

Application of Pattern Recognition Techniques to Hydrogeological Modeling of Mature Oilfields

Leonid Sheremetov, Ana Cosultchi, Ildar Batyrshin, and Jorge Velasco-Hernandez

Mexican Petroleum Institute, Eje Central Lazaro Cardenas 152,
San Bartolo Atepehuacan, 07730, Distrito Federal, Mexico
{sher, acosul, batyr, velascoj}@imp.mx

Abstract. Several pattern recognition techniques are applied for hydrogeological modeling of mature oilfields. Principle component analysis and clustering have become an integral part of microarray data analysis and interpretation. The algorithmic basis of clustering – the application of unsupervised machine-learning techniques to identify the patterns inherent in a data set – is well established. This paper discusses the motivations for and applications of these techniques to integrate water production data with other physicochemical information in order to classify the aquifers of an oilfield. Further, two time series pattern recognition techniques for basic water cut signatures are discussed and integrated within the methodology for water breakthrough mechanism identification.

Keywords: Principal component analysis, clustering, time series pattern, oilfield.

1 Introduction

In engineering practice data-driven approaches usually help to identify unknown facts and inspire further research of the discovered phenomena. In this study the application of pattern recognition techniques (PRT) is illustrated in the context of hydrogeological modeling of mature oilfields. Water associated to hydrocarbon – known as natural brine or formation water – is water trapped in underground formations that is brought to the surface along with oil or gas. It is one of the most effective driving mechanisms for oil production, but when the amount of produced water becomes excessive, oil production drops down reducing the lifespan of most hydrocarbon wells [1]. It is then not surprising that the nature, chemical properties and quantity of produced or injected water during recovery process all have a direct impact on oilfield development. A hydrogeological model deals with the determining the dynamic behavior of fluids within the formation, the identification of the number of aquifers and their impact on oil production, i.e. the origin of the produced water, the mechanism of intrusion and the flow path directions.

Most of physically-based approaches used to build up a dynamic model of the reservoir, such as geochemical models, pressure tests, tracing tests, require expensive and time-consuming procedures and are not risk-free since sometimes may result in the well shut-in. Moreover, domain experts usually use data-driven procedures

analyzing the results of these tests in order to suggest features for complex heterogeneous reservoirs. When studying directly numerical data, different associations between water-related parameters can simply escape because of the amount of information, which makes the manual analysis impractical. In order to solve these difficulties, the data analysis methods and PRT should be used [2].

Data analysis methods can comprise a necessary alternative and an additional tool for hydrogeological modeling. PRT have been widely used for production and pressure data analysis in different studies of reservoir behavior [3]. Nevertheless, to our knowledge, they have not been systematically used to evaluate the dynamic behavior of fluids. The paper discusses the motivations for and applications of PRT in order to classify the aquifers of an oilfield and identification of water behavioral patterns necessary for establishing water control methods.

The proposed methodology consists of several sequential steps of data analysis to describe the dynamic behavior of formation water in an oilfield. First, the principal component analysis (PCA) of various parameters characterizing oilfield formation water was applied. Second, based on this analysis two clusters of petroleum wells were determined applying hierarchical cluster analysis (HCA). Third, a description of these clusters by parameter values discriminating these clusters one from another was given. Fourth, spatial representation of oilfield wells showed that obtained clusters had good spatial grouping. Finally, using new technique of temporal-spatial visualization proposed in this paper, the dynamics of water invasion in petroleum wells related to the spatial clustering of wells was shown.

In the rest of the paper a methodology, which makes use of several PRT is presented and exemplified with a case study of a Mexican oilfield located in the coastal swamps of the Gulf of Mexico.

2 Multivariate Statistical and Pattern Recognition Techniques for Hydrogeological Modeling

The aquifer is described by physicochemical properties of formation water (brine), which also can be used to identify the water likely migration trends. Conventional practice in production engineering usually assumes that for any formation there is only one common and huge aquifer associated. Although that hypothesis may be true, some oilfields have revealed the existence of various aquifers [4]. Aquifer identification methodology described here roughly consists in the application of multivariate statistical techniques to fluid properties and production data.

Dimensionality reduction is the first step. There are two main reasons to keep the dimensionality of the pattern representation as small as possible: measurement cost and classification accuracy. In many practical cases, not all the measured variables from a high-dimensional datasets are important for understanding the underlying phenomena. The best known unsupervised feature extractor is the PCA, which is the most commonly used technique for allowing high-dimensional representations to be compressed into a lower dimensionality without implication of statistical significance testing [5]. In the proposed method, PCA is applied with a Varimax rotation, extracting the factors or principal components. Kaiser criterion was also used (eigenvalue > 1).

The recognition problem is considered as a classification or categorization task, where the classes may be defined by the domain expert (supervised classification) or are learnt based on the similarity of patterns (unsupervised classification). The object is represented by a d -dimensional vector of attributes. To make a classification, the space spanned by these vectors, is subdivided using decision boundaries.

Unsupervised classification method, HCA is used to classify entities with similar properties based on their nearness or similarity. This method has been used to identify aquifers as zones with similar or distinct hydrochemical regimes of ground or superficial water. To establish decision boundaries, concepts from statistical decision theory are utilized.

Once aquifers are identified, the preferential flow directions can be inferred from the water production historical data while the water breakthrough mechanism can be evaluated from the logistic curve parameters of the water-cut curve. Water-to-oil ratio (WOR) or water cut data plots (visual patterns) help to understand the type of flow and the water breakthrough mechanism [6]. Several behavior patterns identified from the basic signatures of WOR are defined in the literature. In [7] three patterns are defined: rapid increase, rapid increase followed by a straight line and gradual increase. Each of these patterns is associated with water invasion problems. For instance, a high water cut following an S-shape curve with a high and positive slope indicates flow behind pipes, fractures or fracture-lake features and specially edge water drive associated to linear flow as defined in [1].

The real behavior usually makes hard to relate the water production curve with these patterns [7,8]. Two methods based on moving approximation (MAP) transform and logistic growth curve (LGC) respectively were developed and tested [9, 10]. A LGC can be used to model functions that increase gradually at first, more rapidly in the middle, followed by a slow increase at the end. LGC is proposed as a mathematical tool to be applied to WOR data in order to determine the type of flow and the water invasion mechanism. The formula for the logistic function follows Gompertz curves:

$$y = ae^{-e^{-k(x-x_c)}} \quad (1)$$

where y represents the water cut in %; x is the time of production in months; a is the upper asymptote, x_c is the time of maximum growth and k is the growth rate. For water influx behavior, k is positive and the logistic function will always increase.

In the second pattern recognition method, the MAP-transform replaces values of the WOR time series $y = (y_i, y_{i+1}, \dots, y_n)$ with the sequence of slope values $MAP(y) = (a_1, \dots, a_N)$, where $N = n - k + 1$, obtained as a result of moving approximations of time series in a moving window of size k . A window W_i of a length $k > 1$ is a sequence of indexes $W_i = (i, i+1, \dots, i+k-1)$, $i \in \{1, \dots, n-k+1\}$. The slope values a_1, \dots, a_{n-k+1} are called local trends of a linear function $f_i = a_i t + b_i$ with parameters $\{a_i, b_i\}$ minimizing the criterion:

$$Q(f_i, y_{W_i}) = \sum_{j=i}^{i+k-1} (f_i(t_j) - y_j)^2 = \sum_{j=i}^{i+k-1} (a_i t_j + b_i - y_j)^2 \quad (2)$$

This function is called a moving (least squares) approximation of y and approximates the data of the time series using a sliding window. Being based on the measurement of a distance between sequences of slope values, the method looks for the pattern (a_i, \dots, a_{i+m}) , where m is a size of the pattern given by the sequence of the slopes, in $MAP(y)$ closest to the searched pattern [9].

3 Application of Pattern Recognition Techniques to the Oilfield Data

Pattern recognition techniques were applied to a data set collected from a Mexican oilfield. All mathematical and statistical computations were made using ExcelTM, OriginProTM and STATISTICATM software while visualization of some maps have been done with VMDPetro software developed in the IMP.

3.1 Cluster Analysis of Physic-chemical Water Properties for Aquifer Identification

First, the analysis of physic-chemical water properties was carried out. Water from 16 wells was sampled and analyzed in the laboratory to fulfill a dataset with up to 147 analyses (observations). The data of each analysis were examined, missing values were imputed and outliers were excluded from the dataset. From 16 properties after correlation analysis, the following ones are kept for the following study: TDS, Ca/Mg, Na/Cl, SO_4 , HCO_3 , Fe and SiO_2 . Oil properties were measured and reported only at the beginning of the well production. These properties include fluid density or API, dynamic or absolute viscosity related to internal resistance of the fluid to flow and cinematic viscosity, a physical property which depends on fluid density and also the carbon content measured by two techniques: Conradson and Ramsbottom.

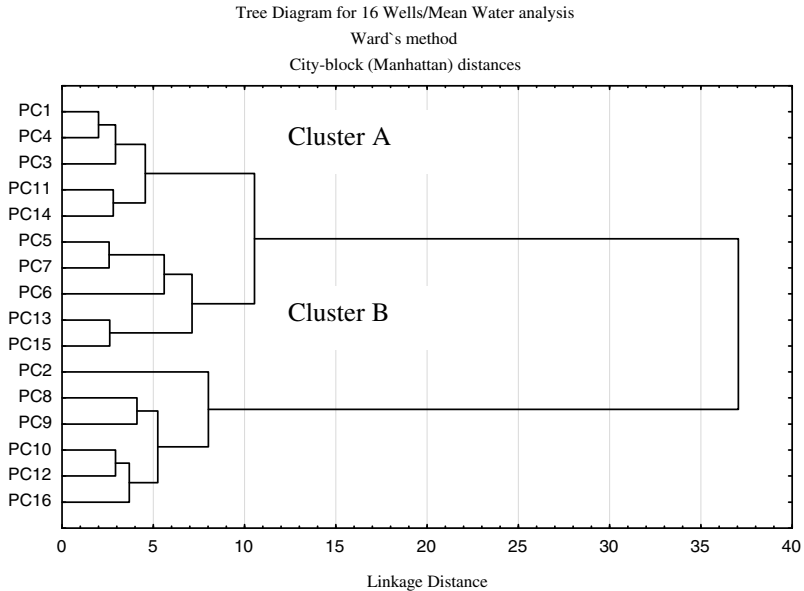
The PCA showed interesting results. While using water and oil properties three principal components were identified: two of them describing water properties and one – for oil, with a total loadings of 70.93%. Nevertheless, results using only water variables show that the system variance can be explained with two PC's with a total loading of 81.72%. It means that oil properties can be excluded from this analysis. Table 1 shows the PC's compositions and loadings after Varimax normalization. PC_i refers to the i -th principal component. PC_1 explains 64.27% (after Varimax rotation - 59%) of the total variance and the loadings are strong and positive for well depth, SO_4^{2-} and SiO_2 , and strong and negative for TDS and Ca/Mg ratio variables. This PC relates to the mineral characteristics of water and is called "water mineralogy factor". PC_2 explains 17.45% (after Varimax rotation - 23%) of the total variance and shows strong and positive loading for Fe ion.

For HCA the set containing the well depth and six water properties – TDS, SO_4^{2-} , SiO_2 , Fe, Ca^{2+}/Mg^{2+} and Na^+/Cl^- – arranged by wells was used. Cluster analysis rendered a dendrogram (Fig. 1) where all samples of the well parameters were grouped into two statistically significant clusters at $(D_{link}/D_{max}) * 100 < 70\%$. The clustering procedure highlighted two groups of wells in a very definite way because the wells in these groups have similar characteristics. Cluster A contains 10 wells, half of them producing from Upper Jurassic Kimmeridgian (UJ_{kim}) and the other half - from Lower Cretaceous (LC) formation; and cluster B contains 6 wells, all of them producing from LC formation.

Table 1. Results of PCA (Eigenvalues and Factor Loadings after Varimax normalization, Marked loadings are $> .7$) and using selected water variables.

	Eigenvalue	% Total	Cumulative	Cumulative
PC1	4.50	64.27	4.50	64.27
PC2	1.22	17.45	5.72	81.72

Variables	PC1	PC2
Well Depth	0.76	0.18
TDS	-0.91	0.35
SO ₄ ²⁻ , meq/l	0.77	-0.46
Fe, mg/l	0.01	0.96
SiO ₂ , mg/L	0.86	0.02
Ca/Mg %	-0.85	0.13
Na/Cl %	0.81	-0.55
Expl.Var	4.10	1.62
Prp.Totl	0.59	0.23

**Fig. 1.** The hierarchical clustering dendrogram of 16 wells of the PC oilfield based on water properties: Ward's method, Manhattan (city-block) distance

Such classification has a clear physical interpretation. Comparison of water properties (not shown here for space limits) shows that Salinity, Electric Conductivity, TDS mean values of Cluster A aquifer are 2.5 times higher than those of Cluster B aquifer and such difference is sustained by the concentration of Cl and Mg ions. Total hardness mean is 5.6 times greater for Cluster A compared to Cluster B and so the concentration of Ca cations. Concentration of SO₄ anions for Cluster B is 2.6 times the concentration of this anion in Cluster A. This difference may be related to the burial conditions of the aquifer or to geochemical process of calcium sulfate mineral dissolution, which is commonly associated to petroleum formations.

The differences between properties are also related to the well depth: Cluster B is grouping deeper wells than Cluster A. Table 2 contains the mean values of all measured variables calculated for each cluster including those, which were not involved in clustering procedure.

Table 2. Statistics of each cluster

Variables	Cluster A		Cluster B	
	Mean value	Std. Desv.	Mean value	Std. Desv.
Well Depth	5889.20	278.31	6363.67	261.03
Water properties				
Density, g/cm ³	1.21	0.02	1.08	0.01
Salinity, mg/L	300750.74	34833.48	116571.30	12250.80
El. Conductivity, mS/cm	605.02	48.97	239.54	28.62
Total hardness, mg/L	137531.58	27510.47	24737.04	3553.79
TDS mg/L	309348.91	21364.85	119294.03	14383.69
Ions, meq/L				
Na ⁺	59450.45	5547.65	34902.48	3451.45
Ca ²⁺	47913.12	9894.68	7610.37	1673.34
Mg ²⁺	4006.76	1289.74	1418.67	433.82
Cl ⁻	186302.42	16292.35	70947.99	7443.87
SO ₄ ²⁻	184.31	97.28	478.38	130.94
HCO ₃ ⁻	330.06	57.53	397.81	128.44
% Ca/ %Mg	7.63	1.80	3.66	1.66
Na%/Cl%	0.49	0.06	0.76	0.02
Solids, mg/L				
Fe	3.94	8.55	0.76	0.80
SiO ₂	46.38	12.72	88.48	37.06
Crude Oil properties				
API	31.36	1.32	31.11	1.20
Dyn. Visc., cP	5.73	1.54	5.81	1.29
Say. Visc., SSU	51.28	6.99	52.67	7.41
RCR, wt%	2.50	0.69	2.77	0.89
RCC, wt%	3.29	0.87	3.23	0.63

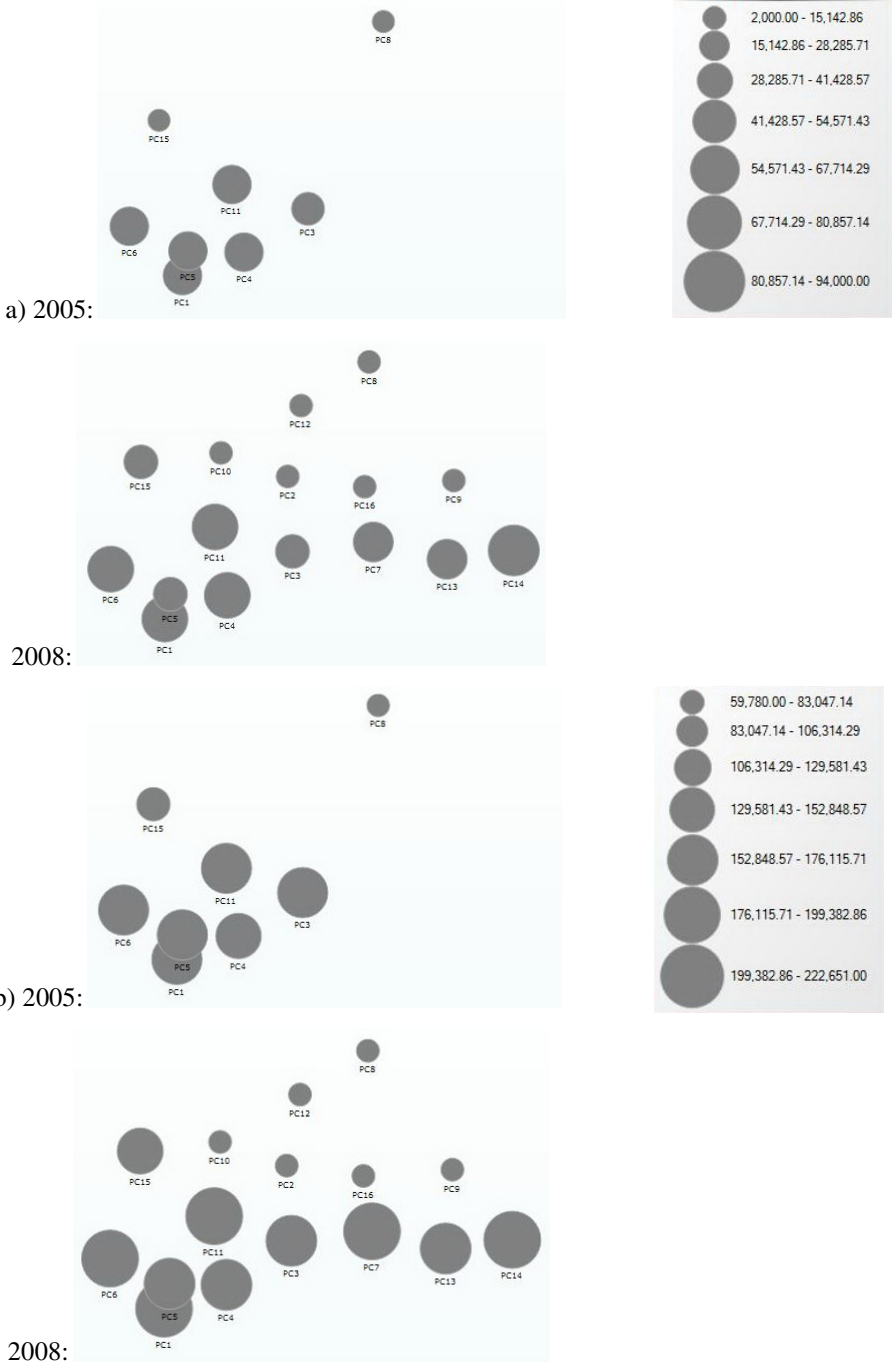


Fig. 2. Spatial and temporal distribution of a) Ca²⁺ and b) Cl⁻ water ion concentrations corresponding to PC wells water analysis dated 2005 and 2008 respectively

For the further interpretation of the HCA results, a geographical distribution of wells was analyzed, showing a clear coincidence with their location dividing the oilfield into two parts: Southern and Northern. For illustrative purposes, only water ion concentrations for Ca^{2+} and Cl^{-} measured during the autumn of 2005 and after four years of production (2008) are presented in Fig. 2.

3.2 Pattern Recognition of Basic WOR Signatures for Water Breakthrough Mechanism Identification

The finding of two aquifers raises two important questions: if there is oil migration between formations through fractures or conductive faults and if water influx is the mechanism connecting these formations? The parameters of the logistic curve, the approximated date of the starting date of water production and the maximum value of water cut are useful for the interpretation of water influx mechanism within each well. For illustrative purposes three most typical identified patterns are shown in Fig. 3. Water cut historical behavior of the wells shows a good fitting of the LGC (Table 3). For the definition of the typical patterns in the form of fuzzy rules [8], the growth rate k and the upper asymptote value a were learnt from the observation data. These parameters can be useful for the field operators and give them a most clear idea of the water problem. Parameter a is very important when compared to the critical water cut calculated from Nodal Analysis used for the analysis of well performance, while k indicates how stressful is the influx problem and what solution should be applied.

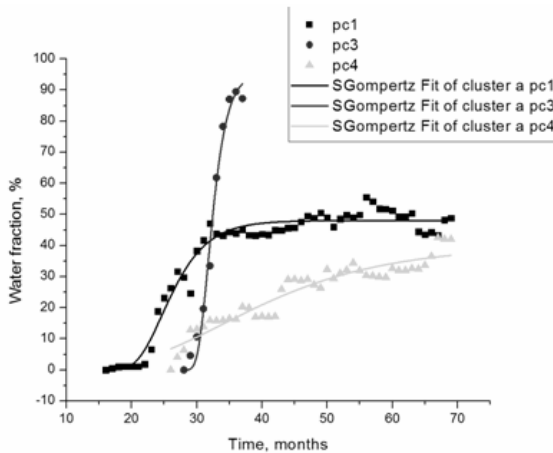


Fig. 3. Three typical behavior WOR patterns with associated logistic Gompertz curve fitting

Additionally to the analysis of the type and distribution of oilfield formation water the dynamics of water invasion of wells has also been studied. For the identification of hydraulic water flow paths within the formation, a new form of representation of spatial-temporal events is proposed, where the earlier events are represented by larger circles (Fig. 4). The spatial distribution of events for each well clearly shows the sequence of water breakthrough within the clusters obtained by HCA. It can be seen

that the size of circles monotonically decreases from wells PC6, PC1 to PC7, PC,13, PC14 for cluster A and from PC8 and PC16 to PC10, PC2 and PC 9 for cluster B. So in these clusters the water invasion is coming from the opposite directions.

Table 3. Parameters of the LGC fitting the water cut data. The parameters are calculated according to the logistic function formula (1).

Model	SGompertz		
Equation	$y = a * \exp(-\exp(-k * (x - xc)))$		
Reduced Chi-Sqr	10.85888	21.5744	10.69689
Adj. R-Square	0.9624	0.98446	0.89447
		Value	Standard Error
pc1	a	48.04807	0.59964
pc1	xc	24.65875	0.28874
pc1	k	0.26092	0.0249
pc3	a	95.53129	5.10802
pc3	xc	31.76059	0.1672
pc3	k	0.62707	0.10336
pc4	a	39.8968	2.63364
pc4	xc	33.91353	1.18578
pc4	k	0.072	0.01226

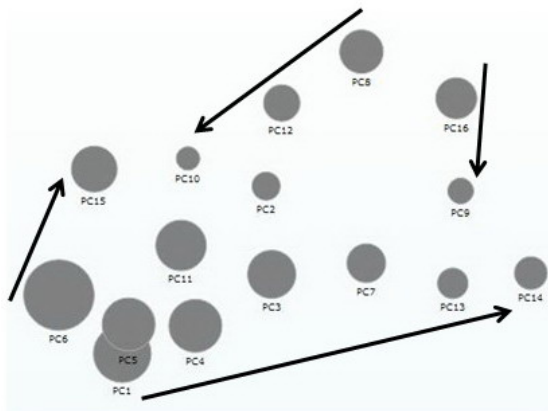


Fig. 4. Spatial-temporal distribution of wells' water breakthrough: the earlier events are represented by larger circles. The arrows show the sequence of these temporal events.

4 Conclusions

The paper shows that PRT can be considered as a powerful exploratory tools to evaluate the dynamic behavior of fluids within a formation; these methods can be

particularly useful when applied previous to any other technique since they can provide important clues not evident in the raw data. The distinctive feature of the proposed approach is the integration of PRT in order to advance with the reservoir hydrogeological modeling. A well-defined and sufficiently constrained recognition problem led to a compact pattern representation and a simple decision making strategy. A varimax rotation of the PCs led to a reduced number of PCs, each of them related to two groups of experimental variables with a respective meaning: water mineralogy factor PC1 and Fe-factor PC2, respectively. HCA has found spatial similarities in wells variations across the oil camp, reducing the number of groups to two although not all the members of the clusters are producing from the same formation. Water behavior PRT resulted in generation of fuzzy rules for diagnostic expert system called SMART-Agua [8].

Existence of large volumes of measurement data is a rather new issue in the oilfield practice, long tradition of using data analysis techniques in that field does not yet exist. The application of the proposed methodology can significantly reduce the time and the costs of petroleum engineering tasks, since the collected data are used for the analysis not requiring additional tests and well shut-ins.

Acknowledgments. The partial support for this work has been provided within the IMP projects D.00507, Y.00102 and Y.00122.

References

1. Seright, R.S., Lane, R.H., Sydansk, R.D.: A strategy for attacking excess water production, paper SPE 84966 (2003)
2. Sheremetov, L., Alvarado, M., Bañares-Alcántara, R., Anminzadeh, F.: Intelligent Computing in Petroleum Engineering (Editorial). *J. of Petroleum Science and Engineering* 47(1-2), 1–3 (2005)
3. Gaskari, R., Mohaghegh, S.D., Jalali, J.: An Integrated Technique for Production Data Analysis with Application to Mature Fields, SPE 100562 (2006)
4. Birkle, P., Martínez-García, B., Milland-Padrón, C.M., Eglinton, B.M.: Origin and evolution of formation water at the Jujo–Tecominoacán oil reservoir, Gulf of Mexico. Part 2: Isotopic and field-production evidence for fluid connectivity. *Applied Geochemistry* 24, 555–573 (2009)
5. Lim, J.S.: Multivariate Statistical Techniques Including PCA and Rule Based Systems for Well Log Correlation. In: Nikraves, M., Aminzadeh, E., Zadeh, L.A. (eds.) *Developments in Petroleum Science*, pp. 673–688. Elsevier, Amsterdam (2003)
6. Yortsos, Y.C., Choi, Y., Yang, Z., Shah, P.C.: Analysis and Interpretation of Water/Oil Ratio in Waterfloods. SPEJ 4, 413–424, paper SPE 59477 (1999)
7. Bailey, B., Crabtree, M., Tyrie, J., Elphick, J., Kuchuk, F., Romano, C., Roodhart, L.: Water control. *Oilfield Review*, Schlumberger (2000)
8. Sheremetov, L., Batyrshin, I., Cosultchi, A., Martínez-Muñoz, J.: SMART-Agua: a Hybrid Intelligent System for Diagnostics. In: INES 2006, London, pp. 238–243 (2006)
9. Batyrshin, I., Sheremetov, L.: Time Series Pattern Recognition Based on MAP Transform and Local Trend Associations. In: Martínez-Trinidad, J.F., Carrasco Ochoa, J.A., Kittler, J. (eds.) *CIARP 2006. LNCS*, vol. 4225, pp. 910–919. Springer, Heidelberg (2006)
10. Cook, L.M.: Oscillation in the simple logistic growth model. *Nature* 207, 316 (1965)