Australian Journal of Basic and Applied Sciences, 5(8): 1090-1095, 2011
ISSN 1991-8178

Multivariate Analysis of Sonic Synthetic Profile in Well

[1]Suzana Leitão Russo, [2]Maria Emilia Camargo, [3]Vitor Hugo Simon

[1]Dept. of Statistic and Actuarial Sciences, Federal University of Sergipe, Brasil.
[2]Dept. of Management, University of Caxias of South, Brasil.
[3]Petrobras/Brasil.

Abstract: This works mains to verify the possibility of the use of multivariate statistical methods in the study of the sonic profile in oil wells. The used statistical methods was analysis of cluster, with the concern in identifying the parameters of the reservoir that more intervene with the draining of fluids and from this shape them with a compatible scale. The used data are the series of the data the wells, called profiles, whose interpretation allows an evaluation of the geologic formation in study, that is, of the petroliferous deposit. The analysis of the data was carried through using software SAS Enterprise Guide 4.

Keys words: Analysis Multivariate, Analysis Cluster, Synthetic Sonic Profile.

INTRODUCTION

The activities and studies that they aim at to determine, in qualitative and quantitative terms, the potential of a petroliferous deposit, that is, its productive capacity and the valuation of its reserves of oil and gas, is a factor of great interest of the petroliferous industry. When a petroliferous deposit is discovered, some procedures are developed with the purpose of raising information, aiming at one better agreement of this deposit. Thus, since the exploration of a field until its abandonment, an enormous amount of data is collected. A way of get the data in the phase of perforation of a well, is the perfilagem methods, through them, is gotten innumerable data of the wells, called data of profiles of wells, whose interpretation allows an evaluation of the formation in bigger intervals and real conditions of the well. The petroliferous industry has been motivated to use renewed techniques of reservoir characterization, for the high investments necessary in the development of the heterogeneous fields. Amongst these techniques, cite the multivariate statistical techniques, with increasing application, especially when it has the use of three-dimensional seismic data. Thus, the objective of this work is the analysis through the methods of cluster for definitions of homogeneous groups between the observations in study.

MATERIALS AND METHODS

Theoretical Review:

Some methods of analysis multivariate with well diverse purposes between itself exist. Therefore, we come back to the first step, that is to know that knowledge if intends to generate. Or better, what it is intended to affirm regarding the data. To show this diversity, we go to consider some objectives and to indicate some possible methods (RENCHER, 2002).

Analysis of Cluster – this term, first used for (TYRON, 1939), in the reality holds a variety of different algorithms of classification, all directed toward an important question in some areas of the research. The process must take in account the possibility of to carry through a hierarchic organization of groups, where to each level of bigger abstraction, the differences between elements contained in each group are also bigger, where they possess little similarity (CORRAR *et al., 2007*).

So that groups of data of profiles of distinct wells of exploitation can be grouped according to its similarity, it is used technique of *clustering* (jain; dubes1988 and BERKHIN, 2002]. This is one generic technique, that is, its application independent of any characteristic of the structure of the grouping. The results of a hierarchical clustering procedure can be displayed graphically using a *tree diagram*, also known as a

Corresponding Author: Suzana Leitão Russo, Dept. of Statistic and Actuarial Sciences, Federal University of Sergipe, Brasil
E-mail: suzana.ufs@hotmail.com

dendrogram, which shows all the steps in the hierarchical procedure, including the distances at which clusters are merged (RENCHER, 2002).

The majority of the methods of analysis of cluster requires a measure of similarity enters the elements to be grouped, normally express as a function metric or distance (HAIR *et al., 1995*).

Dissimilarity Measure (TIM, 2002):

a) Euclidean Distance:

For continuous (ratio scale, interval) variables, the most common dissimilarity measure is the Euclidean distance between two objects.

The Euclidean distance is the geometric distance in the multidimensional space. The Euclidean distance between two rows $X = (X_1, X_2, ..., X_p)$ e $Y = (Y_1, Y_2, ..., Y_p)$ Y = [Y1,Y 2,...,Yp], is defined by:

$$d_{xy} = \sqrt{(X_1 - Y_1)^2 + (X_2 - Y_2)^2 + ... + (X_p - Y_p)^2} = \sqrt{\sum_{i=1}^{p} (X_i - Y_i)^2}$$

b) Square Euclidean Distance:

The Square Euclidean distance is defined by:

$$d_{xy} = (X_1 - Y_1)^2 + (X_2 - Y_2)^2 + ... + (X_p - Y_p)^2 = \sum_{i=1}^{p} (X_i - Y_i)^2$$

c) Manhattan Distance:

The Manhattan Distance is defined by:

$$d_{xy} = |X_1 - Y_1| + |X_2 - Y_2| + ... + |X_p - Y_p| = \sum_{i=1}^{p} |X_i - Y|_i$$

d) Chebychev Distance:

The Chebychev Distance is appropriate in the case to be defined as two elements are different, if only one of the dimensions is different. Its is defined by:

$$d_{xy} = máximo(|X_1 - Y_1|, |X_2 - Y_2|, ..., |X_p - Y_p|)$$

e) Standardized Euclidean Distance

When working with quantitative variables, the Euclidean distance commonly sum distances that are not comparable, as inches, pounds, years, millions, etc.., although the change of one unit can completely change the meaning and the value of the coefficient.

This is one reason for the standardization the variables of the elements x_1, x_2,..., x_p by the vector $\boldsymbol{x}$.

After the choice of variables to be used as criteria of similarity, one of the vital issues of the techniques of cluster analysis is the definition of the coefficient of similarity or dissimilarity.

Agglomerative Hierarchical Clustering Methods:

Agglomerative hierarchical clustering methods use the elements of a proximity matrix togenerate a tree diagram or dendogram.

A variety of agglomerative methods exists, that are characterized in accordance with the used criterion to define the distances between groups. However, the majority of the methods seems to be alternative formularizations of three great concepts of agglomerative grouping (ANDERBERG, 1973):

1. Linkage Methods (*single linkage*, *complete linkage*, *average linkage*, *median linkage*);
2. Centroid Method;
3. Incremental Sum of Squares Method (*Ward* Method) (JAIN *et al., 1999*).

The inter-cluster distances used by three commonly applied hierarchical clustering techniques are, *Single linkage clustering (*distance between their closest observations); *Complete linkage clustering (*distance between the most remote observations); *Average linkage clustering* (average of distances between all pairs of observations, where members of a pair are in different groups)

Single Link (Nearest-Neighbor) Method:

To implement the single link method, one combines objects in clusters using the minimum dissimilarity between clusters: $d_{(UV)W} = \min(d_{UW}, d_{VW})$.

Some features of this method are (ANDERBERG, 1973):

1. In general, very close groups can not be identified;
2. To detect groups of non-elliptical shapes;
3. It shows little tolerance for noise, it tends to incorporate the noise in a existing group;
4. It shows good results both for Euclidean distances and for other distances;
5. Tendency to form long chains (chain).

Complete Link (Farthest-Neighbor) Method:

In the single link method, dissimilarities were replaced using minimum values. For the complete link procedure, maximum values are calculated instead.

$$_{(UV)W} = \max(d_{UW} , d_{VW})$$

Some features of this method are (KAUFMAN, ROUSSEEUW, 1999; ROMESBURG, 1984):

1. It shows good results both for Euclidean distances and for other distances;
2. Tendency to form compact groups;
3. The noises are slow to be incorporated into the group.

Average Link Method:

When comparing two clusters of objects *R* and *S*, the single link and complete link methods of combining clusters depended only upon a single pair of objects within each cluster. Instead of using a minimum or maximum measure, the average link method calculates the distance between two clusters using the average of the dissimilarities in each cluster.

Some features of this method are (KAUFMAN, ROUSSEEUW, 1990):

1. Less sensitivity to noise that the connection methods for nearest neighbor and farthest neighbor;
2. It shows good results both for Euclidean distances and for other distances;
3. Tendency to form groups with similar number of elements.

Centroid Method:

In the average link method, the distance between two clusters is defined as an average of dissimilarity measures.

As characteristics of this method are:

1. Robustness to the presence of noise;
2. Due to the phenomenon of reversal, the method is not widely used.

Median Linkage Method:

In this method, the distance function is given by:

$$d_{(uv)w} = d_{uw} + d_{vw} /2 - (d_{uv} /4)$$

where: d_{UW} e d_{VW} are the distance between the elements *UW* e *VW*, respectivately.

Some features of the reference method are:

1. Presents satisfactory result when groups have different sizes;
2. It can have different result when the elements of the similarity matrix is the permuted;
3. Robustness to the presence of outliers.

Ward's (Incremental Sum of Squares) Method:

Ward's method for forming clusters joins objects based upon minimizing the minimal increment in the within or error sum of squares. At each step of the process, $n(n-1)/2$ pairs of clusters are formed and the two objects that increase the sum of squares for error least are joined. The process is continued until all objects are joined. The dendogram is constructed based upon the minimum increase in the sum of squares for error.

Some features of the reference method are:

1. It shows good results both for Euclidean distances and for other distances;
2. It can provide unsatisfactory results when the number of elements in each group is nearly equal;
3. It tends to combine with a few groups of elements;
4. Sensitive to the presence of outliers

Nonhierarchical Clustering Method:

The non-hierarchical method, or by partitioning have been developed to group elements into K groups, where K is the number of groups previously defined. Not all groups have values of K satisfying, therefore, applies the method several times for different values of K, choosing to present the results better interpretation of the groups or a better graphic representation (BUSSAB *et al.,* 1990).

Compared with the hierarchical method, the method for partitioning is faster because it is not necessary to calculate and store during processing, the similarity matrix. The methods for partitioning best known are the *k-means* method (k-means) and *k*-method (k-medóides), and are described below.

K-Means Method:

The *k-means* method takes an input parameter, *K,* and partitions a set of *N* elements into *K* groups. This method has a time complexity of order *O(nkl)* and a space complexity is the order *O(k+n)*, where *n* is the number of elements, *k* is the number of clusters and *l* is the number of iterations algorithm (Jain *et al.,* 1999).

A number of statistics are generated by cluster analysis programs that may be plotted to heuristically evaluate how many clusters are generated by the clustering process. Some indices generated by the procedure CLUSTER in SAS include:

1. Pseudo *F* and *t*2 statistics (PSEUDO).
2. The root mean square standard deviation (RMSSTD).
3. *R*2 and semi partial *R*2 (RSQUARE).
4. Centroid Distances (NONORM).

Methodology:

The analyzed data set is about information of the variable of profiles of wells that measure properties of the rocks crossed for the well, that it is located in a field of oil of the Basin Sedimentary Sergipe-Alagoas (Brazil). Some of these wells make use of a complete set of profiles, also the sonic one. The possible stratifications of the samples will be identified as: litologics depth, types (compositions), stratigrafics levels, etc… To guarantee the quality of the results it was made specific tests to be carried through parallel to the analysis statistics of the results, such as: test for the parameters; test of significance of the relation between the variable; the verification of the determination coefficient, the significance of the correlation and the regression, as well as the occurrence of aberrant points and crossed validation.

Characterization Of The Variables Of Profiles:

Rays Gamma - GR: It is the measure of the present total radioactivity in the rocks.
Density - RHOB: It measures the density of the rocks.
Sonic - DT: A sonorous wave measures the time necessary to cover a rock foot, this time is called transit time.

The classification will be made hierarchically and will be defined a restricted number of homogeneous class, that is, will be described parsimonious blocks according to its level of similarity, being based, for this, in the characteristics of the stratigraphics levels of the wells.

RESULTS AND DISCUSSIONS

The analyzed data are about information of the variables of profiles of the P1 well that measures properties of the rocks crossed for the wells that are located in a field of oil of the Basin Sedimentary Sergipe Alagoas. To guarantee the quality of the results it was made specific tests to be carried through parallel to the analysis of grouping of the profiles.

Analysis Of The Variables:

When all variables are measured on the same scale, one may compare the standard deviations to evaluate whether or not some variables may dominate the clustering procedure. When variables are measured on different scales, one may estimate the coefficient of variation *(CV = s/x)* for each variable. When variables display large variations, one may need to employ standardized variables. This is accomplished by using the STD option on the PROC CLUSTER statement in SAS (Tim, 2002).

Table 1 shows the descriptive analysis of the P1 well that will serve to detect the homogeneity of the variables. The RHOB and DT variables with CV 3.97% and 10.38% are distinguished as most homogeneous, while the GR variable with CV of 25.87% is distinguished as most heterogeneous. It is observed, that the GR variable presents a negative asymmetry indicating that the distribution presents a tail for the left side. Already RHOB and DT variable present a positive asymmetry indicating that the distribution presents a tail for the right side.

Table 1: Descriptive Analysis of the variable

Variable	Mean	Std dev.	Asymmetry	CV
RHOB	2.3894	0.0949	0.4364	3,97%
DT	76.6602	7.9540	0.0206	10,38%
GR	58.0940	15.0294	-0.4539	25,87%

It was necessary to standardize the variables for later applying the method of cluster hierarchic of Ward. Fig. 1. presents the dendogram produced for the analysis of groupings.

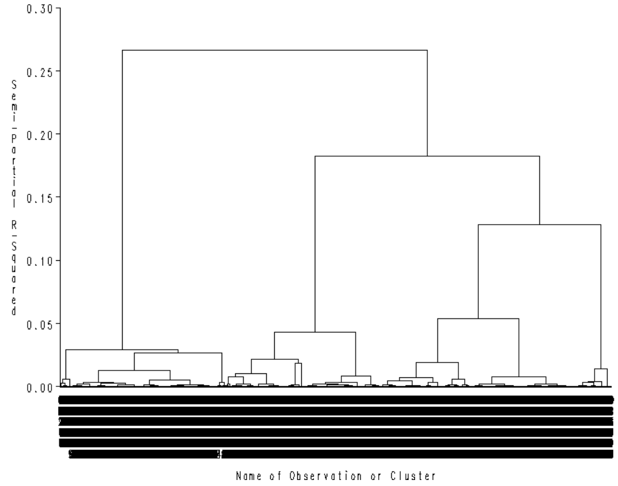

Fig. 1: Dendogram using the Ward method, with standardized data

In Fig. 1 it can be visualized four great main groupings, which if waste in lesser groupings, the test of pseudo F was used to define the number of clusters. In table 3 it is verified existence of four groups, being (CL6, CL2), (CL5, CL3). Of this form, it was grouped the elements next, this occurs in the agglomerative method, each element is initiated representing a group, and to each step, a group or element is on to another one in accordance with its similarity, until the last step, where an only group with all is formed the elements.

Table 3: Cluster analysis used the Ward method

NCL	Clusters	Aderido	Frequencies	Spr^2	R^2	Ersq	Ccc	Psf	Pst2	Tie
1	CL6	CL2	5374	0.2664	.000	.000	0.00	.	1951	
2	CL5	CL3	3756	0.1825	.266	.449	-39	1951	1581	
3	CL4	CL11	2255	0.1284	.449	.613	-47	2188	1805	
4	CL9	CL15	1921	0.0540	.577	.683	-41	2445	1315	

Table 4 shows the covariance matrix between variables, as an alternative measure of similarity, we used Pearson's coefficient, was made also an analysis only with the variables DT, GR and RHOB of P1 well, which shows that GR variable has the largest eigenvalue from the other two variables. The eigenvalues of the correlation matrix for the standardized variables indicate variation in three or four dimensions, accounting for 59,84% to 89,24% of the variance, then the three are come close to 100%. In such a way, the three variables had been analyzed:

Table 4 : Eigenvalues of Covariance Matrix

	Eidenvalues	Difference	Percentage	Accumulated
1	1.7953	0.9135	0.5984	0.5984
2	0.8818	0.5589	0.2939	0.8924
3	0.3228		0.1076	1.0000

It was processed an analysis applying the non-hierarchic method (*K-means*) from the centroids gotten in the Ward method. In all these cases, the results had revealed compatible with the solution proposal, presenting small variations of classification amongst clusters. To follow they are the results after the standardization of DT, GR and RHOB variables where were applied the method non-hierarchic *K-means*.

Table 5: Summary of the initials seeds

Cluster	RHOB standardized	GR standardized	DT standardized
1	2.345977379	1.585758892	1.947669905
2	-2.457067849	-0.454810029	0.085024754
3	-7.528822277	-0.477325917	-2.545786294
4	3.453480787	-2.049585651	-3.159061786

Table 5 shows the standardized variables, where four clusters were identified, in the first cluster all variables are positive in the second cluster only the variable DT is positive relative to the other, the third cluster variables are all negative and the fourth cluster RHOB variable is positive in relation to others. In the Table 6 shows that the clusters resemble one and two best clusters in relation to three and four.

Table 6: Clusters Summary

Cluster	Frequencies Y	RMS Std Dev n	Maximum Distance Observation Seed for n	Neighbor Cluster	Distance between centroids
1	841	0.7508	4.5530	2	1.4385
2	3723	0.7634	3.8672	1	1.4385
3	15	1.4695	5.9788	2	5.1142
4	795	0.7625	2.5676	1	2.6121

Conclusion:

This work aimed to show the usefulness of the technique of cluster analysis on the profiles of the sonic wells P1. To this end, we used the software resources to SAS Enterprise Guide 4, we sought to apply the technique of cluster analysis in Ward method in some variables that capture the performance of wells, the variables chosen were the profiles DT, GR and RHOB. Thus we see that the computational facilities of obtaining dendograms chart and allow a more rapid use of these methods.

REFERENCES

AIN, A.K., M.N. MURTY, P.J. FLYNN, 1999. Data clustering: a review. ACM Computing Surveys, New York, 31(3): 265-323.

ANDERBERG, M.R., 1973. Cluster analysis for applications. New York: Academic Press.

BERKHIN, P., 2002. Survey of clustering data mining techniques. Technical Report, Accrue Software, San Jose, CA.

BUSSAB, W. de O., E.S. MIAZAKI, ANDRADE, D.F. de. 1990. Introdução à análise de agrupamentos. São Paulo: Associação Brasileira de Estatística.

CORRAR, L.J., E. PAULO, J. M. DIAS FILHO, 2007. Análise Multivariada. São Paulo: Ed. Atlas.

HAIR Jr.J.F, R.E. ANDERSON, R.L . TATHAM, BLACK, W.C. 1995. Multivariate data analysis. 4.ed.NJ – USA: Pentice_Hall, Inc.

JAIN A.K., R.C. DUBES, 1988. Algorithms for clustering data, Prentice Hall.

KAUFMAN, L., P.J. ROUSSEEUW, 1990. Finding groups in data: an introduction to cluster analysis. New York: Wiley.

RENCHER, A., 2002. Methods of Multivariate Analysis. 2nd ed. New York: Wiley & Sons.

ROMESBURG, C.H., 1984. Cluster analysis for research*ers.* Belmont: Lifetime Learning Publications.

TIM, N.H., 2002. Applied Multivariate Analysis. New York: Springer-Verlag.

TYRON, R.C., 1939. Cluster Analysis. Ann Arbor, MI: Edwards Brothers., p: 422.