Some Applications of Principal Component Analysis:

Well-to-Well Correlation, Zonation

Istvan Elek
Hungarian Hydrocarbon Institute
Budapest, Hungary

SUMMARY

Principal Component Analysis (PCA) is a multivariate statistical technique often used successfully in various scientific disciplines. This paper aims to show the mathematical principles of PCA and introduce two log analysis applications based on the technique: well-to-well correlation and zonation.

Traditionally well-to-well correlation has been performed using only one log, often a resistivity log or gamma ray log. A better, though usually slower, correlation often can be obtained by using all available wireline logs. This paper describes a method of computing the first principal component, a method which should make correlation easier, more efficient and accurate. The first principal component should contain the largest common part of variances of the input logs. (The Principle Component Analysis technique is described in the Appendix at the end of this paper.)

The second part of this paper deals with a zonation technique based on the first principal component. The technique computes not only boundaries but characteristic values of log responses within a given layer. This computation is based on the ideal case that rock properties are constant within a layer and change suddenly at a layer boundary.

Often in variable groups a common cause-variable or background-variable can be found. In favorable cases the meaning of this common-variable may be easily recognized. However, if the physical meaning of this variable, detected in several examinations, remains unknown, a hypothesis to determine the physical meaning of the background-variable must be established and the veracity of this hypothesis examined.

One benefit of PCA is that the procedure decreases the number of variables. In computing the first principal component, information from the important variables is included while data from unimportant variables are neglected.

WELL-TO-WELL CORRELATION

A thorough well-to-well correlation program ideally includes all the available well log data. Because this can be a very demanding process, interpreters often use only a reduced data set such as the resistivity logs. In many situations data reduction can produce an erroneous or misleading correlation.

A useful compromise between the simple and the complex data sets would be to perform a PCA, an analysis tool which would include all the important well information while leaving the interpreter with a data set much easier to handle.

The PCA technique calls for the interpreter to compute the first principal component for every well, resulting in a dimensionless "log" containing the largest common part of variances of the input logs for each well. The interpreter then has several "logs" suitable for the well-to-well correlation. Previously the interpreter had to assimilate and utilize too many logs; he now needs to perform the pattern recognition necessary for well-to-well correlation with only one "log."

Description of the Technique

The well-to-well correlation program is based on an HP9845/B desktop computer, however the technique can be used on any computer having similar computational and graphics capabilities.

An example of the correlation graphic display is shown in Figure 1. The first principal component for each well is displayed. The zones can be followed from well to well and marked by means of the cursor. These correlations then are stored on a disk or other mass memory device, and the results, such as cross section, displayed on a graphic plotter which can be further enhanced with dipmeter interpretation.

Similar results can be used in seismic interpretation, especially if the input logs are important in the seismic wave propagation (i.e., density, sonic). An example is shown in Figure 2.

ZONATION BY MEANS OF PCA

The identification of rock boundaries is an old problem with well log interpretation, and the lit-

erature on both manual and automatic techniques is extensive. The PCA technique uses only the first principal component log of the input well to identify zones. This layer identification method is not only simple, but fast and reliable. One synthetic and one practical example are given.

Description of the Technique

A flow chart for the technique is in Figure 3. The following steps are included:

- Before processing, determine logs to be included in the PCA and select a minimum thickness of the layers, h_{min}.
- Perform the Principal Component Calculation (see appendix).
- Filter the first principal component with a median filter. The window length of the median filter is equal to h_{min}. The resulting log will be broken (cornered) in some places because of features of this edge-preserving filter (see Figure 4).
- Smooth the broken log to remove angles. The upper limit frequency is related to the window length of the median filter.
- Pick the boundaries at the inflection points (see Figure 5).
- Within each layer, pick a constant characteristic value (c.v.).

— Let the c.v. be the extreme (minimum or maximum) of the smoothed first principal component.

— Let the c.v. be the maximum if the maximum of the smoothed principal component does not occur at the boundary (i.e., the maximum can be found inside the layer).

— Let the c.v. be the minimum if the minimum of the smoothed principal component does not occur at the boundary (i.e., the minimum can be found inside the layer). (See Figure 6).

- So far thin layers have been produced. If thicker units are required, neglect the "weak" boundaries and keep the "strong" boundaries. Define a critical values (EPS) which is dependent on the "strongness" of the boundary. EPS is picked empirically. Presume that the point is in a "big" layer at the i-th data ("i" starts at the beginning of the layer). The average value of the log from the beginning to "i" is equal to "m."

$$m = \frac{1}{i}\sum_{k=1}^{i} a_k$$

where a_k is the long response in the k-th point. Examine the relationship between a_{i+1} and m.

$$\text{if} \quad | a_{i+1} - m | \geq \frac{EPS}{\sqrt{i}}$$

then there is a boundary in the current point.

$$\text{if} \quad | a_{i+1} - m | < \frac{EPS}{\sqrt{i}}$$

then there is not any boundary in the current point. When the critical value (EPS) increases, only "sharp" boundaries occur, and when the

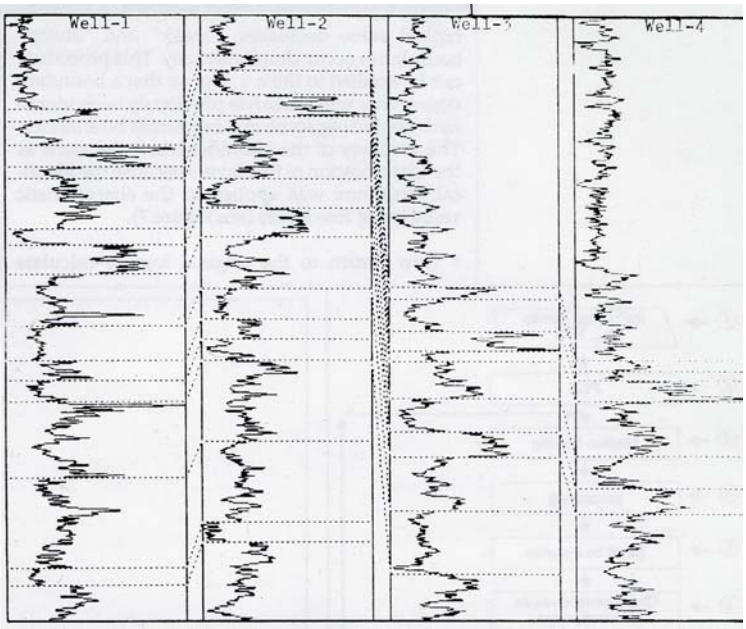


Figure 1: *First principal components of wells to be correlated.*

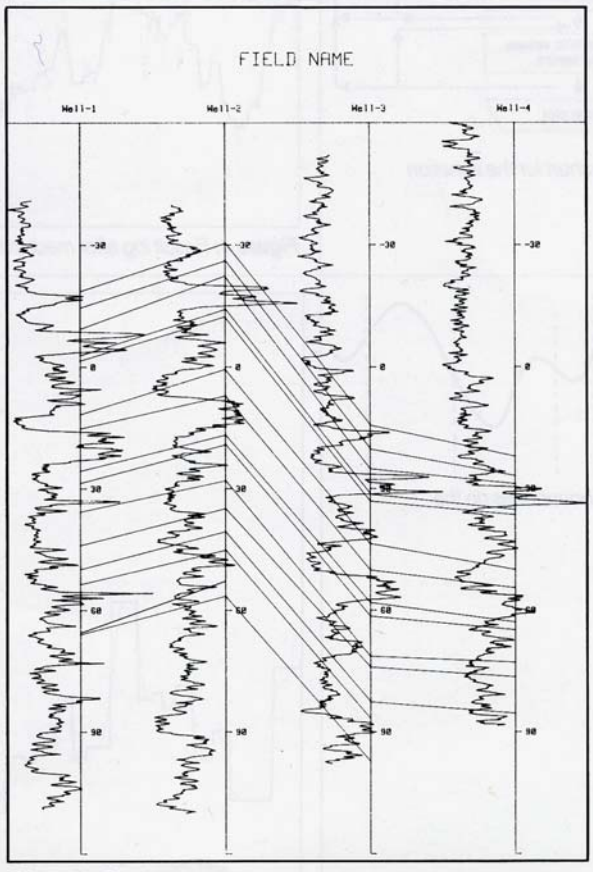


Figure 2: *Result plot of well-to-well correlation taking elevations into consideration.*

critical value decreases, "weak" and "sharp" boundaries occur simultaneously. This procedure can be applied to the c.v. log, so that a boundary occurs only when one has previously been identified, neglecting some less important boundaries. The accuracy of the identification is the same as the identification of fine layers because the identical procedure was applied to the characteristic value log of fine layers (see Figure 7).

• Now return to the original logs to calculate characteristic values of the larger layers. The variations in log response within a big layer can be large as there may be several local minima and maxima which can cause difficulties. To solve these problems, divide the c.v. identification into two options:

a. c.v. can be the average log response within the layer.

b. c.v. can be the extreme (minimum or maximum value) of the log within the layer.

The program selects a or b automatically

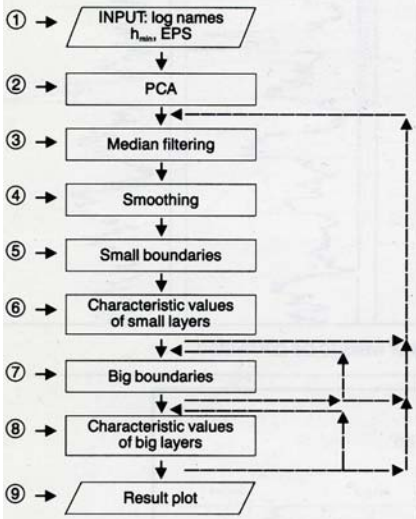


Figure 3: *Flow-chart for the zonation technique.*

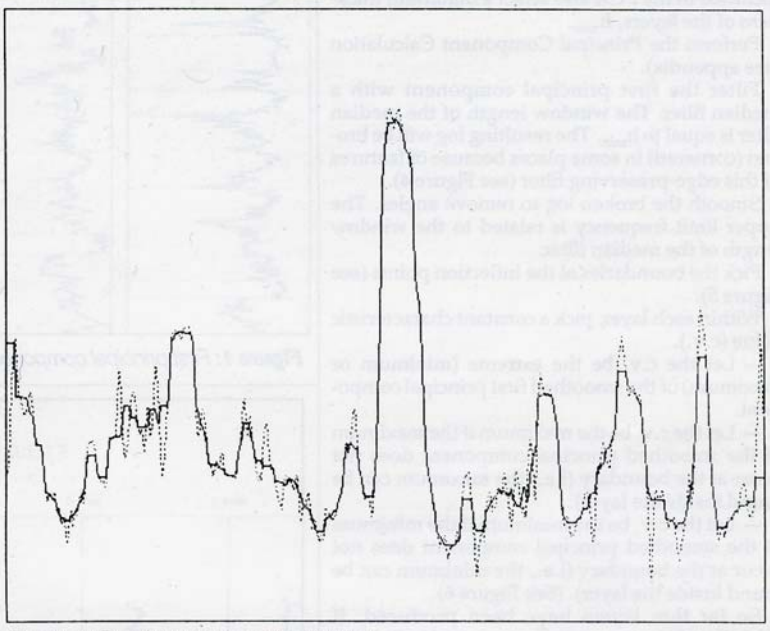


Figure 4: *Result log after median filtering.*

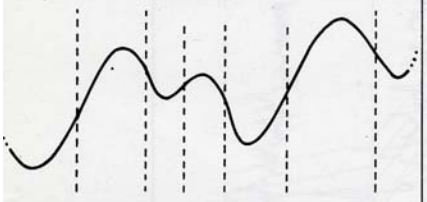


Figure 5: *Fine boundaries on the smoothed log.*

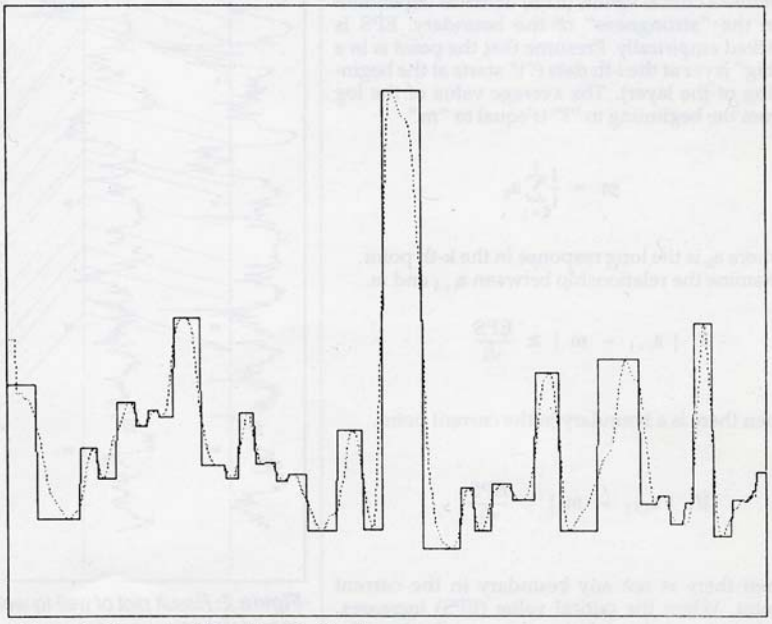
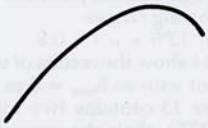


Figure 6: *Characteristic values for fine layers.*

depending on the shape of the curve within the layer. Some simple examples follow:

—If the curve shape resembles this:

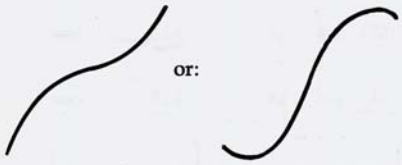


the c.v. is the maximum (or something around maximum).

—If the curve shape resembles this:

then the c.v. is the minimum (or something around minimum).

—If the curve shape resembles this:



then the c.v. is the average.

There are several more complicated cases (i.e., very noisy log), but these can be traced back to the shown simple cases (after using robust statistical estimations and filtering).

Examination of Synthetic Logs

Some synthetic logs were constructed to examine the boundary identification method.

At first several layers were constructed and given five physical values: resistivity, gamma ray level, density, neutron porosity, and acoustic travel time. These are the model logs (see Figure 8).

Uncorrelated noise of varying amplitude was then added to these model logs (see the red curves in Figures 9 and 10).

In Figures 9 and 10 the following logs are depicted:

The green curves are the model logs (MLLD, MGR, MDEL, MFIN, MATL). The MBOUND curve is the "true" boundary log which was used to compute the squared logs. These are the "ideal" boundaries. The red curves are the "noised" logs (NLLD, NGR, NDEL, NFIN, NATL). The blues curves are the zoned logs computed from the boundary identification processing.

The green and blue curves should overlay each other if the boundary identification process has been successful. There are two or three places where the blue zoned logs and green model log curves are separated. In some other cases, only one dark curve can be seen because of the overlay.

The first principle component of the "noised" logs can be seen on Figure 9 in the right column (red curve). The "ideal" model boundaries (green - MBOUND) and the computed resultant boundaries (blue - RBOUND) are the results of the processing and are shown in Figure 9.

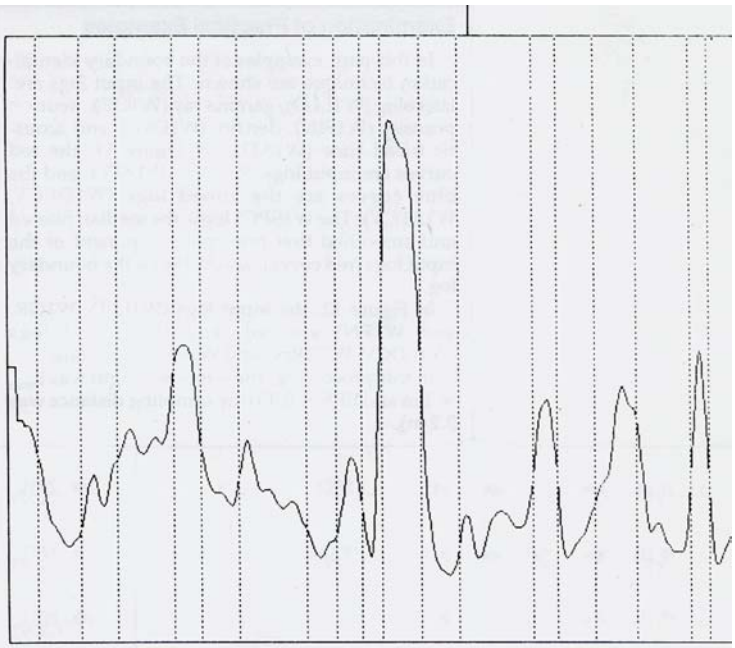


Figure 7: *Boundaries of big layers.*

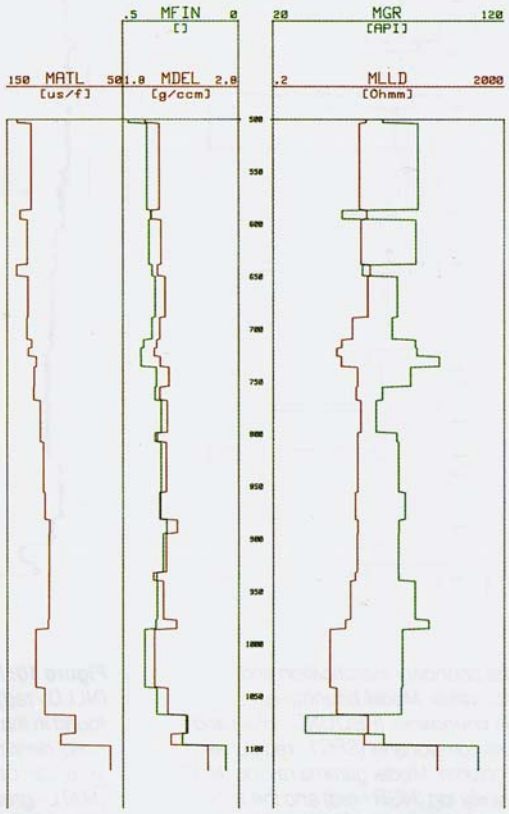


Figure 8: *Synthetic logs: "M" prefix means "model."*
MLLD: laterolog MGR: gamma ray MDEL: density
MFIN: neutron porosity MATL: acoustic

Examination of Practical Examples

In this part, examples of the boundary identification technique are shown. The input logs are: laterolog (W1LLD), gamma ray (W1GR), neutron porosity (W1FIN), density (W1DEL), and acoustic travel time (W1ATL). In Figure 11, the red curves are input logs (W1DEL, W1ATL), and the blue curves are the zoned logs (W1DECV, W1ATCV). The W1SPC1 log is the median filtered and smoothed first principal component of the input logs (red curve), and W1BO is the boundary log.

In Figure 12, the input logs (W1LLD, W1GR, and W1FN) are red, and the zoned logs (W1LDCV, W1GRcv, and W1FNCV) are blue.

In this processing, the window length was h_{min} = 1 m and EPS = 0.15 (the sampling distance was 0.2 m).

The results will be fine resolution if parameters are chosen within the following ranges:
h_{min} = 0.4 m – 2 m, EPS = 0.05 – 0.3.

To obtain coarser layers, the parameter should fall within the following ranges:
h_{min} = 2 m – 10 m, EPS = 0.3 – 0.8.

Figures 13 and 14 show the results of using the same input logs but with an h_{min} = 5 m and and EPS = 0.3. Figure 13 contains two input logs (W1DEL and W1ATL), their characteristic value logs (W1DECV and W1ATCV), the median filtered and smoothed first principal component log W2SPC1, and the boundary log W2BO.

Figure 14 shows three input logs (W1LLD, W1GR, and W1FIN) and three c.v. logs (W2LDCV, W2GRCV, and W2FNCV).

A major problem with the technique is its sensitivity to correctly depth-matched input logs. If

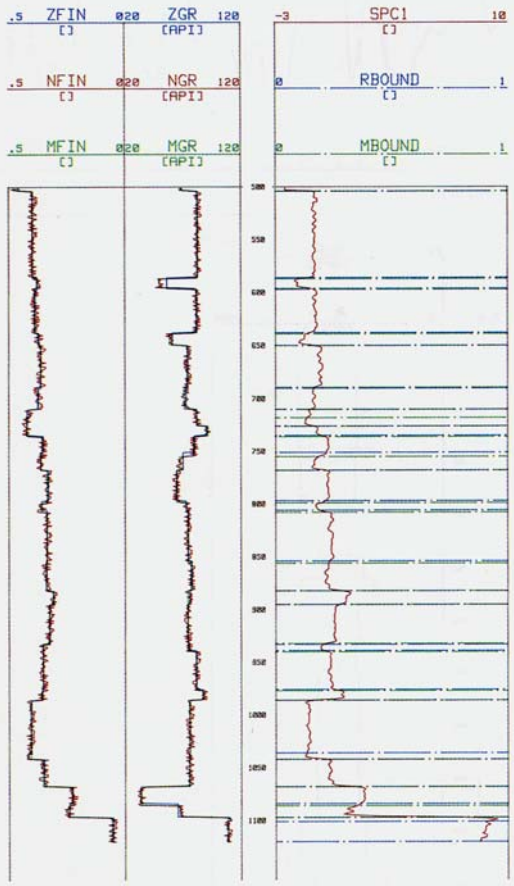


Figure 9: *Result plot of the boundary identification and calculation of characteristic value. Model boundaries (MBOUND - green), result boundaries (RBOUND - blue) and the smoothed first principal components (SPC1 - red curve) can be found in the right column. Model gamma ray log (MGR - green), the noisy gamma ray log (NGR - red) and the zoned gamma ray log (ZGR - blue) can be found in the middle column. Model neutron porosity log (MFIN - green), noisy neutron log (NFIN - red) and zoned neutron log (ZFIN - blue) can be found in the left column.*

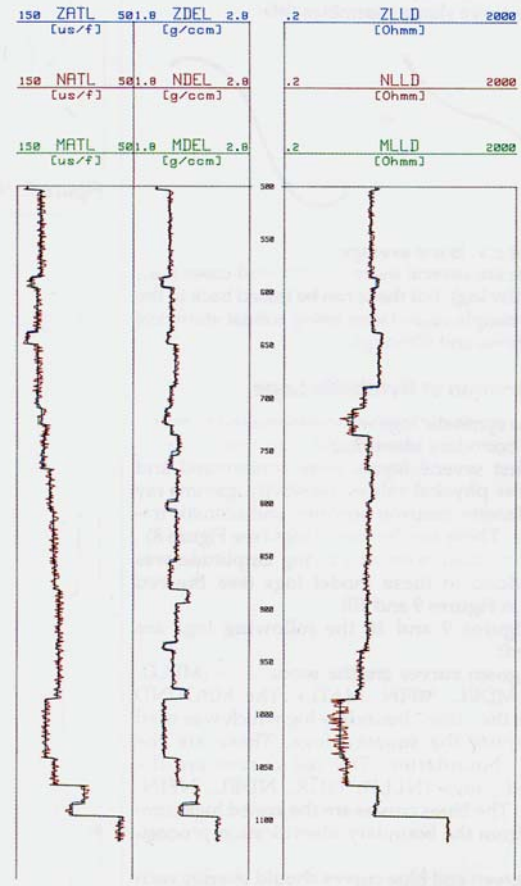


Figure 10: *Model laterolog (MLLD - green), noisy laterolog (NLLD - red) and the zoned laterolog (ZLLD - blue) can be found in the right column. Model density (MDEL - green), noisy density (NDEL - red) and the zoned density (ZDEL - blue) can be found on the middle column. Model acoustic log (MATL - green), noisy acoustic (NATL - red) and zoned acoustic log (ZATL - blue) can be found on the left column.*

logs are not correctly depth-matched, then the first principal component will also contain this mismatch and compute incorrect boundaries. This problem does not cause much error if only large layers need to be identified. (Median filtering and smoothing may be used to reduce this error.)

APPENDIX

Mathematical Bases of the PCA

Assume that the hypothetical database has the following structure: "p" observation units and each unit has "N" data (i.e., there are "p" observation vectors).

$\underline{x}^{(1)}$	$\underline{x}^{(2)}$	. . .	$\underline{x}^{(p)}$
$x_1^{(1)}$	$x_1^{(2)}$	. . .	$x_1^{(p)}$
.	.		.
.	.		.
$x_n^{(1)}$	$x_n^{(2)}$	. . .	$x_n^{(p)}$

Let $\underline{X}^{(j)}$ vectors be random variables. So elements of $\underline{X}^{(j)}$ are realizations of a random variable. In this way each observation unit corresponds to a vector random variable. Since observation units can be of different physical quantities, standardize them:

$$\underline{x}_1^{(j)} = \frac{x_1^{(j)} - \bar{x}^{(j)}}{s^{(j)}}$$

where $\bar{X}^{(j)}$ is the average of the j-th factors, $s^{(j)}$ is the empirical scatter of the j-th vector. (The average of the $\underline{X}^{(j)}$ is equal to zero and its empirical scatter is equal to 1 because of the standardization.) Next, compute the correlation matrix of the database the following way:

$$\underline{\underline{R}} = \begin{bmatrix} r_{11} & r_{12} & . & r_{1p} \\ r_{21} & . & . & . \\ . & . & . & . \\ r_{p1} & . & . & r_{pp} \end{bmatrix}$$

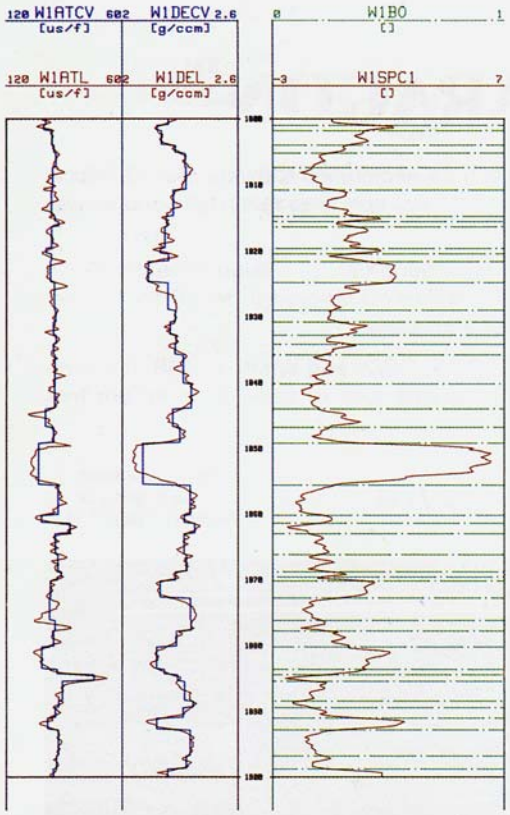


Figure 11: *Result plot of the boundary identification and calculation of the characteristic values.*
H_{min} = 1 m, EPS = 0.15.
W1SPC1 is the median filtered and smoothed first principal component log, W1BO is the boundary log, W1DEL - density, W1DECV - c.v. of the density log, W1ATL - sonic, W1ATCV - c.v. of the sonic log.

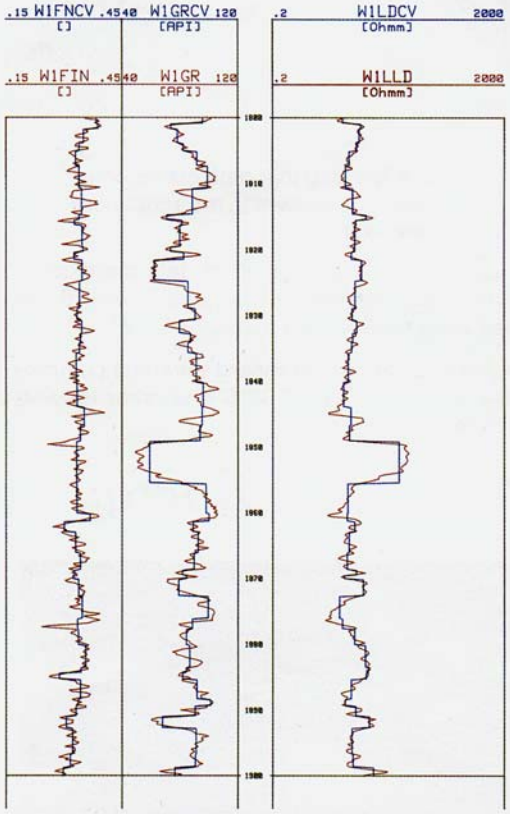


Figure 12: *Result plot of the boundary identification and calculation of the characteristic values.*
H_{min} = 1 m, EPS = 0.15.
The well is the same which was shown in Figure 11. W1LLD - laterolog, W1LDCV - c.v. log of laterolog, W1GR - gamma ray log, W1GRCV - c.v. log of gamma ray, W1FIN - neutron porosity log, W1FNCV - c.v. log of the neutron porosity log.

where r_{ij} = correlation ($\underline{X}^{(i)}$, $\underline{X}^{(j)}$) and i = 1,2,...p, j = 1,2...p

Calculate eigenvalues of the correlation matrix and its eigenvectors. Let $|\lambda_1| > |\lambda_2| > \ldots |\lambda_p|$ be these eigenvalues and $\underline{v}^{(1)}, \underline{v}^{(2)}, \ldots \underline{v}^{(p)}$ eigenvector. The j-th principal component can be computed:

$$c_i^{(j)} = x_i^{(1)} \cdot v_1^{(j)} + x_i^{(2)} \cdot v_2^{(j)} + \cdots + x_i^{(p)} \cdot v_p^{(j)}$$

where

$$i = 1,2 \cdots n \quad i = 1,2 \cdots p \quad \text{and} \quad v_1^{(j)} \cdot v_2^{(j)} \cdots v_p^{(j)}$$

are principal component weights.

Obviously not more than "p" pieces principal components can be computed (p is the number of input vectors).

Principal components have some important properties:

—The correlation matrix of principal components is a diagonal matrix meaning that principal components are uncorrelated.

—Their average is zero and their scatter is equal to the proper eigenvalue.

$$\underline{\underline{R}}_c = \begin{bmatrix} \lambda_1 & & 0 \\ & \lambda_2 \ddots & \\ 0 & & \lambda_p \end{bmatrix}$$

—Eigenvectors are normalized, i.e.,

$$\langle v^{(j)} \cdot v^{i)} \rangle = \delta_{ji}$$

where $\langle\rangle$ means scaler product and δ_{ji} is the Kronecker-delta.

$$\delta_{ji} = \begin{cases} 1 \text{ if } j = i \\ 0 \text{ else} \end{cases}$$

—Sum of eigenvalues is equal to the number of the input logs.

—Sum of variances of the standardized logs is equal to the number of the input logs.

$$\sum_{j=1}^{p} \lambda j = \sum_{i=1}^{p} S_i^2 = \sum_{k=1}^{p} S_k^2 = p$$

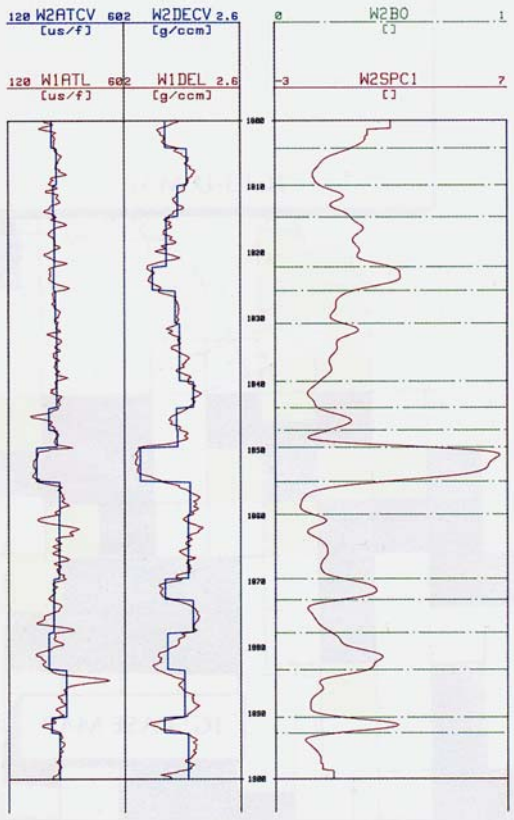


Figure 13: *The same well which was seen in Figures 11 and 12. The difference between the two results is only in the values of h_{min} and EPS. In this case h_{min} = 4 m, EPS = 0.6. W1SPC1 - median filtered and smoothed first principal component log, W1BO - boundary log, W1DEL - density log, W1DECV - c.v. log of density, W2DECV - c.v. log of the density log, W1ATL - sonic log, W2ATCV - c.v. log of the sonic log.*

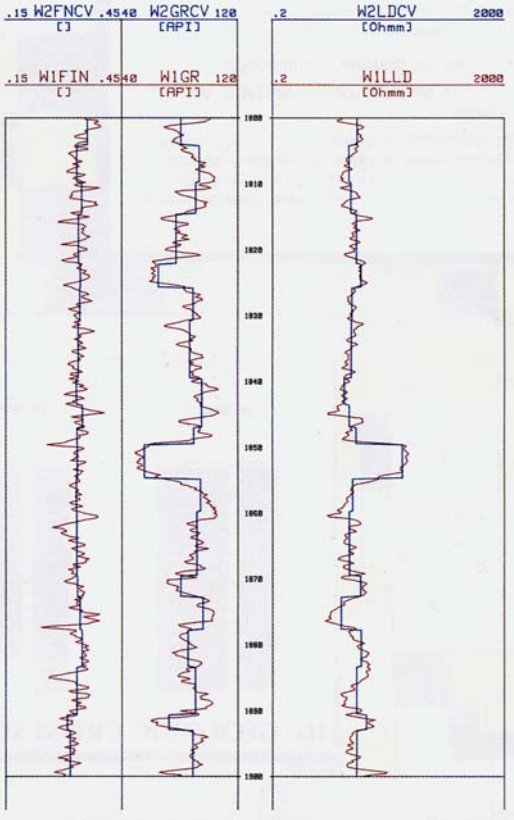


Figure 14: *Result plot of the boundary identification and calculation of the characteristic values.*
H_{min} = 1 m, EPS = 0.6.
W1LLD - laterolog, W1LDCV - c.v. log of laterolog, W1GR - gamma ray log, W2GRCV - c.v. log of gamma ray log, W1FIN - neutron porosity log, W1FNCV - c.v. log of the neutron porosity log.

where λ_j are eigenvalues, s_i are the variances of the standardized logs meaning that the PCA only rearranged variances but did not change the quantity of variances.

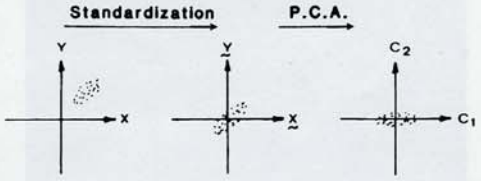
Demonstration of Meanings of Principal Components

Variables are rearranged by means of Principal Component Analysis. Each standardized variable has the same importance by reason of variances, but the first principal component includes the largest common part of all the variables. The second principal component includes the largest common part of the residual variances, etc.

The last principal component has hardly any variances at all and so may be considered unimportant. Therefore, the first principal component may be used instead of the whole database in many cases, through the number of principal components to be used depends on the problem being examined.

Geometrical Meaning of PCA

In the simple case where two variables ($\underline{X}$, $\underline{V}$ vectors), exist, the first step of Principal Component Analysis is standardization (scaling from -1 to 1 around a value of 0), and the second step is the computation of eigenvalues and eigenvectors and calculation of principal components.



The processing consists of a coordinate transformation, which moves the point cloud to the origin (standardization), and then rotates the coordinate axes until the main direction of the cloud is aligned with the first principal component axis.

Inversion of the PCA

The computation of principal components can be inverted and the standardized variables can be retrieved from the principal components. Remember how the principal components were computed:

$$c_i^{(j)} = x_i^{(1)} \cdot v_1^{(j)} + x_i^{(2)} \cdot v_2^{(j)} + \cdots + x_i^{(p)} \cdot v_p^{(j)}$$

where $i = 1,2,\ldots n$ $\quad j = 1,2,\ldots p$
and $x^{(j)}_i$ are the standardized variables and
$c^{(j)}_i$ are the principal components.
Now try to retrieve the standardized variables:

$$x_i^{(j)} = c_i^{(1)} \cdot v_j^{(1)} + c_i^{(2)} \cdot v_j^{(2)} + \cdots + c_i^{(p)} \cdot v_j^{(p)}$$

If a matrix from principal component weights is constructed, then the direct and inverse computation can be easily seen.

DIRECT

INVERSE

$$\begin{matrix} v_1^{(1)} & v_2^{(1)} & \cdots & v_p^{(1)} \\ v_1^{(2)} & v_2^{(2)} & \cdots & v_p^{(2)} \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ v_1^{(p)} & v_2^{(p)} & \cdots & v_p^{(p)} \end{matrix}$$

Direct transformation:
first principal component computation - coefficients are elements of the first *line.*
second principal component computation - coefficients are elements of the second *line.*

.
.
.

p-th principal component computation - coefficients are elements of the p-th *line.*

Inverse transformation:
first standardized variables computation - coefficients are elements of the first *column.*
second standardized variables computation - coefficients are elements of the second *column.*

.
.
.

p-th standardized variables computation - coefficients are elements of the p-th *column.*

References

Biemond, Jan, and J. J. Gerbrands, 1979, An edge-preserving recursive noise-smoothing algorithm for image data: IEEE Transaction systems man and cybernetics, vol. smc. 9, no. 10.

Elek, Istvan, 1986, Some applications of the principal component analysis in the well log interpretation: Hungarian Geophysics No. 1 (in Hungarian).

——— and Gyorgy Kovacs, 1984, An application of the Walsh-transformation in the well log interpretation; manuscript SZKFI (in Hungarian).

Janos, Svab, 1984, Multivariate methods in the biometry: Agricultural Publishing House (in Hungarian).

Krezner, Mark G., and Elton Frost, Jr., Blocking - a new technique for well log interpretation: SPE 11093.

Morrison, Donald F., 1979, Multivariate statistical methods: McGraw-Hill, New York.

Pomalaza-Raez, Carlos A., and Clare D. McGilllen, 1984, An adaptive edge-preserving filter: IEE Transaction on speech and signal processing, vol. ASSP-32, no. 3.

Schlumberger Well Evaluation Conference Proceedings, Algiers, 1979.

Serra, O., and H. T. Abbot, 1982, The contribution of logging data to sedimentology and stratigraphy: SPE 9270.

Vincent, P., J. E. Gartner, and G. Attali, 1977, "GEODIP": an approach to detailed dip determination using correlation by pattern recognition: Paper SPE 6823.

Acknowledgments

The author is thankful to the Cities Service Company and its employees for helping him. He particularly wishes to thank Robert Y. Elphick who corrected the manuscript and helped to publish this article. In the end, thanks are also extended to the Hungarian Hydrocarbon Institute for granting permission to publish the article and providing computer facilities.

G