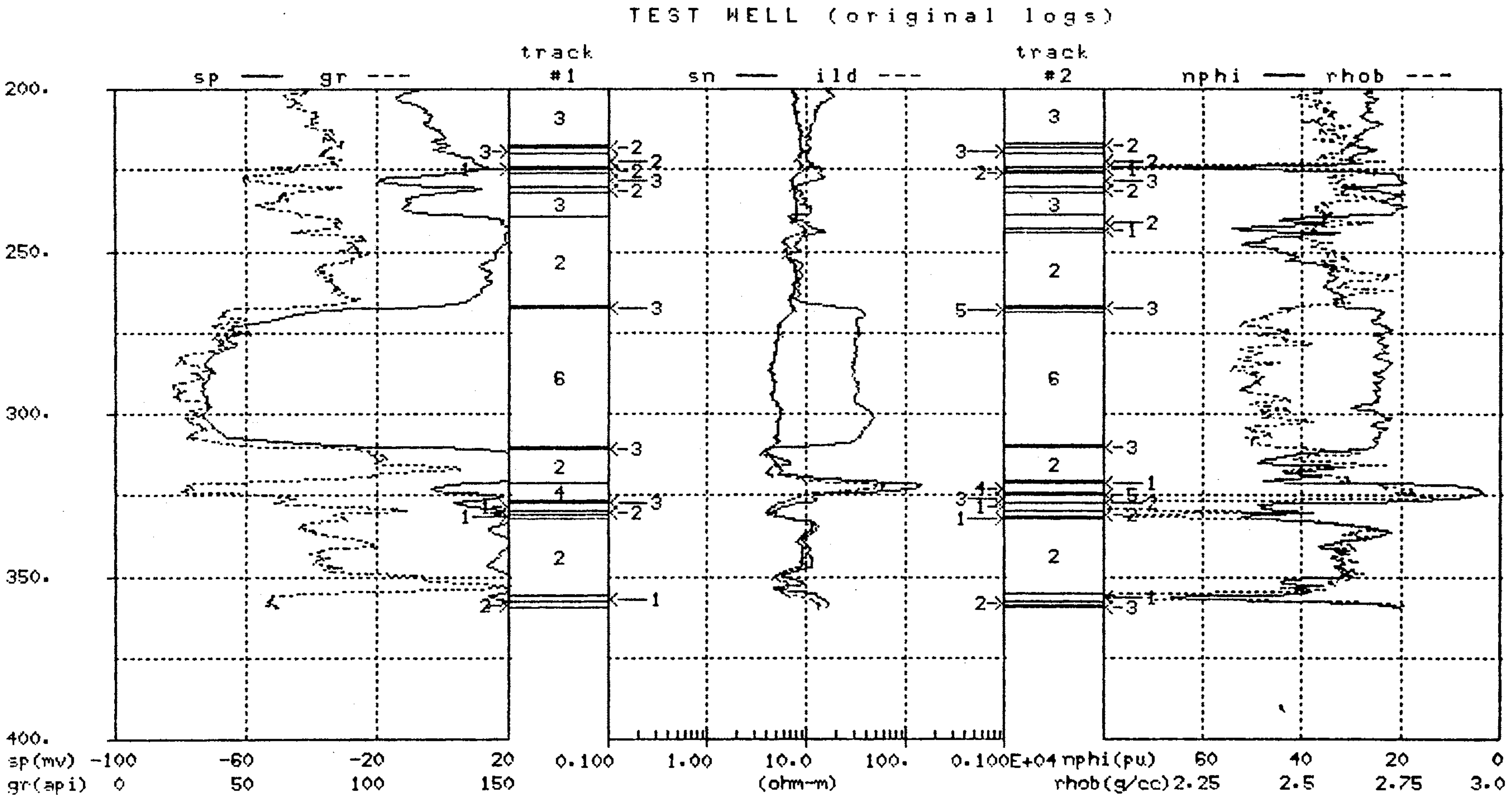


Figure 36. (continued)

segments may be attributed to the increased weighting of the resistivity logs in Example 4.

In conclusion, the increased weighting of the resistivity logs in Example 4 detracts from the ability of the cluster validity measures, F, G and ΔJ, to detect good data structure. A similar observation was made for the six and three dimensional cases in Example 1. Also, the scaling procedure of Example 4 tends to make the isolation of the coal samples by the FCM clustering algorithm even more difficult. In general, the overall physical interpretation of the segmentation results for Example 4 is less consistent than the physical interpretation for the Example 1 segmentation results.

3.3 Chapter Summary

The examples in this chapter have illustrated the ability of a specific segmentation algorithm to segment a borehole based upon wireline log responses. The essential steps of the segmentation algorithm are: 1) a linear scaling process which scales the input logs to zero mean signals bounded by plus and minus one, 2) the calculation of PC logs based upon the scaled input logs, 3) the clustering of the PC log data using an FCM clustering algorithm with Euclidean norm and m=1.5 and 4) the interpretation of the FCM algorithm output using cluster validity measures F, G and ΔJ.

This segmentation algorithm was applied to a test interval which consists of four main lithology types; shale, sandstone, limestone and coal. The test interval also included hydrocarbon bearing and water bearing zones. Three general observations about the segmentation algorithm examples are given below.

1. Except for Example 4, the basic structure of the log data is carried by the first two PC logs.
2. The basic data structure indicated by the cluster validity measures has a reasonable physical interpretation.
3. The ability of the cluster validity measures to detect good substructure in the data and the physical interpretation of the borehole segments is best when the influence of the resistivity logs is limited.

In addition, Example 2 illustrated how physically important information is sometimes carried by the higher numbered PC logs.

The physical interpretation of the segmentation algorithm output has two recurring inconsistencies. One is the fact that the algorithm labels certain shale segments the same as the coal intervals. The segmentation results for the four dimensional case in Example 2 are the best in this respect, labelling only two shale segments the same as the coal segments. The second inconsistency is the grouping which includes both limestone samples and sandstone samples from the hydrocarbon bearing sandstone unit. The other inconsistencies in the physical interpretation of the segmentation algorithm's output occur in transition zones from one lithology to the next lithology and it does not seem to matter whether the transitions occur abruptly or gradually. For example, the transition may be an abrupt shale-coal-sandstone transition or a gradual transition from sandstone to shale. In either event, it is likely that certain discrepancies will exist between the core description information and the output of the segmentation algorithm.

CHAPTER IV

MULTIWELL APPLICATION OF THE SEGMENTATION ALGORITHM

4.0 Introduction

Chapter IV extends the segmentation algorithm to a multiwell environment using data from eight wells in the Hartzog-Draw Field, Wyoming. The application of the segmentation algorithm to multiple wells is motivated, in part, by the question, " Is it possible to identify segments between wells in the same field which have similar wireline log characteristics?" The next question is, "Do these segments have geological significance?" and if so, is it possible to design a classifier to reliably identify similar segments in other wells in the same field?

The eight wells are arbitrarily labeled as Wells #1-#8 and Section 4.1 describes the data base from the Shannon Sandstone of the Hartzog-Draw Field which is used for this part of the study. Section 4.2 uses log data from Wells #6, #7 and #8 in a multiwell example which investigates the ability of the segmentation algorithm to identify segments with similar wireline log characteristics between the three wells. Section 4.3 compares the segmentation algorithm output for Well #8 to a well accepted geological facies description of the Shannon Sandstone interval. Finally, Section 4.4 investigates the possibility of designing a nearest prototype classifier to identify segments with particular wireline log characteristics in the other five wells within the available Hartzog-Draw data base.

4.1 Data Base

The Hartzog-Draw Field is located in the Powder River Basin of northeastern Wyoming. Since its discovery in 1975, a wealth of information has been accumulated in the form of digitized wireline logs, cores and core descriptions and various types of reservoir analyses. Most of this information has been collected with respect to the

Shannon Sandstone because of its good reservoir properties and it is the Shannon Sandstone which is considered here.

Digital wireline log data for eight wells in the Shannon Sandstone, Hartzog-Draw Field was provided by OXY Inc., of Tulsa OK. These wells are referred to generically as Wells #1-#8. This wireline log data includes: gamma ray(gr), spontaneous potential(SP), spherically focused(SFL), medium induction(ILM), deep induction(ILD), neutron porosity (NPHI), bulk density(RHOB) and interval transit time(DT) information for six of the eight wells. Well #4 is lacking the SFL information and Well #5 is lacking the SFL and SP information. This log data provides the necessary inputs to the segmentation algorithm for the multiwell example in Section 4.2.

In addition to the wireline log data, geological documentation was also provided to describe the major geological facies within the Shannon Sandstone. There are six facies observed in cores: 1) interbar, 2) bar margin I, 3) bar margin II, 4) bioturbated siltstone, 5) central bar and 6) shelf silty shale. This geological information is contrasted, in Section 4.3, to the segmentation algorithm results for Well #8 from Section 4.2. The Shannon Sandstone contains up to the first 5 facies types with the shelf silty shale facies overlying and underlying the Shannon Sandstone interval. The geological environment associated with the Shannon Sandstone is basically a shale-sand-shale sequence and is much simpler than the geological environment for the test well data used in Chapter III.

Similar log and core data was used by Almon[2] in an application of discriminant function analysis to discriminate between the six facies types(see Section 1.3.2). Also of importance is a multiwell Faciolog evaluation of wells within the Hartzog-Draw Field which will provide another basis of comparison for some of the results in Section 4.2.

4.2 Multiwell Example

Wells #6, #7, and #8 were chosen from the data base for use in the multiwell example. These three wells were chosen for three reasons: 1) each well has a full complement of logs, 2) the wells lie on an east to west line across the Hartzog-Draw Field with Well #6 lying approximately halfway between Well #7 and Well #8 and 3) two of the three wells were used in the multiwell Faciolog evaluation reported by Widdicombe, et. al.[61].

Initially, a 250 ft. interval, which includes the Shannon Sandstone and the overlying and underlying shale units, was selected for analysis. The basic methodology of Section 3.0 is applied, in turn, to Wells #6, #7 and #8. Let's consider Well #6. All eight logs are used as inputs to the segmentation algorithm. The input logs are scaled using the linear scaling procedure described as method 3 in Section 2.2.2. The input logs are transformed into PC logs according to Equation (31) and then the FCM algorithm is applied to the two dimensional PC log data. The choice to use only the first two PC logs

$$\begin{bmatrix} \text{pc\#1} \\ \text{pc\#2} \\ \text{pc\#3} \\ \text{pc\#4} \\ \text{pc\#5} \\ \text{pc\#6} \\ \text{pc\#7} \\ \text{pc\#8} \end{bmatrix} = \overset{\underline{\text{KLT Matrix}}}{\begin{bmatrix} 0.502 & 0.462 & -0.367 & -0.424 & -0.417 & 0.071 & 0.189 & 0.069 \\ 0.285 & -0.234 & 0.047 & 0.016 & 0.022 & 0.633 & -0.424 & 0.528 \\ 0.352 & -0.241 & 0.229 & 0.188 & 0.180 & 0.151 & 0.810 & 0.136 \\ 0.733 & -0.126 & 0.167 & 0.149 & 0.170 & -0.326 & -0.352 & -0.370 \\ 0.007 & -0.722 & -0.271 & -0.314 & -0.253 & -0.426 & 0.026 & 0.247 \\ 0.004 & 0.354 & 0.351 & 0.108 & 0.018 & -0.521 & -0.058 & 0.681 \\ 0.050 & 0.098 & -0.705 & 0.158 & 0.648 & -0.106 & 0.0147 & 0.188 \\ 0.030 & -0.020 & -0.299 & 0.792 & -0.530 & -0.026 & -0.001 & 0.030 \\ & & & & & & & \end{bmatrix}} \begin{bmatrix} \text{gr} \\ \text{sp} \\ \text{sfl} \\ \text{ilm} \\ \text{ild} \\ \text{nphi} \\ \text{rhob} \\ \text{dt} \end{bmatrix} \qquad (31)$$

is based on the results in Chapter III which indicate that the general structure of the data is captured by the first two PC logs. Notice that the scaled GR, SP, ILM and ILD values have the largest weights in the calculation of the first principal component and the three porosity logs have the greatest weights in the calculation of the second principal component. Equation (31) is specific to Well #6 but the same observations hold for the computation of the PC logs for Wells #7 and #8.

The FCM algorithm, with m=1.5 and ε = 0.01, is applied to the two dimensional PC log data using the Euclidean norm and c=2,3,4,5,6,7 and 8. Table XI lists the cluster validity measures for the various alternatives for Well #6. The best indication of good substructure is judged to be when c=4. The F values for c=2,3 and 4 are all of comparable magnitude and c=4 is preferred over 2 and 3 because ΔJ is minimum and G indicates good spatial separation of the maximum membership clusters shown in Figure 37. The segmentation results corresponding to Figure 37 are shown in Figure 38 along with six of the eight input logs. (The plotting routine is limited to two signals per grid.)

An identical process is applied to the log data of Wells #7 and #8. Figures 39(a) and 39(b) show the FCM clusters determined for Wells #7 and #8 respectively and Figures 40 and 41 show the corresponding segmentation results. For this relatively large 250 ft. interval there is good agreement among the segmentation results for Wells #6, #7 and #8. The following observations are generally true for the segmentation

TABLE XI.

FCM VALIDITY MEASURES FOR HARTZOG DRAW WELL #6 WITH EUCLIDEAN NORM,ε = 0.01 AND m =1.5

c	F	G	ΔJ
2	0.974	−0.239	2.894
3	0.979	−1.328	1.562
4	0.978	−0.289	0.665
5	0.957	−0.862	0.958
6	0.877	−3.218	1.079
7	0.892	−2.935	0.720
8	0.893	−2.798	0.681

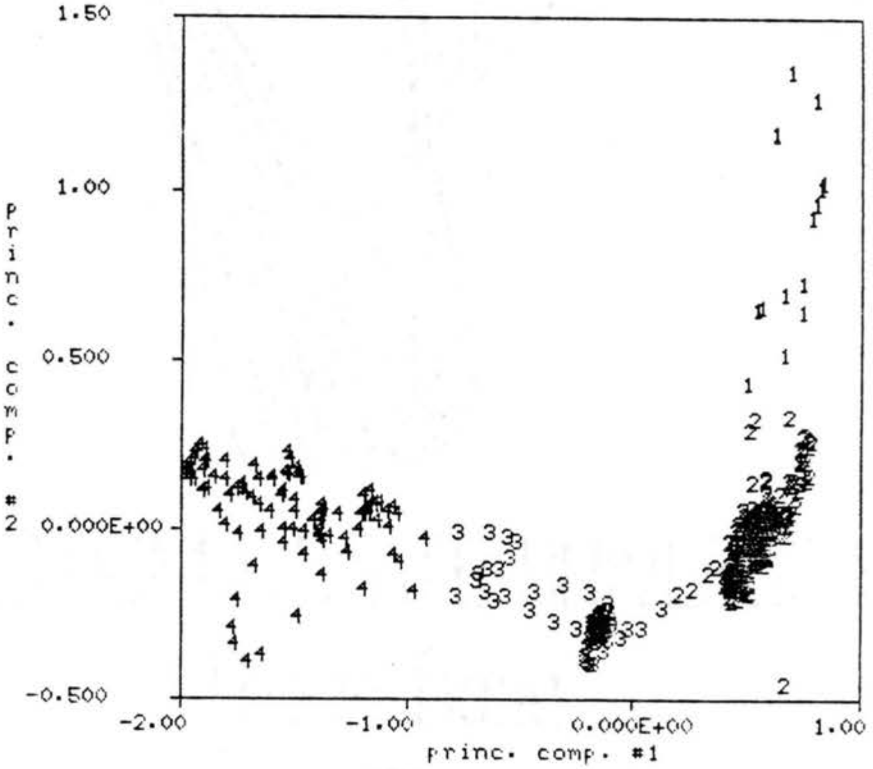


Figure 37. Maximum Membership Clusters, Well #6, c=4

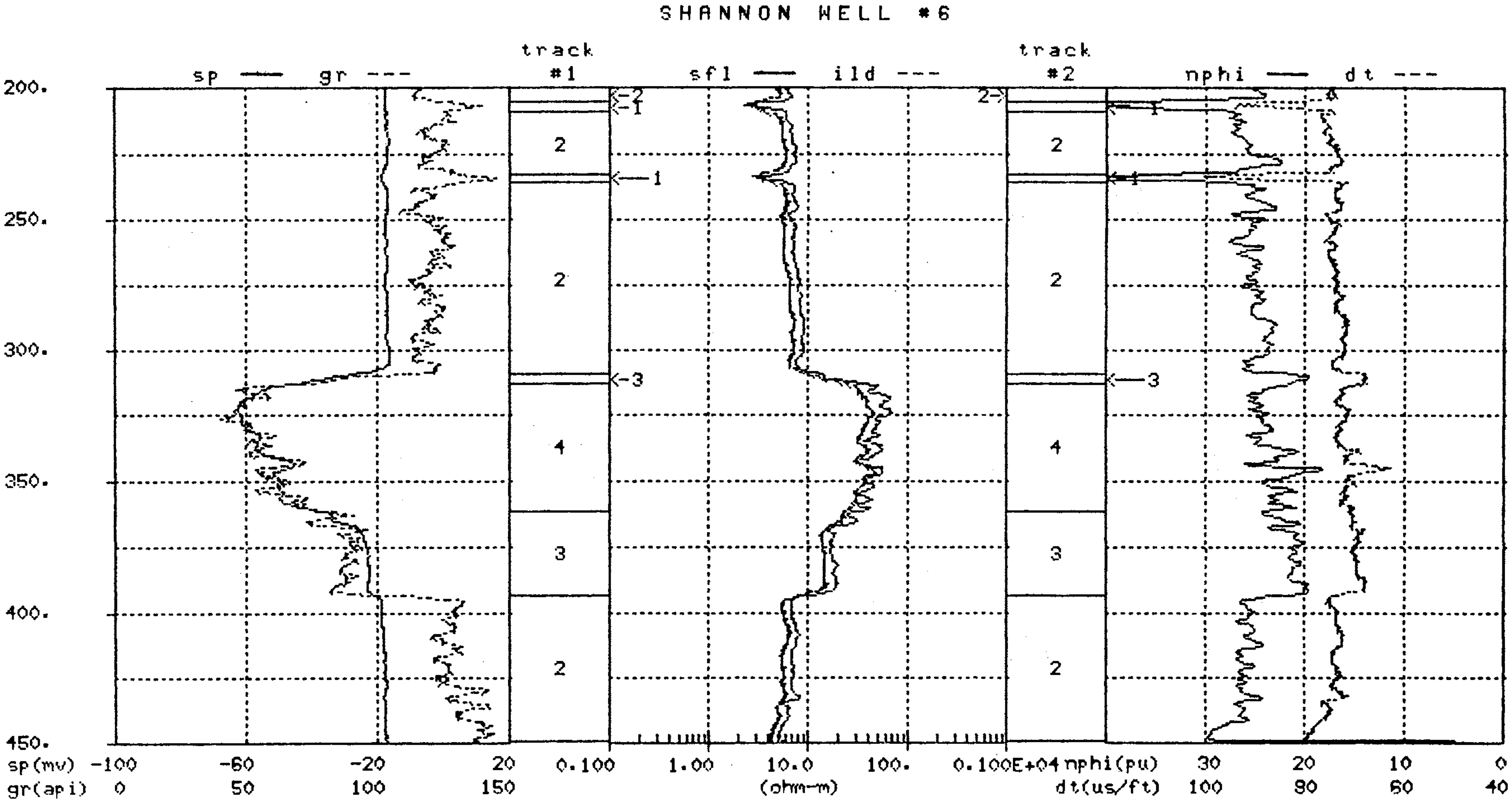
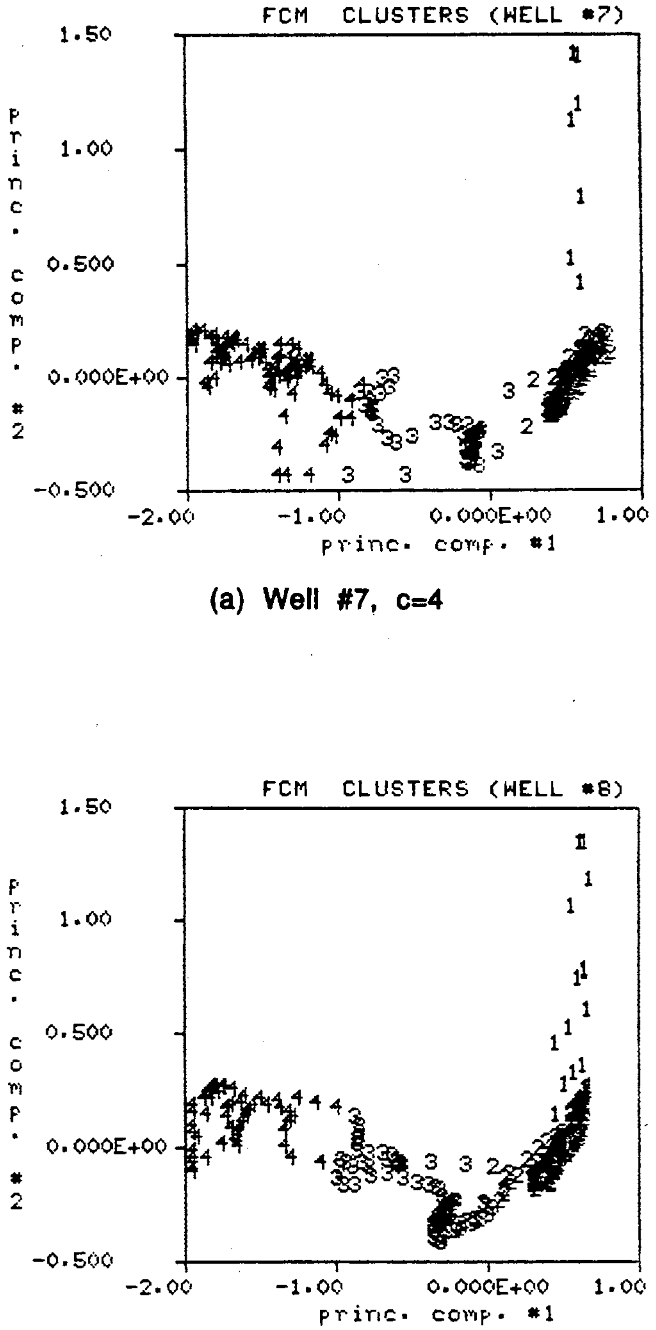


Figure 38. Segmentation Algorithm Output for Well #6



(a) Well #7, c=4

(b) Well #8, c=4

Figure 39. Maximum Membership Clusters for Well #7, c=4 and Well #8, c=4

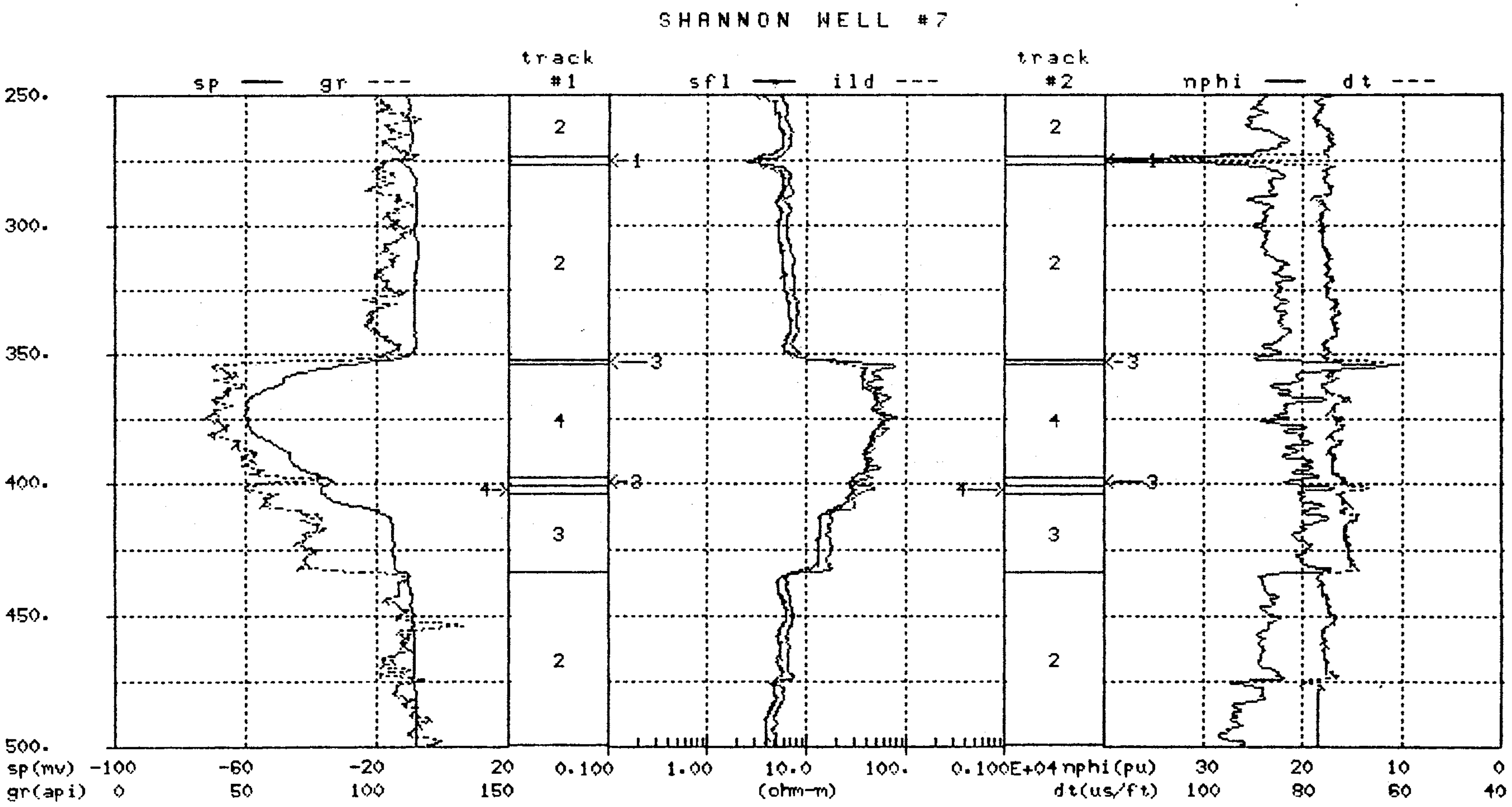


Figure 40. Segmentation Algorithm Output for Well #7

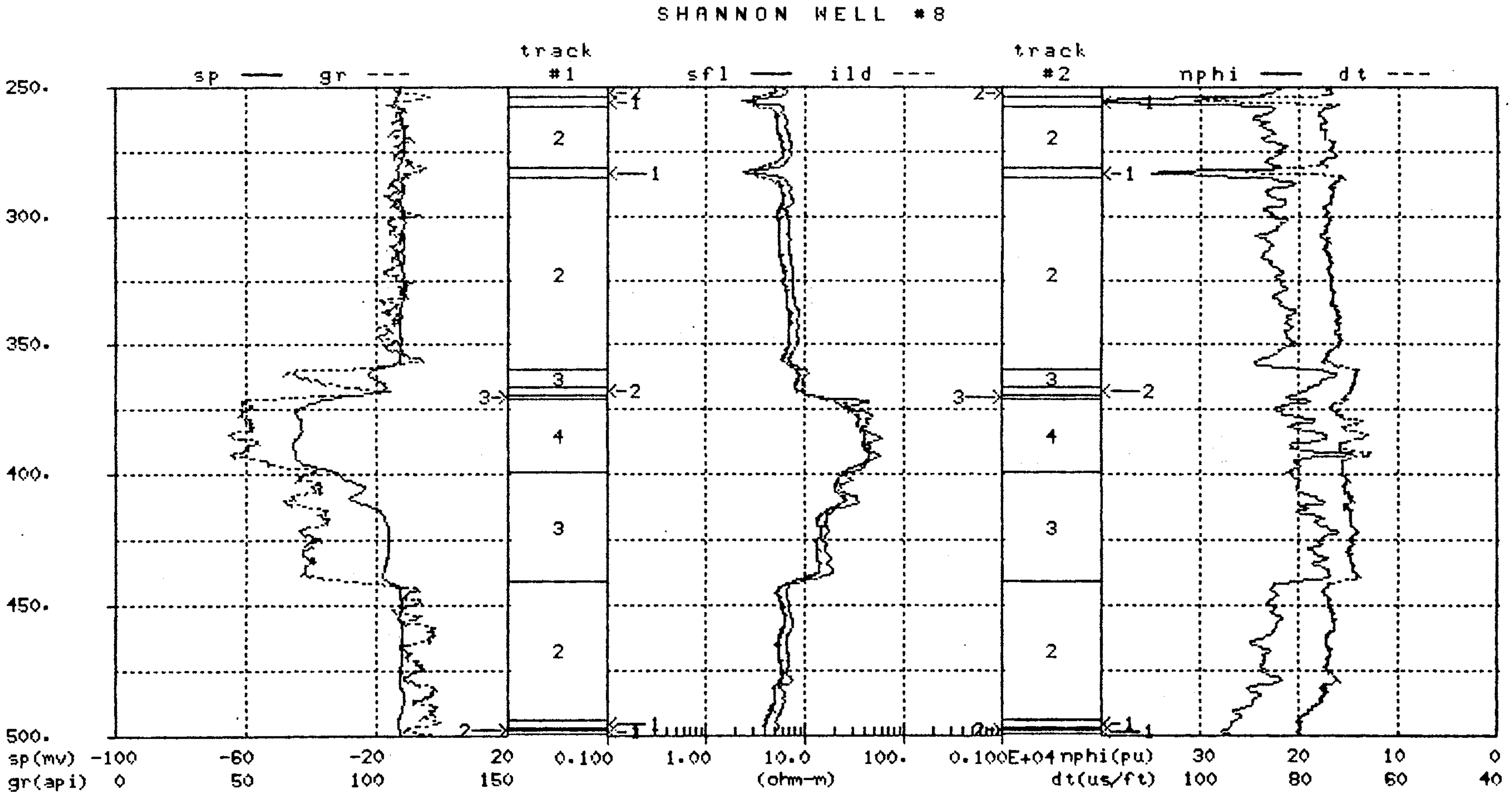


Figure 41. Segmentation Algorithm Output for Well #8

results displayed in Figures 38, 40 and 41. Segment #1 corresponds to the very distinctive bentonitic shale marker beds which occur near the top of each interval. Segment #2 corresponds to the shelf silty shale facies which flanks the Shannon Sandstone both above and below. Segments #3 and #4 encompass the Shannon Sandstone with segment #4 relating to the central bar facies and segment #3 incorporates the remainder of the Shannon Sandstone interval. The only exceptions to this description of the segmentation results occur in Well #8. There are two segments labeled '1' at the bottom of the interval for Well #8 which are part of the shelf silty shale facies and there is a segment labeled '2' within the Shannon Sandstone interval which is not part of the shelf silty shale facies. A more detailed look at the Shannon Sandstone is forthcoming in Section 4.3 but first the segmentation procedure is repeated for the smaller intervals bounded by segments #3 and #4 within each of the three wells.

More specifically, the intervals of interest for this second application of the segmentation algorithm are 310-392 ft., 352-431 ft. and 359-441 ft. for Wells #6, #7 and #8 respectively. This time the inputs to the segmentation algorithm include the GR, SP, NPHI, RHOB and DT logs. The resistivity logs are eliminated from the inputs to the segmentation algorithm based on the observation in Chapter III that their influence tends to detract from the ability of the segmentation algorithm to detect good data structure and due to the resistivity logs sensitivity to hydrocarbons. The inputs are scaled and transformed using the KLT. All FCM parameters remain unchanged from the first application of the segmentation algorithm and c is varied from 2 to 10. Table XII shows the cluster validity measures for all three wells. The initial indication of good data substructure in Well #6 is for c=5, with secondary substructure indicated for c=8. Recall the guideline for interpretation of the validity measures is to use F to divide the c values into groups then use G and ΔJ to pick a 'best' value of c from each group. In Table XII(a) c=2 is alone in one group, c=3, 4, 5 and 6 form a second group, c=7 and 8 form a third group and c=9 and 10 form a final group. The large ΔJ value detracts from choosing c=2 as the primary indication of substructure in the data. From the second group, c=5 is chosen as the primary indication of good data substructure and c=8 is chosen from the third group as an indication of secondary substructure in the Well #6 data. Figure 42 shows the maximum membership clusters for Well #6, c=5. For Well #7, c=6 is the primary indication of data substructure and there is no obvious indication of secondary substructure in the data. Figure 43 shows the maximum membership clusters for Well #7, c=6. In Well #8, it seems reasonable to pick c=3 as

TABLE XII.

FCM VALIDITY MEASURES FOR SHANNON SANDSTONE, WELLS #6 ,#7 AND #8 WITH EUCLIDEAN NORM, ε = 0.01 AND m = 1.5

c	F	G	ΔJ
2	0.948	–0.511	4.831
3	0.893	–1.266	4.139
4	0.889	–1.057	2.945
5	0.908	–0.840	1.533
6	0.893	–0.921	1.226
7	0.874	–0.954	1.368
8	0.879	–0.837	0.981
9	0.841	–1.916	1.269
10	0.859	–1.504	0.759

(a) Well #6

c	F	G	ΔJ
2	0.927	–1.124	4.986
3	0.925	–0.984	2.363
4	0.904	–1.240	2.442
5	0.919	–0.901	0.923
6	0.915	–0.652	0.623
7	0.877	–1.309	0.960
8	0.873	–1.309	0.569
9	0.868	–1.309	0.527
10	0.854	–0.650	0.496

(b) Well #7

c	F	G	ΔJ
2	0.931	–0.823	4.034
3	0.928	–0.196	1.726
4	0.921	–1.244	1.600
5	0.884	–1.471	1.681
6	0.877	–1.873	1.400
7	0.876	–1.832	1.172
8	0.878	–1.832	1.008
9	0.870	–1.232	0.710
10	0.852	–1.232	0.742

(c) Well #8

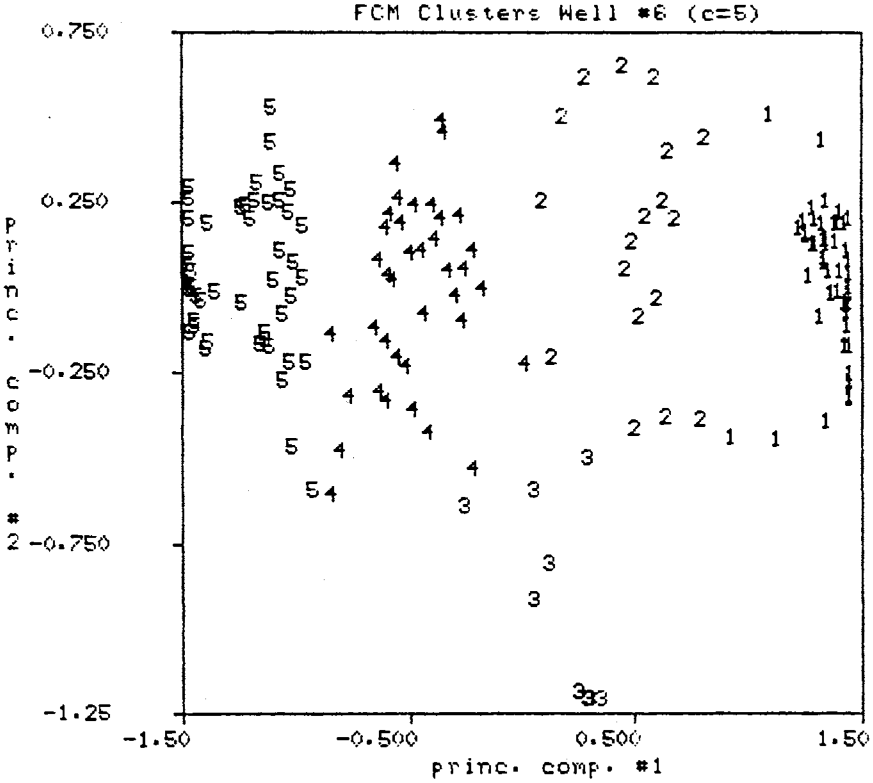


Figure 42. Maximum Membership Clusters Shannon Sandstone, Well #6, c=5

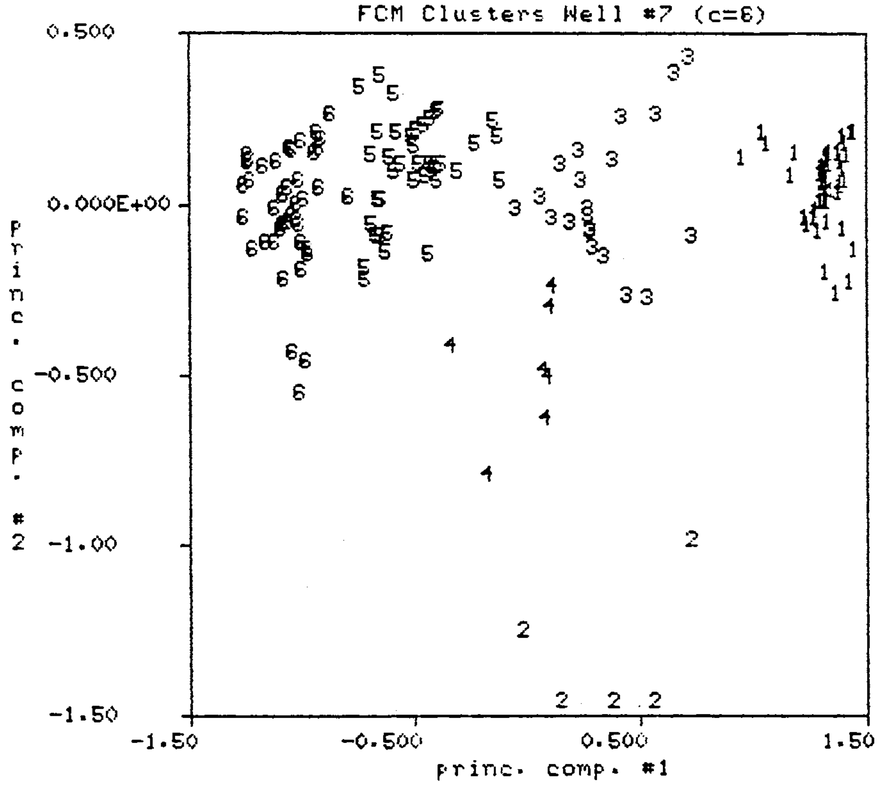


Figure 43. Maximum Membership Clusters Shannon Sandstone, Well #7, c=6

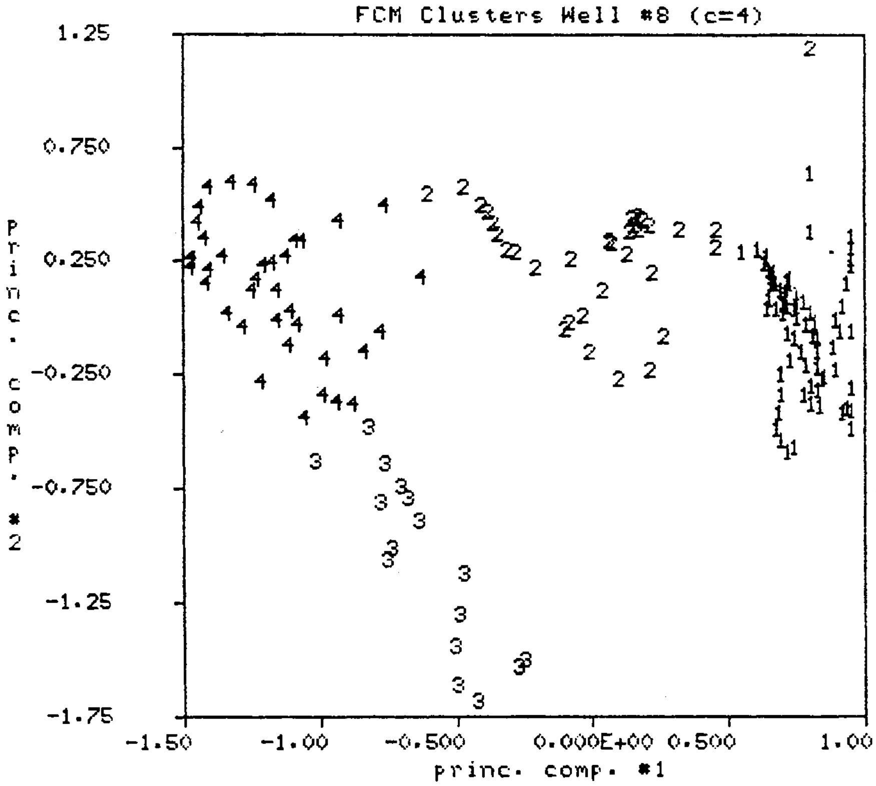


Figure 44. Maximum Membership Clusters Shannon Sandstone, Well #8, c=4

the primary indication of data substructure. Following the stated guideline for interpretation of the validity measures, c=2, 3, 4 are grouped together based upon the comparable F values. The value, c=3, is preferred over 2 and 4 because G indicates 'much better' spatial separation between the worst case pair of maximum membership clusters for c=3. However, validity measure G has the problem of being very sensitive to 'outliers' within the clusters. Observe the maximum membership clusters for Well #8, c=4 in Figure 44 and note the sample labeled '2' in the upper right hand corner of the figure. It is the judgement of this observer that this 'outlier' in cluster #2 distorts G and that c=4 is the primary indication of substructure for Well #8. This points to the need to visually verify the indications of the validity measures. Secondary substructure for Well #8 is indicated for c=9.

For the initial comparison c=5, c=6 and c=4 are used for Wells #6, #7 and #8 respectively, and the corresponding segmentation results are shown in track #1 of Figures 46, 47 and 48. The secondary substructure for Well #6, (c=8) and Well #8, (c=9) is shown in track #2 of Figures 46 and 48 respectively. The primary and secondary data structures shown for Well #7 in Figure 47 are identical. Two main factors are used to make the comparison between the segments in the three wells. First, the vertical position of the segments within the respective boreholes is taken into account and second, the spatial distribution of the clusters in Figures 42, 43 and 44 help establish the correspondence shown in Figure 45. The cluster centers for the clusters in Figures 42, 43 and 44 are expressed in the domain of the original wireline logs in Table XIII. The initial segmentation results in Figures 38, 40 and 41 have already established a correspondence between certain intervals between wells. Working from the premise that segment #4 in Figure 38 relates to segments #4 in Figures 40 and 41 leads to the conclusion that segments #3, #4 and #5 in Figure 46(track #1) correspond in some fashion to segments #4, #5 and #6 in Figure 47(track #1) which in turn correspond to segments #3 and #4 in Figure 48(track #1). The cluster center information in Table XIII helps characterize the attributes of the segments in each well with respect to the original wireline log data. Notice that center #1 for all three cases in Table XIII corresponds to relatively high GR reading, a suppressed SP value and a moderate porosity indication on the porosity logs. Also, the vertical position of segment #1 within each well indicates that these segments correspond to each other. Segment #2 in Well #7 has a significantly lower porosity indication than the other segments in Well #7 and is judged to be a separate segment without counterpart in the adjacent wells. Segment #2 in Well #6 has a slightly lower GR value and a slightly more developed SP than segment #1 in the same well. Similar characteristics exist for segment #3 in

Well #7 and segment #2 in Well #8. Again, the vertical position of these segments within the borehole helps establish this correspondence. In a similar fashion, a correspondence among the remaining segments in the three wells is established and shown in Figure 45.

This example provides evidence that it is feasible to detect segments with similar wireline log characteristics among wells in the same field. The next question is, "How do these segments relate to the geological facies known to exist in the Shannon Sandstone?"

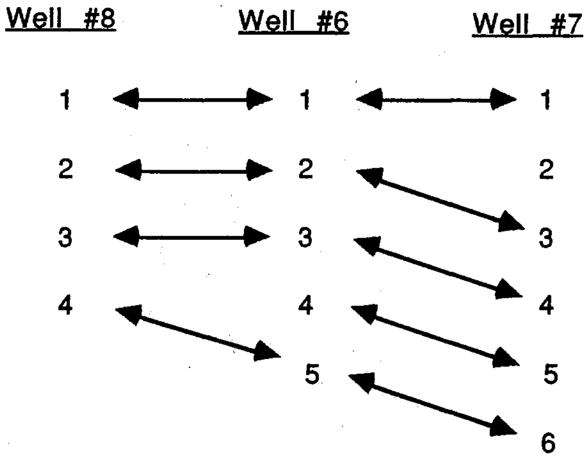


Figure 45. Manual Correlation of Segments Between Wells

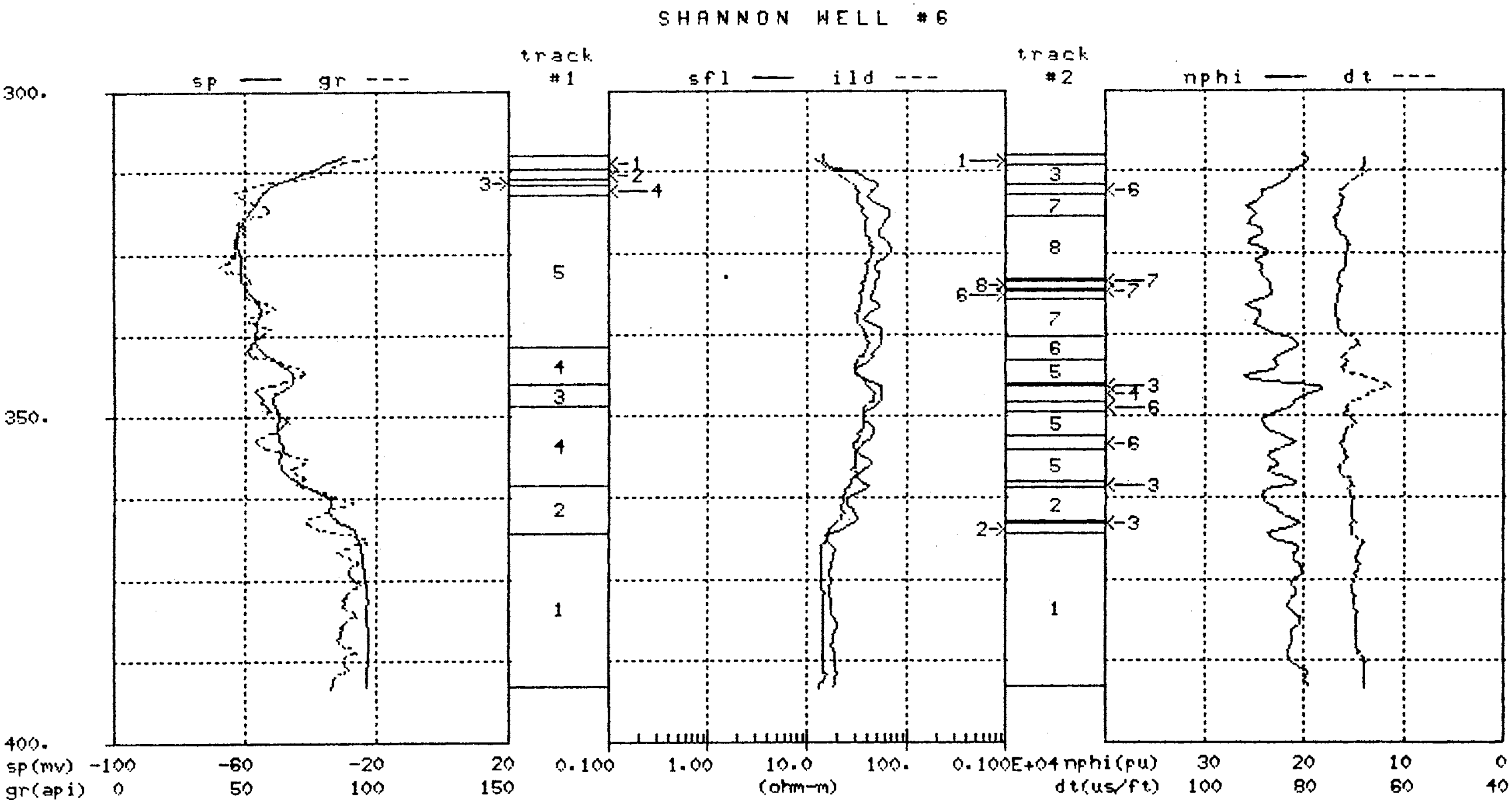


Figure 46. Segmentation Algorithm Output Showing Primary(c=5, track #1) and Secondary(c=8, track #2) Structure for the Shannon Sandstone, Well #6

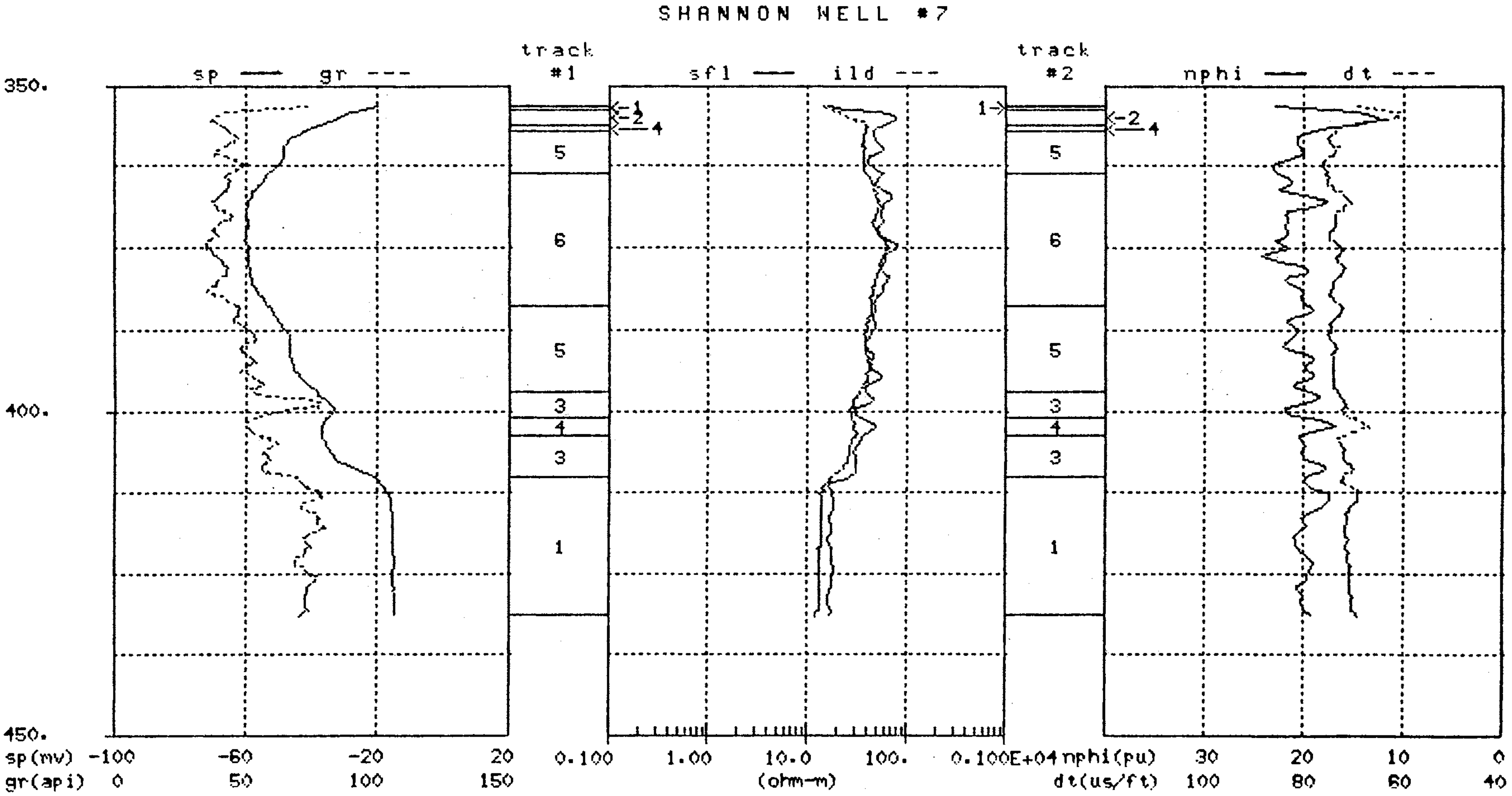


Figure 47. Segmentation Algorithm Output Showing Primary(c=6,track #1) and Secondary(c=6,track #2) Structure for the Shannon Sandstone, Well #7

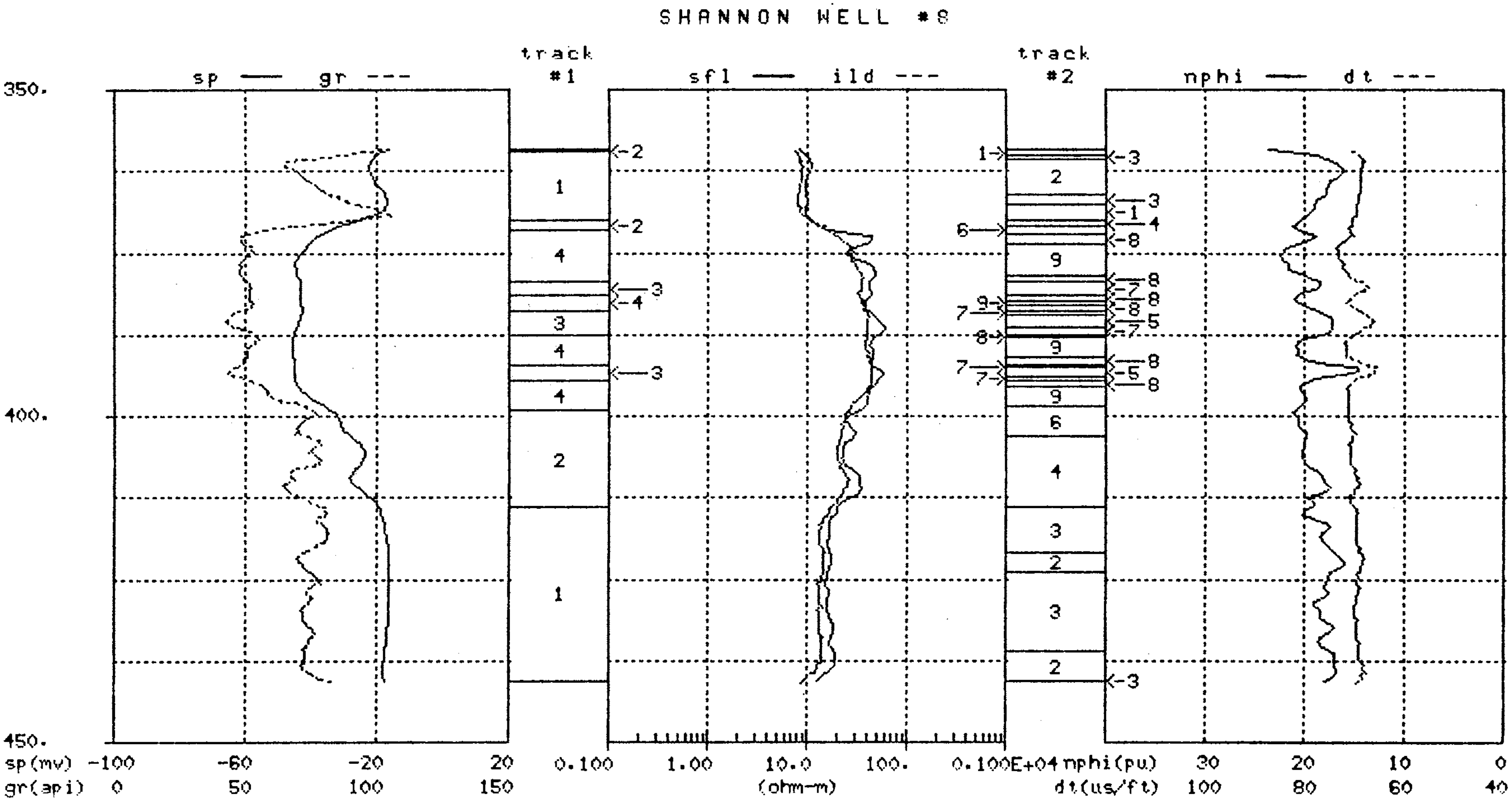


Figure 48. Segmentation Algorithm Output Showing Primary(c=4, track #1) and Secondary(c=9, track #2) Structure for the Shannon Sandstone, Well #8

TABLE XIII.

CLUSTER CENTERS FOR WELL #6, WELL #7 AND WELL #8
IN THE DOMAIN OF THE ORIGINAL WIRELINE LOGS

Cluster Center	# of Samples	GR	SP	NPHI	RHOB	DT
1	52	89.3	–24.6	20.9	2.54	69.1
2	18	78.4	–35.1	22.4	2.50	70.5
3	9	58.8	–47.6	19.5	2.52	67.8
4	38	61.8	–49.3	23.1	2.46	71.3
5	47	51.7	–58.9	24.2	2.42	72.5

(a) Well #6

Cluster Center	# of Samples	GR	SP	NPHI	RHOB	DT
1	45	74.0	–16.0	19.5	2.56	70.8
2	5	41.9	–29.9	13.6	2.62	63.8
3	21	60.0	–32.9	20.1	2.51	72.0
4	7	48.7	–38.1	18.1	2.53	69.7
5	39	49.1	–47.8	21.0	2.48	73.7
6	41	39.9	–57.3	20.8	2.44	73.8

(b) Well #7

Cluster Center	# of Samples	GR	SP	NPHI	RHOB	DT
1	75	77.8	–18.4	18.0	2.58	69.2
2	34	72.5	–26.4	19.7	2.53	70.6
3	16	44.8	–42.5	16.9	2.54	67.8
4	39	54.5	–42.9	20.6	2.47	71.2

(c) Well #8

4.3 Geological Description of The Shannon Sandstone

The comparison of segmentation results for the Shannon Sandstone with geological descriptions of the same interval is restricted to Well #8 but is judged to be representative of similar comparisons for Well #6 and Well #7. Figure 49 shows the geological facies in track #1 and the rock types in track #2 for the Shannon Sandstone in Well #8. The GR, SP, SFL, ILD, NPHI and DT logs are included in Figure 49 for reference purposes. The facies and rock types were determined by a qualified geologist with reference to the proper documentation[27,34,45,54,57,61]. The respective facies and rock descriptions are given in Table XIV. The borehole segmentation results will be compared primarily to the facies descriptions given in Table XIV and shown in track #1 of Figure 49. These facies descriptions are well documented geological descriptions of the Shannon Sandstone taken from core analyses. The rock type definitions in Table XIV correspond to track #2 of Figure 49 and are taken from the "electrofacies" determined by Widdicombe et., al.[61] in their multiwell Faciolog evaluation of four wells in the Hartzog-Draw Field. These "electrofacies" were determined from the GR, NPHI, and DT wireline logs and the volume of clay(VCL) computed log. The rock type definitions in Table XIV are determined by relating the Faciolog "electrofacies" in track #2 of Figure 49 back to core descriptions of the same interval. Since the origin of the rock type information is from log information rather then core information, it is presented here merely as an example of the Faciolog procedure applied to the Shannon Sandstone interval. Figure 50 duplicates the facies of Figure 49, track #1 and contrasts these facies to the segmentation results for the same interval. It is interesting to note that the segmentation of the borehole based on the wireline logs does not compare as favorably to the geological facies as one might hope. A similar statement could be made with reference to the rock types shown in track #2 of Figure 49. Figure 51 shows the Well #8 facies and rock types crossplotted using the first two PC logs determined for Well #8. The zeros appearing in these crossplots correspond to sample points in the interval without either a facies or rock type assigned to them. Facies #5, the central bar facies, in Figure 51(a) tend to cluster and relates to clusters #3 and #4 in Figure 44. This is the best comparison between facies type and FCM clusters for Well #8. Other attempts to relate facies or rock types to the FCM clusters in Figure 44 are tenuous at best. However, facies #5 is the central bar facies and has the best reservoir potential, so it is beneficial that this is the one geological

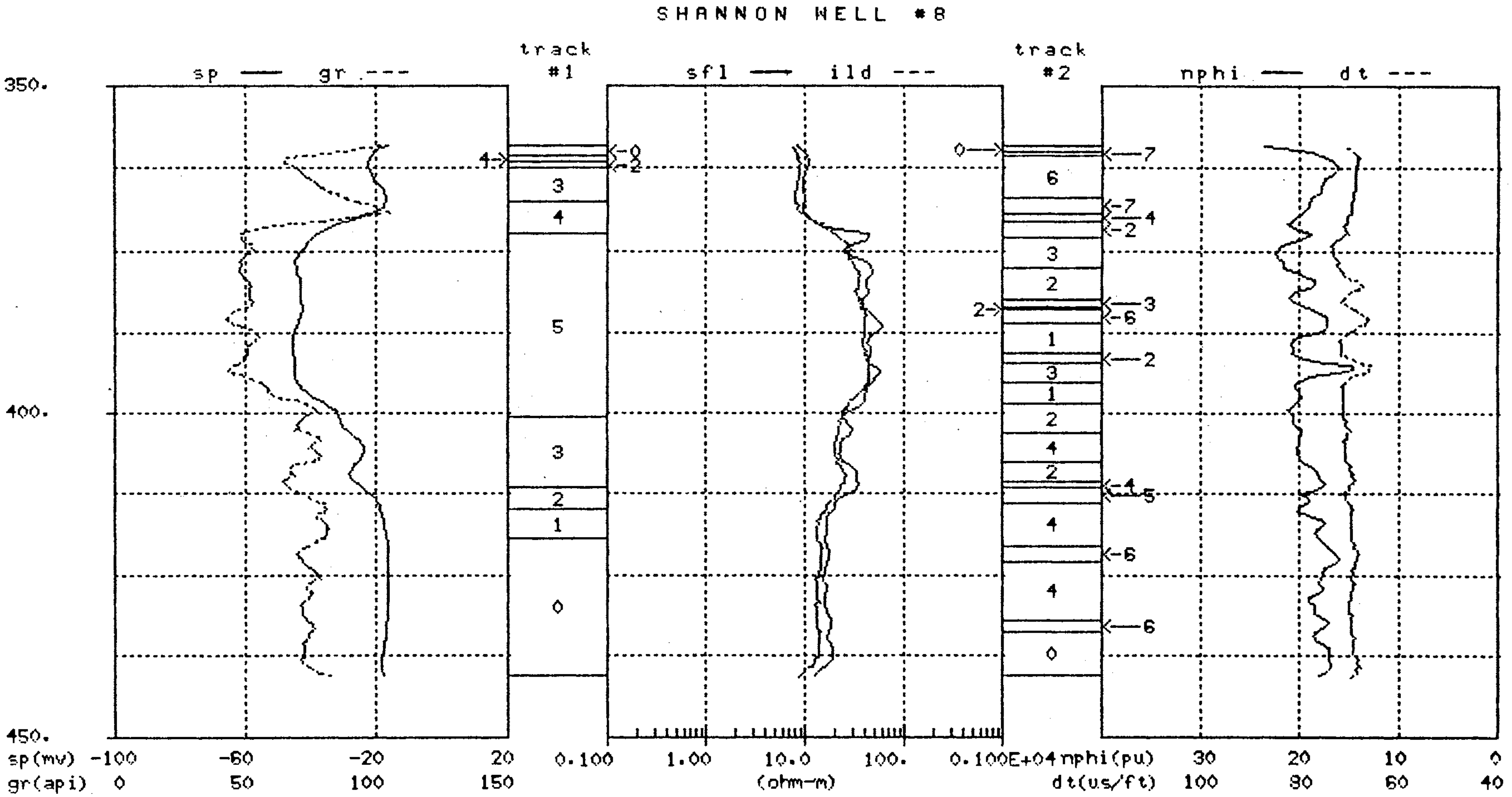


Figure 49. Geologic Facies(track #1) and Rock Types(track #2) for the Shannon Sandstone, Well #8

TABLE XIV.

FACIES AND ROCK TYPES FOR THE SHANNON SANSTONE

Geologic Facies #	Description
0	Undefined
1	Interbar facies - burrowed, ripple bedded, very shale laminated, fine grained sandstone
2	Bar margin facies I - cross bedded and rippled, coarse to medium grain sandstone with clasts of shale and sideritic mudstone
3	Bar margin facies II - moderately burrowed, horizontally laminated and ripple bedded, shaly sandstone
4	Bioturbated siltstone - thin, intensely burrowed, shaly siltstone
5	Central bar facies - cross bedded, medium to fine grain sandstone with scattered gravel size clasts of sideritic mudstone and minor laminae of shale

Rock Type #	Description
0	Undefined
1	Fine grain sandstone; shale volume < 3%
2	Fine to medium grain sandstone, shale volume < 5%; 3-12 % glauconite
3	Fine to medium grain sandstone; shale volume 2-5%; 5-10% glauconite
4	Fine grained sandstone; shale volume 8-20%; 2-8% glauconite
5	Fine grained sandstone; shale volume < 5%; less than 5% glauconite
6	Medium to coarse grain sandstone, 10-25% shale volume; up to 10% glauconite
7	Very shaly and silty, very fine grained sandstone, shale volume 40-60%; 10% glauconite

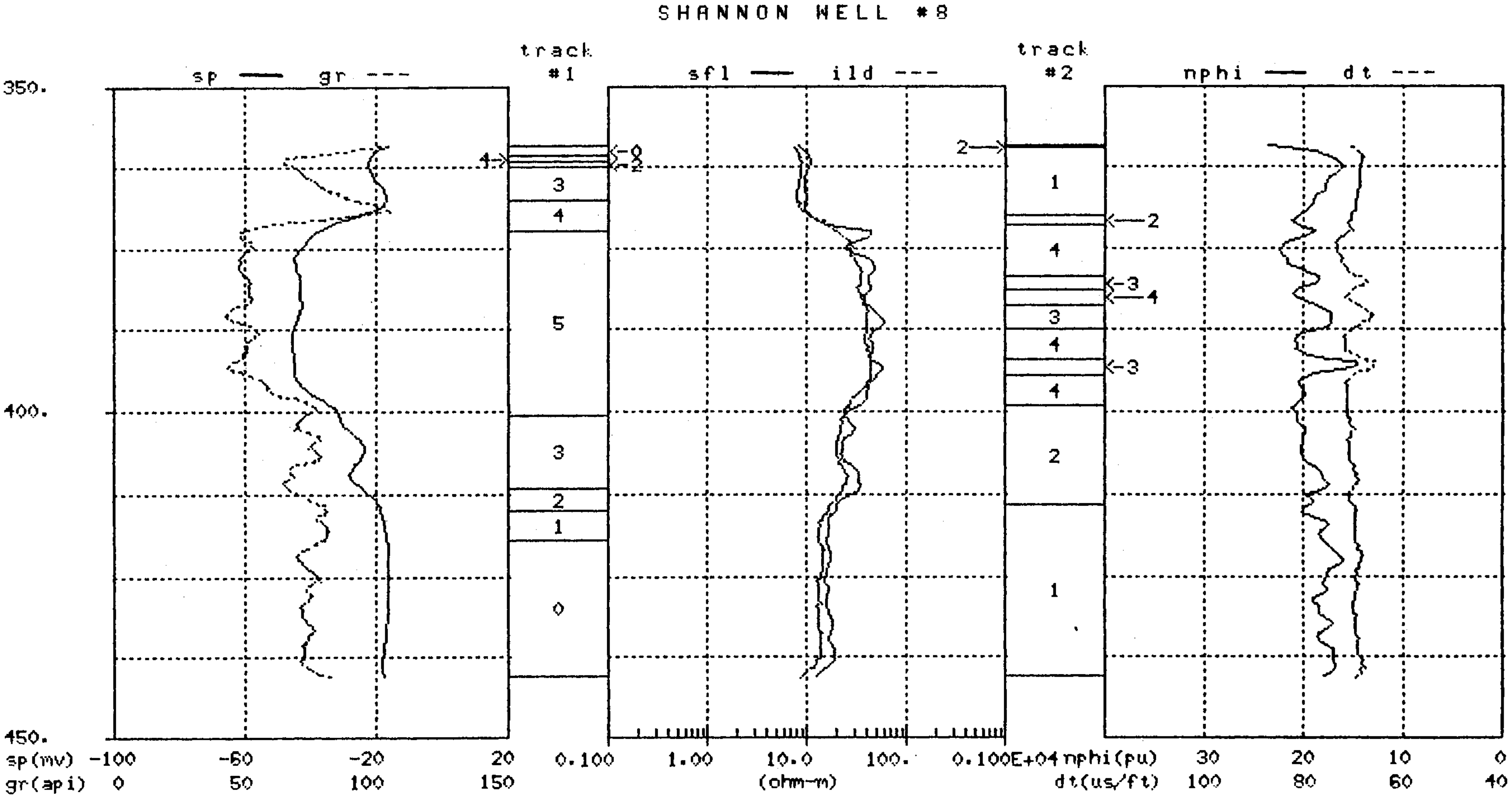
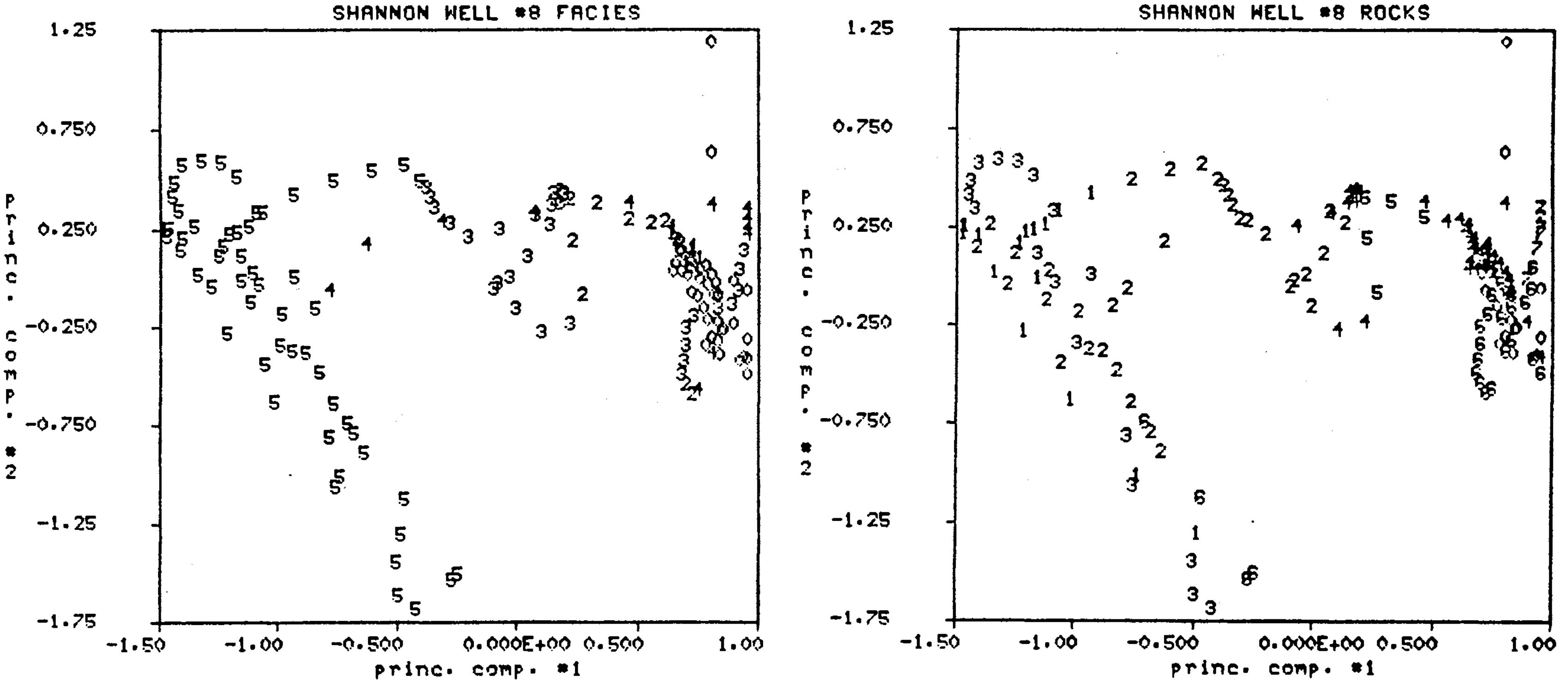


Figure 50. Comparison of Geologic Facies(track #1) with Segmentation Algorithm Output (track #2) for the Shannon Sandstone, Well #8



(a)

(b)

Figure 51. Crossplot of (a) Geological Facies and (b) Rock Types for the Shannon Sandstone, Well #8

facies which is captured by the wireline log information and detected by the segmentation algorithm.

Several reasons might be given to help explain why the other facies are not detected by the segmentation process. One reason is that the discrimination between facies types is often made on visual evidence like color or the presence of fossils and this type of information does not necessarily relate directly to any geological parameter captured by the wireline log data. A second reason could be the small sample size of many of the facies. For example, facies #2 in Figure 51(a) has only a few log values associated with it and the FCM algorithm is known to have difficulty isolating small populations even when they have very distinctive log characteristics. A third reason could be the gradational changes from one facies type to the next facies type make it difficult to detect the individual facies types. Whatever the reason, the given facies and rock types tend not to group into clusters detectable by the segmentation algorithm. Figure 51 would suggest that, with the exception of facies #5(central bar facies), the given facies and rock type information is not readily extracted from the wireline log data. It is rather interesting that the Faciolog "electrofacies" (Figure 49, track #2) were generated using similar tools as the segmentation algorithm in this work,(PC analysis and cluster analysis), and yet the resulting borehole segments are significantly different from the borehole segments for Well #8 in Figure 48. Also, the Faciolog results do not seem to relate very well to the geological facies for the same interval.

4.4 Nearest Prototype Classifiers

The problem of classification is basically one of partioning the feature space into regions, one for each class of data. There exist a variety of approaches to the classification problem that can be broadly categorized as either Bayesian or non-Bayesian. At this point either approach is viable. The Bayesian approach is attractive in the sense that it is statistically optimum with respect to the mean square error provided the distribution of the data is known. In practice, the distribution of the data is seldom known and in this case no assumptions have been made with regard to the distrubution of the data. The best that can be done in this case is an approximate Bayesian classifier using statistical estimates taken from a training set of data. However, there is no theoretical basis that assures that an approximate Bayesian classifier will out perform a classifier of non-Bayesian design. This fact coupled with the fact that the segmentation algorithm lends itself very nicely to the design of a non-Bayesian nearest prototype classifier leads to the following discussion on the design and testing of a nearest prototype classifier.

This section discusses the application of the segmentation algorithm in the design and testing of a nearest prototype classifier. The digitized logs for the Shannon Sandstone interval in Wells #6, #7 and #8 comprise the training set from which the classifier is designed. The 'prototypes' for the classifier are generated by running the segmentation algorithm on the log data in the training set. This was done in Section 4.2. See Figure 45 for the correspondence between segments for the three wells in the training set. The classifier design is tested in two steps. The first step applies the classifier to the training set data and the second step applies the classifier to log data outside the training set. The segmentation algorithm results will be used to evaluate the relative performance of the classifier in both steps of testing.

The results of Section 4.2 suggest that there are six different classes of log information for the Shannon Sandstone interval. Not all six classes are present in each well, in fact, only Well #7 has all six classes, Well #6 has five classes and Well #8 has four of the six classes of log information. Principal component prototypes for these six classes of data are listed in Table XV and plotted in Figure 52. These prototypes are simply the respective cluster centers determined by the segmentation algorithm. Notice in Figure 52 that class #2 has a single prototype, class #5 has two prototypes and the remaining classes have three prototypes. Figure 52 also relates position in the two dimensional PC space to the physical parameters of permeability(SP), porosity (NPHI, RHOB and DT), and radioactivity(GR). A relative indication of permeability, porosity and radioactivity can be obtained by doing a perpendicular projection of a given point in the two dimensional PC space to each of the three lines representing permeability, porosity and radioactivity. Figure 52 is very helpful in translating position in the PC space into physical meaning. For example, if one compares the protypes of class #3 to those of class #5 , it is easily observed that class #3 is less permeable, more radioactive and, in general, less porous than class #5. It should be noted that porosity is the least discriminating of the three physical variables.

Nearest protoype classifiers are conceptually very simple. The classifier computes the distance from a pattern $\underline{X}$ of unknown classification to the protoypes of each class and assigns $\underline{X}$ to the class to which it is closest. In this application, the nearest prototype classifier uses the protypes listed in Table XV and the Euclidean metric to measure distance in the two dimensional principal component space. Inputs to the classifier are PC logs #1 and #2 and the classifier outputs the class to which each sample point belongs based on a minimum distance criteria.

TABLE XV.

SHANNON SANDSTONE PRINCIPAL COMPONENT PROTOTYPES
FOR: WELL #6, WELL #7 AND WELL #8

Prototype	pc #1	pc #2
1	1.330	0.053
3	0.472	0.180
4	0.133	–0.975
5	–0.510	–0.002
6	–1.260	0.061

(a) Well #6

Prototype	pc #1	pc #2
1	1.28	0.079
2	0.330	–1.420
3	0.307	0.054
4	–0.041	–0.455
5	–0.535	0.130
6	–1.08	0.012

(b) Well #7

Prototype	pc #1	pc #2
1	0.755	–0.013
3	0.001	0.326
4	–0.641	–1.050
6	–1.190	0.171

(c) Well #8

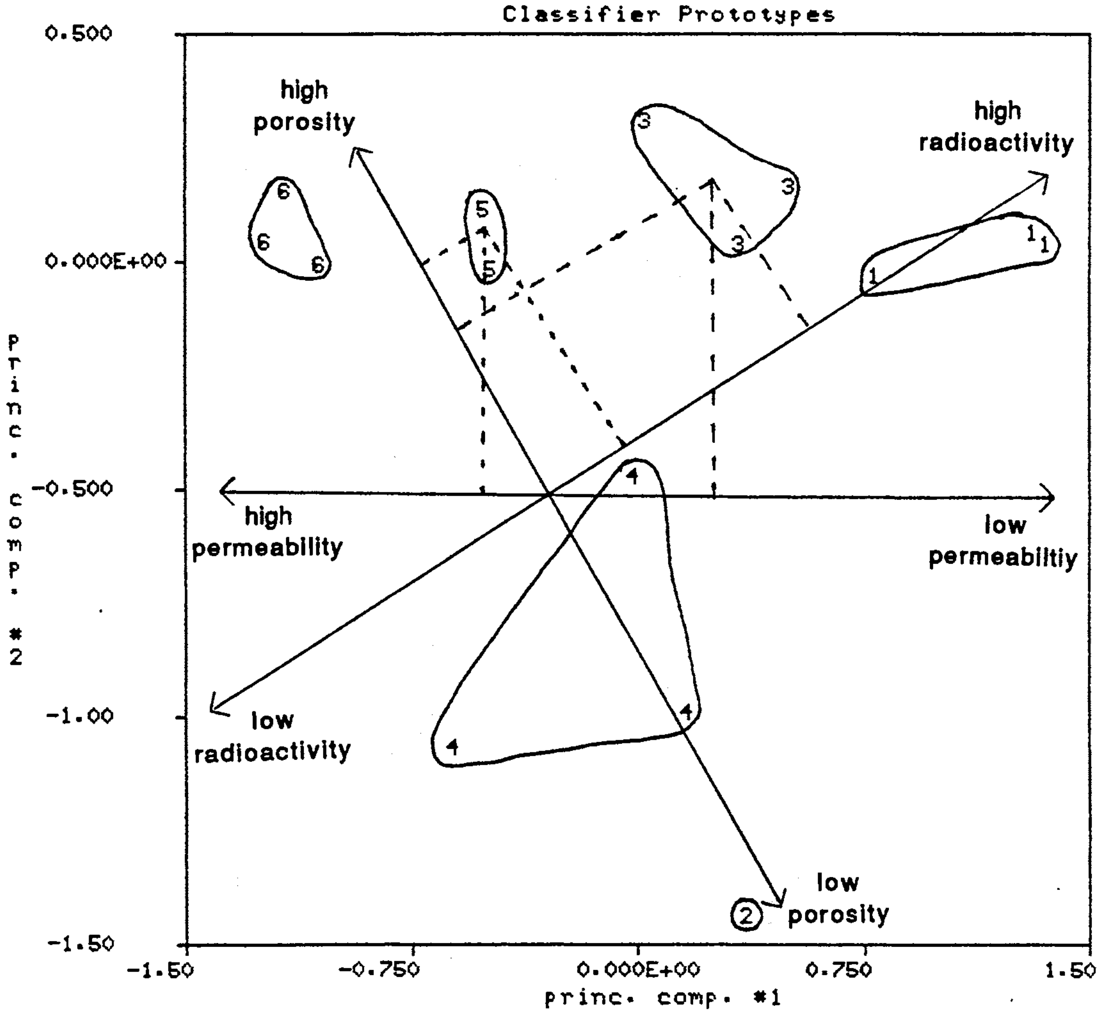


Figure 52. Principal Component Prototypes for the Shannon Sandstone from Table XV

The classifier performance is first tested by applying the classifier to the training set data. Performance is evaluated by comparing the classifier output to the segmentation algorithm results for the same data. Figure 53 displays the segmentation results(track #1) and the classifier results(track #2) for Shannon Well #6. There is generally good agreement between the two results with the classifier differing on 14 of the 164 or 8.54% of the points classified. In a similar fashion, Figures 54 and 55 display the classifier results for Wells #7 and #8.respectively. The classifier differs from the segmentation algorithm results on 5.06% of the points in Well #7 and 15.8% of the points in Well #8. Overall there is 90% agreement between the two results and where there are differences they are minor differences. It should be noted that the majority of the differences involve classes #3 and #5.

In some instances the classifier yields a more probable result than the segmentation algorithm. For example, consider the classifier result for Well #6 at 345-349 ft. in Figure 53. The segmentation algorithm(track #1) shows this interval to be class #4 while the classifier result shows class #2 for part of this interval that coincides with a noticeable decrease in the porosity logs. A quick reference to Figure 52 indicates this is a very reasonable classification since porosity is one of the main discriminating factors between class #2 and class #4.

Now consider classifier performance on log data outside the training set. Log data from Shannon Well #3 is used to illustrate the performance of the classifier on log data outside the training set. In order to maintain a similar method for evaluating classifier performance, the segmentation algorithm is also applied to the log data from Well #3. A procedure exactly analogous to the one used is Section 4.2 is used for Shannon Well #3. All algorithm parameters are identical to those used for Wells #6, #7 and #8. Figure 56 shows the segmentation results for a relatively large 250 ft. interval of Well #3. The Shannon Sandstone interval is the actual interval of interest and is marked by segments #2, #3 and #4 and lies between 291-344 ft. Following the procedure of Section 4.2, the segmentation algorithm is applied to the smaller interval of interest and Table XVI lists the cluster validity measures for this interval. Primary data structure is judged to exist for c = 5 with secondary data structure for c = 8. Figure 57 compares the segmentation algorithm results for c = 5 to the classifier results for the same interval. Notice that both results indicate the presence of five of the six classes of log information in the Shannon Sandstone interval for Well #3. Class #2 is not detected by either the classifier or the segmentation algorithm. The classifier results differ from the segmentation algorithm results on 22 of the 106 or 20.75% of the data points in the interval. The majority of the differences exist in the interval 310-325 ft. and

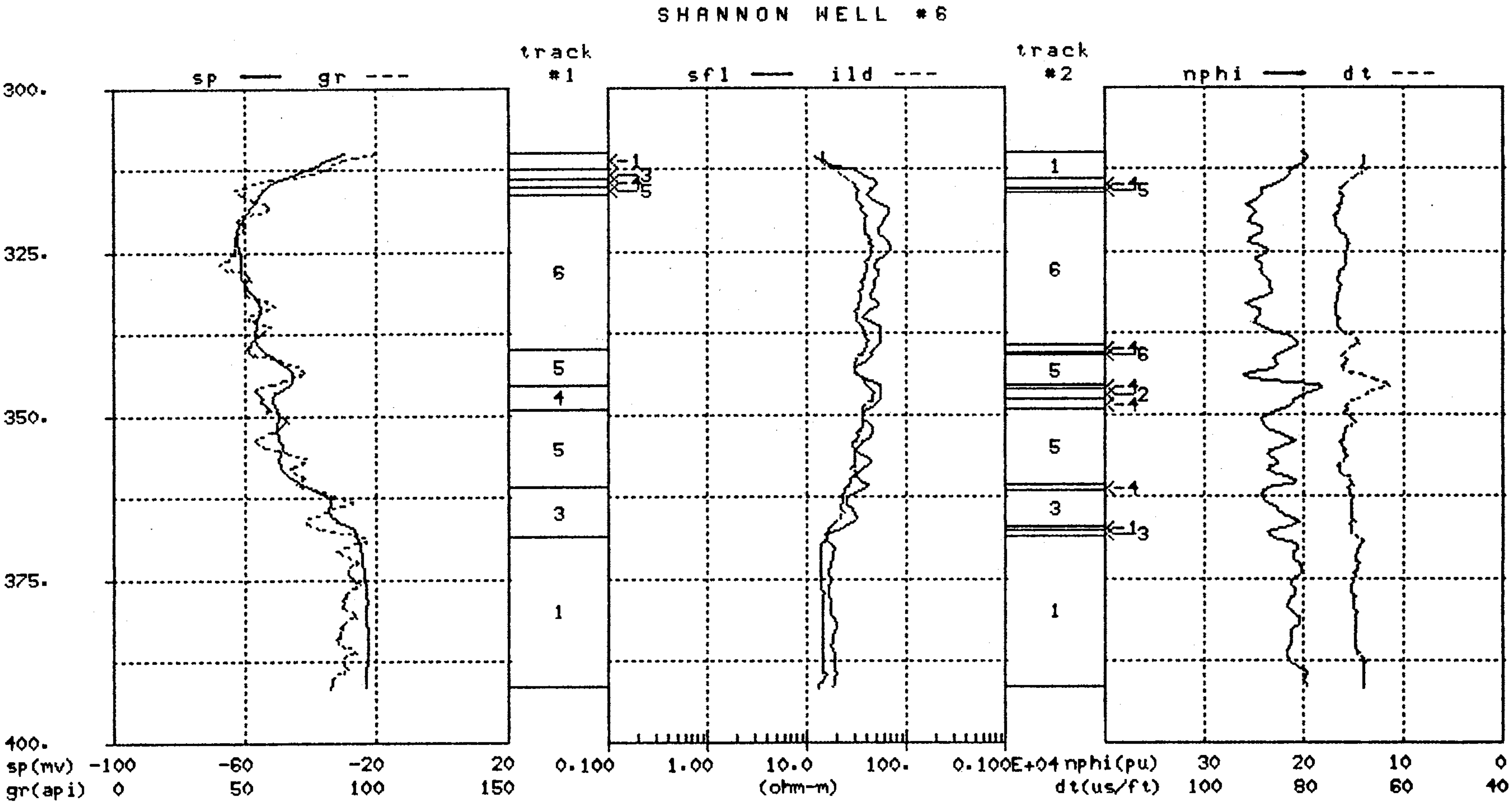


Figure 53. Comparison of Segmentation Results(track #1) and the Classifier Results(track #2) for the Shannon Sandstone, Well #6

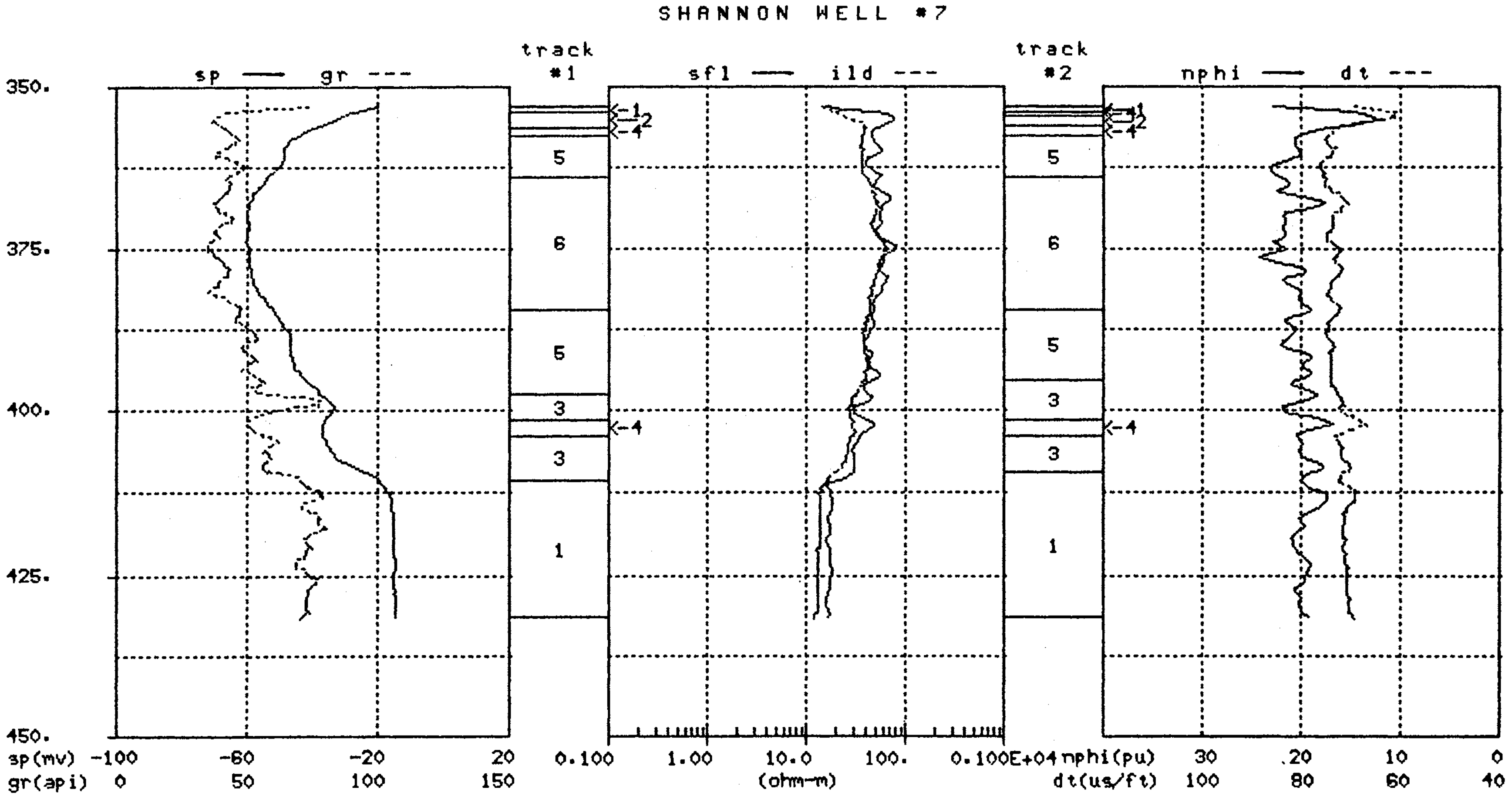


Figure 54. Comparison of Segmentation Results(track #1) and the Classifier Results(track #2) for the Shannon Sandstone, Well #7

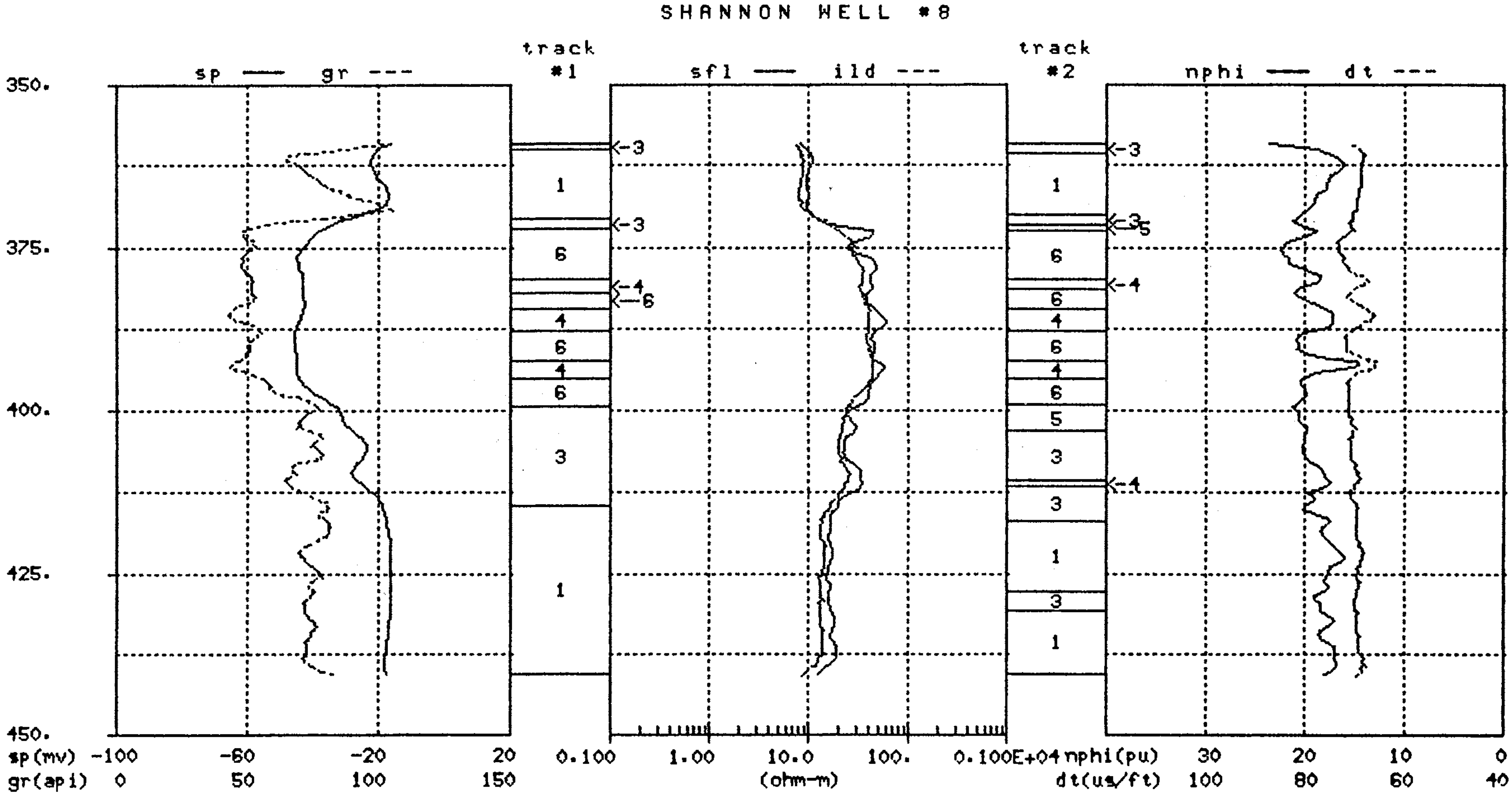


Figure 55. Comparison of Segmentation Results(track #1) and the Classifier Results(track #2) for the Shannon Sandstone, Well #8

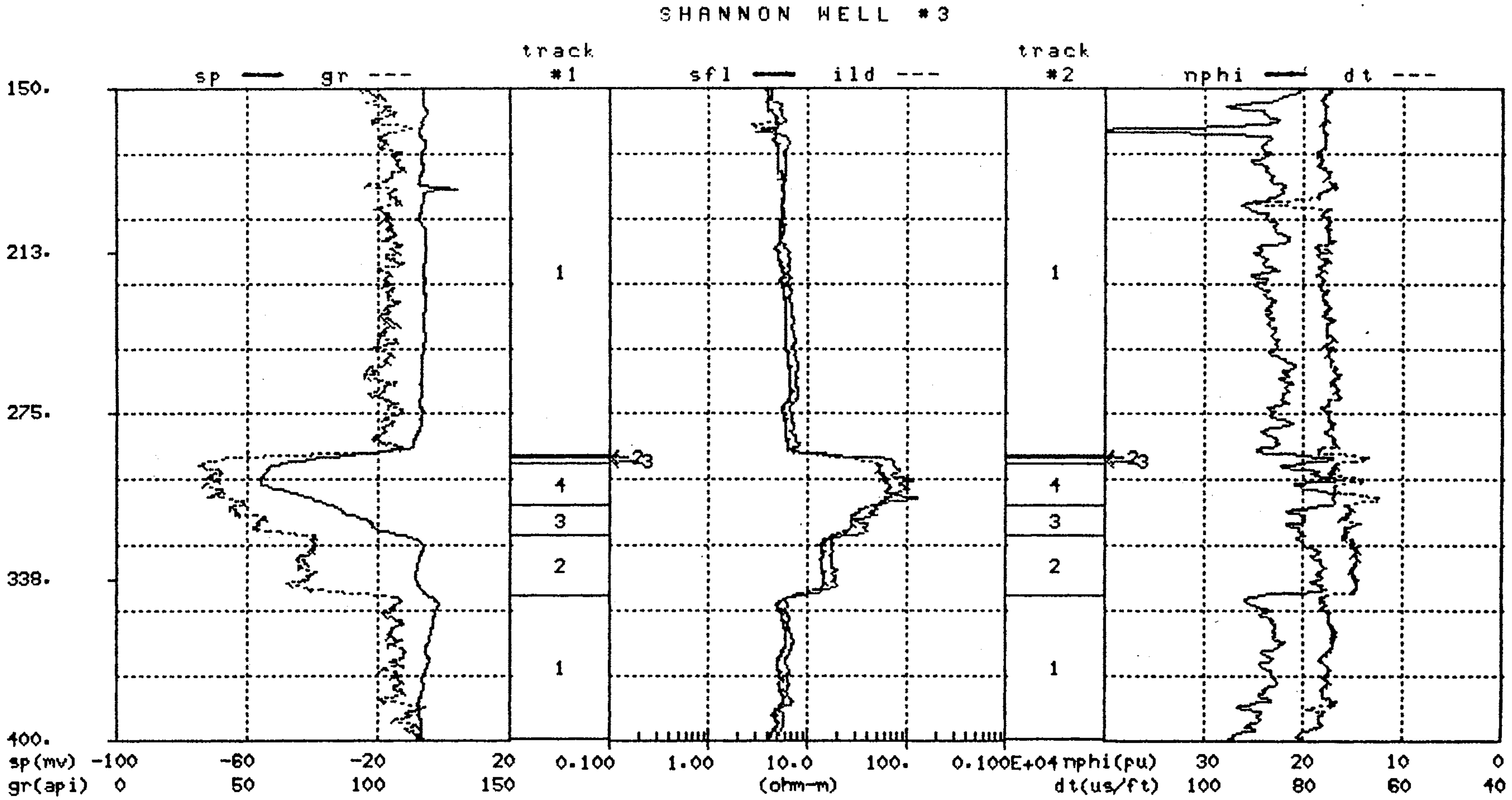


Figure 56. Segmentation Algorithm Output for Well #3

TABLE XVI.

FCM VALIDITY MEASURES FOR SHANNON SANDSTONE, WELL #3 WITH EUCLIDEAN NORM, $\varepsilon = 0.01$ AND $m = 1.5$

c	F	G	ΔJ
2	0.911	–0.278	2.561
3	0.869	–0.820	3.558
4	0.915	–0.667	1.159
5	0.912	–0.423	0.845
6	0.904	–0.778	0.718
7	0.896	–0.569	0.598
8	0.898	–0.569	0.527
9	0.855	–1.194	0.590
10	0.848	–1.276	0.544

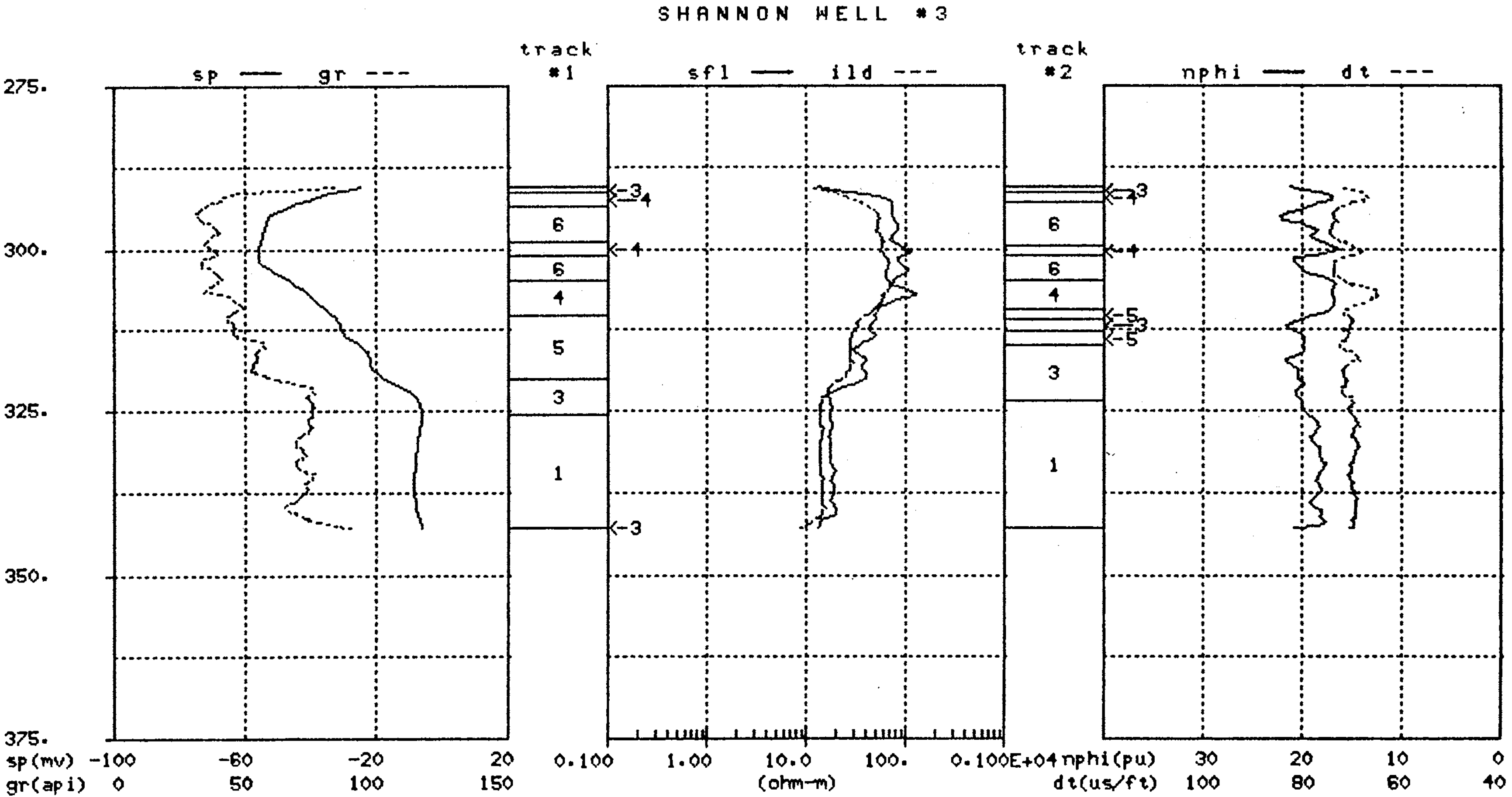


Figure 57. Comparison of Segmentation Results(track #1) and the Classifier Results(track #2) for the Shannon Sandstone, Well #3

involve the boundaries between classes #3~and #5. It is not possible to state absolutely which segmentation of the Shannon Sandstone in Figure 57 is best, but it is the opinion of this observer that the segmentation algorithm(track #1) is the better result. This is based on the following observations. Recall the observation made earlier in this section that class #3 is more radioactive and less permeable than class #5. It is not readily apparent from the GR and SP signal traces in Figure 57 that any such change occurs at a depth of 310 ft., therefore the classifier result is suspect at this point. Also, the segmentation algorithm(track #1) boundary between class #5 and class #3 at 320 ft. coincides with an obvious transition in both the GR and SP signals and no similar boundary exists for the classifier result in track #2.

The classifier result for Well #3 indicates that the classifier can operate reliably on data outside the training set, but may not provide as consistent a classification of the log data as the segmentation algorithm itself. However, in the case of Well #3 the classifier result is certainly good enough to warrant its use over the segmentation algorithm simply because it is simple and easy to use. The classifier uses an automated one pass process to classify the log data, whereas the segmentation algorithm uses an iterative approach, requires the interpretation of cluster validity measures and requires the manual correlation of segments between wells. More extensive testing of the classifier is needed to better understand its reliability in identifying different classes of data in wells outside the training set.

4.5 Chapter Summary

The multiwell example in Section 4.2 establishes the feasibility of determining segments between wells with similar wireline log characteristics and correlating the segments manually by using the vertical position of the segments within the borehole and the spatial distribution of the maximum membership clusters. The segments determined by the segmentation algorithm have a marginal relationship to the geological facies within the Shannon Sandstone interval but certainly no worse than the relationship between the geological facies and the Faciolog "electrofacies". The shelf silty shale facies and central bar facies are the only two facies which exhibit a consistent relationship to the borehole segmentation results. The fact that the output of the segmentation algorithm exhibits only a marginal relationship to the geological facies should not discount the utility of the algorithm. The geological basis for discrimination between facies types is often dependent upon visual evidence found in the core and this visual evidence may or may not relate to the geological parameters effecting the logs. Conversely, the wireline logs may be responsive to geological parameters that may be bypassed or overlooked in conventional core analyses. If one accepts the premise that neither the log information or the core information incorporates all of the geological information for a given interval, then it may be prudent to merge the two descriptions looking for a more comprehensive description of the given interval. For this reason, design of a classifier to reliably indentify segments between wells with similar wireline log characteristics is still potentially useful.

The segmentation algorithm results of Section 4.2 indicated six classes of log information exist for the Shannon Sandstone interval. Figure 52 shows a relative comparison of these six classes in terms of porosity, permeability and radioactivity. The classes range from class #1 which has relatively high radioactivity, low permeability and moderate porosity, to class #6 which has relatively high porosity, high permeabiltiy and low radioactivity. The respective principal component cluster centers are used as prototypes in the design of a nearest prototype classifier. When tested on the training set data the classifier performed reliably agreeing with the segmentation algorithm results on 90% of the points classified. There were no major differences between the two results. The classifier was tested on a well outside the training set with 79% agreement between the classifier results and the segmentation algorithm results. Further testing of the classifier is necessary to determine its ability to correctly classify log data outside the training set. One obvious application for the nearest prototype classifier is the automatic correlation of segments from well to well.

CHAPTER V

SUMMARY AND EXTENSIONS

This study provides evidence that principal components analysis and fuzzy objective function clustering algorithms can be applied in the analysis of wireline log data. These pattern recognition tools are used to develop a segmentation algorithm that relates segments with similar wireline log characteristics within a single well or among multiple wells. In a multiple well setting, a training set of log information may be used to design a nearest prototype classifier for the automatic recognition of similar segments in nearby wells not necessarily in the training set.

The interpretation of the output of the borehole segmentation model has been primarily lithological in nature. This interpretation approach is used because of the availability of lithology/facies information corresponding to the wireline log data used in this study. Output of the segmentation model generally has a good physical interpretation except in the cases of thin beds and transition zones. The evidence in this study supports the premise that the basic lithologic information in the original wireline logs is carried by the first two PC logs formed from linear combinations of the original logs. The best interpretation results were obtained when the influence of the resistivity logs, on the calculation of the PC logs, was limited. Although it was not done in the present model, the input logs could be weighted, so that the user has some control over the linear combination of the inputs that form the PC logs. In the model's present form, weighting of the inputs is done automatically by a linear scaling of the original wireline logs. There has been no attempt to use the existing model to perform any type of reservoir analysis. In the event such a goal is pursued, the model is general enough to allow derived logs, such as percentage shale volume, effective porosity and water saturation to be used as inputs.

Any further investigation in the area of borehole segmentation models should concentrate in two areas. The first is to deconvolve the logging tool response from the wireline log data to obtain log values more representative of the geologic formation cut by the borehole. This would help reduce the problem that the existing model has with thin beds and transition zones, assuming that the deconvolution problem can be solved. The second area involves the development of another model which would use the Bayesian approach to unsupervised classification. This approach involves a computer implementation called Autoclass[15]; which determines the most probable number of classes present in real-valued or discrete data, the most probable description of those classes and each sample's probability of membership in each class. This approach has yielded good results on some standard test data sets, including Anderson's iris data. The Bayesian approach would provide an interesting comparative study to this study.

One other problem of general interest that merits further investigation is the problem of cluster validity. The concept of cluster validity is fundamental to the clustering problem. The present model uses three measures of validity that require an interpretation on the part of the user to determine which clustering is 'best'. Although guidelines have been established for the interpretation of the validity measures, a more objective means of determining validity is desired.

REFERENCES

1. Aitchinson, J., "Reducing the Dimensionality of Compositional Data Sets." Mathematical Geology, V. 16, No. 6, 1984, pp. 617-635.

2. Almon, William R., Statistical Differentiation of Sedimentary Facies on the Basis of Wireline Log Responses, Shannon Sandstone, Hartzog Draw Field, Wyoming, Masters Thesis, University of Tulsa, 1980.

3. Asquith, George, Basic Well Log Analysis for Geologists. American Association of Petroleum Geologists, 1982.

4. Backer, E., Cluster Analysis by Optimal Decomposition of Induced Fuzzy Sets. Delft Univ. Press, 1978.

5. Basic Open Hole Seminar, Gearhart Industries, Inc., 1981.

6. Bateman, Richard, Log Quality Control, International Human Resources Development Corporation, 1985.

7. Beauchamp, J. J., et. al., "Application of Discriminant Analysis and Generalized Distance Measures to Uranium Exploration." Mathematical Geology, V. 12, No. 6, 1980, pp. 539-558.

8. Bedwell, John L., Textural Parameters of Clastic Rocks from Borehole Measurements and Their Application in Determining Depositional Environments, Ph.D. Thesis, Colorado School Of Mines, 1974.

9. Bezdek, James C., "Numerical Taxonomy with Fuzzy Sets." Journal of Mathematical Biology, V. 1-1, 1974, pp. 57-71.

10. Bezdek, James C., "Partition Structures: A Tutorial." The Analysis of Fuzzy Information, CRC Press, 1987, V. 3, ch. 6.

11. Bezdek, James C., Pattern Recognition with Fuzzy Objective Function Algorithms, Plenum Press, 1981.

12. Bezdek, James C., Hathaway, R.J., Sabin, M.J. and Tucker, W.T., "Convergence Theory for Fuzzy c-Means: Counterexamples and Repairs." IEEE Transactions on Systems,Man and Cybernetics, Sept./Oct. 1987, pp. 873-877.

13. Carloss, James C., Depositional Environments From Borehole Measurements Lower Cretaceous, Peoria Field and Adjacent Areas, Arapahoe County, Colorado, Masters Thesis, Colorado School of Mines, 1974.

14. Cartinhour, Jack W., Pattern Recognition for a Class of Warped Waveforms with Applications to Well Log Signature Recognition, Ph.D. Thesis, Oklahoma State University, 1987.

15. Cheeseman, Peter, et.al., "Bayesian Classification." Proceedings of the 7th National Conference on Artificial Intelligence, Sponsored by American Association of A. I., 1988.

16. Chien, Y. T. and Fu, K. S., "On the Generalized Karhunen-Loeve Expansion." IEEE Transactions on Information Theory, July 1967, pp. 518-520.

17. Collyer, P. L. and Merriam, D. F., "An Application of Cluster Analysis in Mineral Exploration." Mathematical Geology, V. 5, No. 3, 1973, pp. 213-223.

18. Davis, John C., Statistics and Data Analysis in Geology, John Wiley and Sons, 1986.

19. Drapeau, George, "Factor Analysis: How It Copes with Complex Geological Problems. "Mathematical Geology, V. 5, No. 4, 1973, pp. 351-363.

20. Duda, R. and Hart, P., Pattern Classification and Scene Analysis, John Wiley and Sons, 1973.

21. Einstein, Bruce A. and Fehlauer, John, "Signal Processing for Feature Extraction and Pattern Recognition." IEEE International Conference on Acoustic, Speech,and Signal Processing, 1976.

22. Fischer, R. A., "The Use of Multiple Measurements in Taxonomic Problems." Annals of Eugenics, V. 7, 1936, pp. 179-188.

23. Friedman, H. and Rubin, J., "On Some Invariant Criteria for Grouping Data." Journal of Am. Stat. Assoc., V. 62, 1967, pp. 1159-1178.

24. Gill, D., "Application of a Statistical Zonation Method to Reservoir Evaluation and Digitized Log Analysis." Am. Assoc. Petroleum Geologists Bulletin, V. 54, No. 5, 1970, pp. 719-729.

25. Glossary of Terms and Expressions Used in Well Logging, Society of Professional Well Log Analysts, 1975.

26. Gonzalez, Rafael C. and Wintz, Paul, Digital Image Processing, Addison-Wesley, 1977.

27. Goodwin, D. W., "Hartzog Draw CO2 Feasibility Study: Petrophysical Analysis." Unpublished Cities Service Report, July 1983.

28. Hawkins, D. M. and Merriam, D. F., "Optimal Zonation of Digitized Sequential Data."Mathematical Geology, V. 5, No. 4, 1973, pp. 389-395.

29. Hawkins, D. M. and Merriam, D. F., "Segmentation of Discrete Sequences of Geologic Data." Geological Soc. America Mem. 142, 1975, pp. 311-315.

30. Hawkins, D. M. and Merriam, D. F., Zonation of Multivariate Sequences of Digitized Geologic Data." Mathematical Geology, V. 6, No. 3, 1974, pp. 263-269.

31. Hawkins, D. M., and ten Krooden, J. A., "A Review of Several Methods of Segmentation." Geomathematical and Petrophysical Studies in Sedimentalogy, Pergamon Press, 1979, pp. 117-126.

32. Hilchie, Douglas W., Advanced Well Log Interpretation, Douglas W. Hilchie, Inc., 1982.

33. Hinch, H. H., "Evaluation and Description of Core from the APC SHADS No. 1 Well." Unpublished AMOCO Report, April, 1987.

34. Hobson, J. P. and Fowler, M. L., "Field-Wide Distribution of Facies, Flow Units, and Geologic Areas, Shannon Sandstone, Hartzog Draw Field, Wyoming." Unpublished Cities Service Report, Dec. 1985.

35. Hotelling, H., "Analysis of a Complex of Statistical Variables into Principal Components." Journal of Educational Psychology, V. 24, 1933, pp. 417-441 and 498-520.

36. Kendall, M. G., "Discrimination and Classification." Multivariate Analysis, Academic Press, 1966, pp. 165-185.

37. Makhoul, J., Roucos, S. and Gish, H., "Vector Quantization in Speech Coding." Proceedings of the IEEE, V. 73, No. 11, Nov. 1985, pp. 1551-1588.

38. McCammon, Richard B., "Principal Component Analysis and Its Application in Large Scale Correlation Studies." Journal of Geology, V. 74, No. 5, pt. 2, 1966, pp. 721-733.

39. Meyer, B. L., and Nederlof, M. H., "Identification of Source Rocks on Wireline Logs by Density/ Resistivity and Sonic Transit Time/Resistivity Crossplots." American Association of Petroleum Geologists Bulletin, V. 68, 1984, pp. 121-129.

40. Miesch, A. T., "Scaling Variables and Interpretation of Eigenvalues in Principal Component Analysis of Geologic Data." Mathematical Geology, V. 12, No. 6, 1980, pp. 523-538.

41. Parks, James M., "Cluster Analysis Applied to Multivariate Geologic Problems." Journal of Geology, V. 74, No. 5, pt. 2, 1966, pp. 703-715.

42. Patchett, J., "The Determination of the Properties of Clays in Shales from Logs with an Example of One Interpretation Technique." Paper Obtained from Author, AMOCO, Tulsa, OK. 1986.

43. Patterson, David, et. al., "Discriminant Analysis Applied to Aerial Radiometric Data and Its Application to Uranium Favorability in South Texas." Mathematical Geology, V. 13, No, 6, 1981, pp. 535-568.

44. Priisholm, S. and Michelsen, O., "The Use of Porosity Logs in Lithology Determination, Lithostratigraphy and Basin Analysis." Geomathematical and Petrophysical Studies in Sedimentology, Pergamon Press, 1979, pp. 71-79.

45. Ranganathan, V. and Tye, R. S., "Petrography and Reservoir Quality of the Shannon Sandstone, Hartzog Draw Field, Campbell and Johnson Counties, Wyoming." Unpublished Cities Service Report, Oct. 1983.

46. Rao, K. R. and Ahmad, N., "Orthogonal Transforms for Digital Signal Processing." IEEE International Conference on Acoustic, Speech and Signal Processing, 1976.

47. Sabin, Michael J., "Convergence and Consistency of Fuzzy c-Means/Isodata Algorithms." IEEE Trans. on Pattern Analysis and Machine Intelligence, V. 9, No. 5, 1987, pp. 661-668.

48. Sabin, Michael J., and Gray, R.M., "Global Convergence and Empirical Consistency of the Generalized Lloyd Algorithm." IEEE Transactions on Information Theory, March, 1986, pp. 148-155.

49. Savre, W. C., "Determination of a More Accurate Porosity and Mineral Composition in Complex Lithologies with the Use of Sonic, Neutron and Density Surveys." Journal of Petroleum Technology, Sep. 1963, pp. 945-959.

50. Scott, A. and Symons, M., "Clustering Methods Based on Likelihood Ratio Criteria." Biometrics, V. 27, 1971, pp. 387-397.

51. Serra, O., Fundamentals of Well-Log Interpretation, Elsevier, 1984.

52. Shaw, B.R. and Cubitt, J.M., "Stratigraphic Correlation of Well Logs: An Automated Approach."Geomathematical and Petrophysical Studies in Sedimentology, Pergamon Press, 1979, pp. 127-148.

53. Size, William B., "Interpretation of Factor Analysis on Modal Data from the Red Hill Syenitic Complex." Mathematical Geology, V. 5, No. 2, 1973, pp. 191-197.

54. Stewart, Gary F., Professor of Geology, Oklahoma State University, 1988.

55. Testerman, J.D., "A Statistical Reservoir-Zonation Technique." Journal of Petroleum Technology, V. 14, No. 8, 1962, pp. 889-893.

56. Till, R. and Colley, H., "Thoughts on the Use of Principal Component Analysis in Petrogenic Problems." Mathematical Geology, V. 5, No. 4, 1973, pp. 341-350.

57. Tillman, R. W. and Martinsen, R. S., "Hartzog Draw Field, Wyoming: Facies Analysis and Sedimentology." Unpublished Cities Service Report, Dec. 1978.

58. Tou, Julius T. and Gonzales, Rafael C., Pattern Recognition Principles, Addison-Wesley, 1974.

59. Watney, W. Lynn, "Gamma ray-Neutron Crossplots as an Aid in Sedimentological Analysis."Geomathematical and Petrophysical Studies in Sedimentology, Pergamon Press, 1979, pp. 81-99.

60. Well Logging and Interpretation Techniques, Dresser Atlas, Dresser Industries, Inc., 1982.

61. Widdicombe, R. E., Noon, P. and Best, D. L., "Multiwell Faciolog Evaluation, Hartzog Draw Field Powder River Basin, Wyoming." SPWLA 25th Annual Logging Symposium, 1984.

62. Wolfe, J., "Pattern Clustering by Multivariate Mixture Analysis." Multivariate Behavior Research, V. 5, 1970, pp. 329-350.

63. Wolfe, Martin and Pelissier-Combescure Jacques, "Faciolog-Automatic Electrofacies Determination." SPWLA 23rd Annual Logging Symposium, 1982.

APPENDIX

This detailed core description is supplied by AMOCO of Tulsa, Oklahoma and is compiled from an unpublished AMOCO report prepared by H. H. Hinch [33]. This information is used in Chapter III to evaluate the physical significance of the segmentation algorithm output.

Depth Interval(ft.)	Lithology	Description
2.0-6.0	Shale(#1)	Black, organic rich, contains high angle mineralized fractures.
6.0-29.0	Limestone(#1)	Gray to tan, stylolitic, dense to finely crystalline, argillaceous, grading into very calcareous fossiliferous shale in both the top two feet and the bottom 1ft
	Petroliferous(#1)	(1.5 net ft. of oil showing in 6.5 ft.of core.) Oil occurs discontinuously in apparently isolated limestone vugs from17.3-17.7 ft. and from 22.4-23.1 ft. Also, oil occurs in a high angle fracture from 24.1-24.3 ft. (Oil is indicated by both a light tan stain and yellow fluorescence.)
29.0-32.9	Shale(#2)	Black, organic rich, vertical fractures from 31.0-33.0 ft.
32.9-42.6	Limestone(#2)	Gray to tan, argillaceous, finely crystalline, Interbedded with calcareous, dark gray shale(38.8-39.8 ft.), grading into calcareous fossiliferous shale in both the top 2.5 ft. and the bottom 1 ft.
	Petroliferous(#2)	(4.0 net ft of oil showing in 4.0 ft. of core.) Oil occurs discontinuously in limestone pores adjacent to a high anglefracture from 34.3-38.3 ft. (Oil indicated by both tan stain and light yellowfluorescence.) Orange mineral fluorescence occurs in the more argillaceous beds from 37.2-42.6 ft.
42.6-44.8	Sandstone(#1)	Light gray, very argillaceous, carbonaceous micaceous, contains pyritized plant fragments.

Depth Interval(ft.)	Lithology	Description
44.8-45.6	Shale(#3)	Black, carbonaceous, grading into thin argillaceous sandstone at 45.4-45.6 ft.
45.6-46.7	Coal(#1)	
46.7-49.4	Sandstone(#2)	Light gray, very argillaceous, micaceous, grading downward into gray shale.
49.4-51.7	Missing core	
51.7-52.3	Shale(#4)	Gray, very calcareous, fossiliferous, arenaceous, pyritic, bioturbated.
52.3-59.2	Shale(#5)	Gray, pyritic, arenaceous in top 1 ft.
59.2-66.3	Sandstone(#3)	Light gray, ripple laminated, micaceoussandstone, interlaminated with dark gray shale. Sandstone is best developed (>70%)between 63.0 ft. and 64.9 ft.
66.3-68.0	Missing core	
68.0-75.3	Sandstone(#4)	Dark gray, very argillaceous, micaceous carbonaceous, interbedded with dark gray arenaceous shale.
75.3-77.5	Sandstone(#5)	Dark gray, very argillaceous, carbonaceous, micaceous, interlaminated with(1-10 mm. thick) light gray, less argillaceous bioturbated sandstone containing pyritized plant fragments.
77.5-79.2	Missing core	
79.2-99.9	Shale(#6)	Dark gray.
99.9-101.5	Missing core	
101.5-107.3	Limestone(#3)	Dark gray to black, very argillaceous, very fossiliferous grading into very calcareous shale at bottom. A high angle mineralized fracture occurs from 377.0-379.5 ft.

Depth Interval(ft.)	Lithology	Description
107.3-111.5	Limestone(#4)	Dark gray to tan, very fossiliferous, very argillaceous.
111.5-114.7	Shale(#7)	Black, organic rich, contains small phosphatic nodules.
114.7-141.5	Shale(#8)	Gray, slightly pyritic, containes siderite nodules from 129.5-141.5 ft.
141.5-143.0	Coal(#2)	
143.0-144.0	Sandstone(#6)	Tan, very argillaceous, carbonaceous at top, calcareous and fossiliferous toward bottom, grading downward into green slickensided bioturbated shale which contains small limestone nodules.
144.0-144.5	Shale(#9)	Green, abundant slickensides.
144.5-145.5	Sandstone(#7)	Tan, very argillaceous, calcareous, fossiliferous, containing small limestone nodules.
145.5-146.4	Shale(#10)	Green to gray, arenaceous at top and bottom.
146.4-151.6	Sandstone(#8)	Light gray, coarse grained, micaceous, cross bedded (20^o to 30^o dips), abundant shale partings.
	Petroliferous(#3)	(5.4 net feet of oil showing in 9.3 ft. of core from 146.4-155.7 ft.) Oil occurs continuously in (coarse grained) sandstone pores from 146.4-146.8 ft., from 146.9-148.3 ft., from 148.5-149.3 ft. and from 149.4-151.6 ft. (Oil indicated throughout the total interval by both light tan stain andyellow fluorescence.)
151.6-153.7	Sandstone(#9)	Light gray, fine grained, ripple laminated, micaceous, interlaminated with 1-10 mm. thick dark gray shale (>70% sandstone).

Depth Interval(ft.)	Lithology	Description
	Petroliferous(#4)	Oil occurs continously from 151.8-151.9 ft. in the pores of a thin coarse grained sandstone that has the same characteristic as the sand from 146.4-151.6 ft.
153.7-154.1	Sandstone(#10)	Light gray, coarse grained, micaceous, cross bedded, abundant shale partings.
	Petroliferous(#5)	Oil occurs in sandstone pores from 153.9-154.1 ft.
154.1-154.7	Sandstone(#11)	Light gray, fine grained, ripple laminated, micaceous, interlaminated with 1-10 mm. thick dark gray shale (>70% sandstone).
154.7-155.7	Sandstone(#12)	Light gray, coarse grained, micaceous cross bedded, shale partings.
	Petroliferous(#6)	Oil in sandstone pores from 154.7-154.8 ft. and from 155.5-155.7 ft.
155.7-157.5	Sandstone(#13)	Light gray ripple laminated, micaceous,interlaminated with 1-10 mm. thick dark gray shale.
157.5-164.5	Shale(#11)	Dark gray, (interlaminated with a very minor amount (<10%) of 1-5mm. thick ripple laminated light gray, micaceous sandstone.
164.5-182.0	Shale(#12)	Light gray, micaceous, locally arenaceous. Tan siderite nodules occur from 176.7- 178.5 ft.
182.0-183.2	Shale(#13)	Black, calcareous, fossiliferous (pyritized brachiopods).
183.2-187.6	Shale(#14)	Light gray, micaceous, silty, contains tan siderite nodules.
187.6-197.9	Shale(#15)	Dark gray, interlaminated with <20% light gray, 1-5 mm. thick micaceous, ripple laminated, slightly bioturbated sandstone which contains tan siderite nodules.

Depth Interval(ft.)	Lithology	Description
197.9-212.0	Sandstone(#14)	Light gray, 0.1-0.7 ft. thick beds of medium to coarse grained, argillaceous, micaceous, carbonaceous, nodular(sideritic) sandstone,interbedded with dark gray shale that is interlaminated with light gray 1-5 mm. thick micaceous, ripple laminated, slightly bioturbated sandstone (>75% sandstone at top grading downward gradually to ≈50% sandstone at bottom).
212.0-224.0	Shale(#16)	Dark gray, interlaminated with light gray, 1-5 mm. thick micaceous, ripple laminated, slightly bioturbated sandstone (≈50% sandstone at top grading downward gradually to <10% sandstone at 219.5 ft., grading downward further to 100% shale at 224.0 ft.
224.0-224.7	Coal(#3)	
224.7-227.1	Missing core	
227.1-227.9	Sandstone(#15)	Light gray, silty, slightly carbonaceous,argillaceous.
227.9-232.3	Sandstone(#16)	Light gray, micaceous, argillaceous, cross bedded with ripple laminations at bottom.
232.3-233.3	Shale(#17)	Light gray, high clay content, numerous slickensides. Shale grades downward gradually in underlying sandstone.
233.3-241.0	Sandstone(#17)	Light gray, argillaceous, micaceous, very carbonaceous, interlaminated with dark gray shale. Shale percentage increases gradually downward to 100% at bottom.
241.0-244.5	Shale(#18)	Dark gray, grading downward gradually into black shale.

Depth Interval(ft.)	Lithology	Description
244.5-255.5	Shale(#19)	Black, carbonaceous, numerous slickensides. Partially mineralized high angle fractures occur from 245.5-246.0 ft. and from 247.5-248.2 ft.
255.5-267.1	Shale(#20)	Gray. Contact with underlying sandstone is abrupt.
267.1-310.8	Sandstone(#18)	Light gray to tan, micaceous, carbonaceous cross bedded (low angle at top, high angle in middle), ripple laminated at the bottom.
310.8-320.9	Shale(#21)	Dark gray at top, grading gradually downward to black, organic rich, pyritic shale at bottom. A high angle fracture occurs at 317.5 ft.
320.9-322.1	Shale(#22)	Black, fossiliferous, very calcareous, grading gradually downward into underlying limestone.
322.1-327.0	Limestone(#5)	Gray, dense, argillaceous, fossiliferous,grading in shale at bottom.
327.0-328.7	Shale(#23)	Black, fossiliferous, very calcareous.
328.7-331.4	Shale(#24)	Gray to green, high clay content, calcareous, numerous slickensides. Shale contains small limestone nodules.
331.4-332.3	Shale(#25)	Black, fossiliferous, bioturbated. Shale contains small limestone nodules and green shale clasts.
332.3-333.6	Shale(#26)	Green, high clay content, numerous slickensides, bioturbated.
333.6-340.1	Sandstone(#19)	Gray, argillaceous, fine grained, massive. Sandstone grades gradually downward into underlying shale.
340.1-349.5	Shale(#27)	Dark gray, silty. Shale grades gradually downward into underlying black shale(The shale next to a vertical fracture from340.5-342.4 ft. exhibits an unusual green zonation.)

Depth Interval(ft.)	Lithology	Description
349.5-356.1	Shale(#28)	Black, organic rich, slightly fossiliferous(pyritized brachiopods). Shale contains both phosphate and pyrite nodules. Bottom 0.1 ft. is coal which is in abrupt contact with the underlying sandstone.

VITA

Robert Gene Hayes

Candidate for the Degree of

Doctor of Philosophy

Thesis: AN APPLICATION OF PATTERN RECOGNITION PRINCIPLES TO THE ANALYSIS OF WIRELINE LOGS

Major Field: Electrical Engineering

Biographical:

Personal Data: Born in Enid, Oklahoma, August 10, 1955, the son of Emerson and Elouise L. Hayes. Married to Anita J. Murphy on March 17, 1979. Father of two children: Isaac Leigh, 4; and Corrie Elaine, 1.

Education: Graduated from Covington-Douglas High School, Covington, Oklahoma, in May, 1973; received the Bachelor of Science degree in Mathematics from Oklahoma State University in May, 1977; received the Master of Science degree in Electrical Engineering from Oklahoma State University in May, 1978; completed requirements for the Doctor of Philosophy degree in Electrical and Computer Engineering from Oklahoma State University in December, 1989.

Professional Experience: Senior Engineer, Boeing Military Airplane Company, Wichita, Kansas, June, 1978 to June, 1980; Assistant Professor, Electrical Engineering Technology, Oklahoma State University, September, 1982 to June, 1985; Instructor, Electrical and Computer Engineering, Oklahoma State University, June, 1985 to May, 1988.

Awards/Affiliations: Mathematics Achievement Award, member Pi Mu Epsilon, Tau Beta Pi, Eta Kappa Nu, Institute of Electrical and Electronics Engineers and American Society for Engineering Education.