Detection of Geologic Anomalies with Monte Carlo Clustering Assemblies

Simon Katz[1]*, Fred Aminzadeh[2], George Chilingar[3], Leonid Khilyuk[4] and Matin Lockpour[5]

[1,4]*Russian Academy of Natural Sciences, US Branch, CA.*
[2,3,5]*Petroleum Engineering Program, School of Engineering, University of Southern California, Los Angeles, CA*

Received October 08, 2018; Accepted January 7, 2019

Abstract Authors present new clustering-based algorithms for detection of geologic anomalies and results of their testing on the data containing anomalies of two types: (a) high permeability anomaly with regular records containing smaller permeability values, and (b) gas-filled sand anomaly with regular records containing data from brine-filled sands. Results of algorithms testing, presented in the paper, demonstrate high stability of anomaly detection with false discovery rate below 20% and with the true discovery rate exceeding 73%.

Keywords Anomaly detection, clustering assembly seismic velocities, rock density, brine, gas, true discovery, false discovery

1 Introduction

The goal of this paper is to present new clustering assembly algorithms for detection of geologic anomalies of various types. These algorithms may be utilized as one of the steps in location of oil and gas reservoirs, overpressure zones, and other geologic anomalies. List of publications on detection of geologic anomalies includes finding the location of fractured carbonates filled with gas [4], finding the location of an overpressure zone [5], and one-class methodology for anomaly detection within the homogeneous geologic area [7]. Key element of the proposed methodology is repeated clustering of analyzed data. Clustering technique has been used for geologic applications [2] and for detection of single outliers or clusters of small size [8, 10, 11]. In geological applications, size of anomaly area may not be necessarily small. Besides, records in the training set may form several clusters of different size.

**Corresponding author*: simonkatz2000@yahoo.com

DOI: 10.7569/JSEE.2018.629516

This makes the problem of detection of geologic anomaly more complex, compared to the problem of outlier detection. The authors overcome this problem via construction of multiple randomized train and test sets and clustering them. As a consequence, obtained multiple cluster sets form clustering assembly. The clustering assembly is used for calculation of irregularity index of an individual clusters and anomaly indexes for each cluster set, each cluster, and anomaly index of individual records in the test set. A decision for anomaly identification is done via analysis of anomaly indexes and selection of the threshold for anomaly identification.

Evaluation of efficiency of algorithms was done using data with anomalous records of two types. First one is the dataset with anomalous records containing high permeability values, exceeding 1000 mD, and regular records with smaller permeability values. Permeability dataset was published by Aase et al. in [1] and posted as an open source at the pubs.usgs.gov website. It contains 99 records. with eleven of them with permeability exceeding 1000 mD. Parameters used for clustering this dataset are porosity and grain size. Another anomaly type is anomaly with records collected from gas-saturated sands and regular records collected from brine-filled sands. In this case, number of anomalous and regular records both equal 25, and each record contains three parameters – V_p, V_s, and rock density ρ. These data were published by Ramos. and Castagna, in S9].

Authors used NbClust R clustering package [3]. This package includes 9 clustering methods and a large number of criteria for selection of the optimal number of clusters. Euclidean distance in detection of anomalies of both types was used. In the case of high permebility anomaly, decision about optimal number of clusters was defined by majority of all criteria, and by using clustering criterion "silhouette" in the case of gas-sand anomaly.

2 Analysis of Inhomogeneity of the Training and Test Sets and Instability of Clustering

Ideally, all regular records in the union of the train and test sets form a single cluster, so that records outside this cluster are anomalous. In fact, this usually does not happen, and both the train and the test sets may be broken into several clusters. Another issue, that complicates anomaly detection clustering methodology, is instability of clustering process, where minor change in the clustered data may lead to significant change in the number of clusters and in their size. This is illustrated by Figure 1 that shows histograms of the number of clusters and their sizes in repeated clusteing of the dataset with records collected from brine-filled sands areas. Clustering was performed on the same dataset with repetedly randomized reordering of the clustered records and removing five records from the dataset. One can observe, that the dataset 'brine' is inhomogeneous containing clusters of different size, which number varies in the wide range. Figure 1 also shows, that there are clusters as small as one or two records. Therefore, using small cluster size as indicator of geologic anomaly is problematic. To overcome

DOI: 10.7569/JSEE.2018.629516

the problem of clustering instability and questional relations between cluster size and its potentional anomaly, authors developed anomaly detection algorithms, that rely on the use of clustering assemblies containing large number of individual cluster sets. Each cluster set is generated by clustering the randomized test set, that contains records, some of which labeled as regular and others - unlabeled records. Clustering assembly is utilized to formulate criteria for identification of anomalous cluster sets, clusters, and anomalous records.

Due to instability of individual clusters, anomaly detection, that rely on the use of individual cluster set, produce unreliable results. This is illustrated by Figure 2, that shows three clusters and ellipses, drawn around clusters with confidence level 0.95 (R function dataEllipse). Clustering was done on randomized test set, that includes anomalous and regular permeability data. Two of the clusters are regular and one is anomalous. One can observe, that cluster ellipses overlap each other and poorly separated.

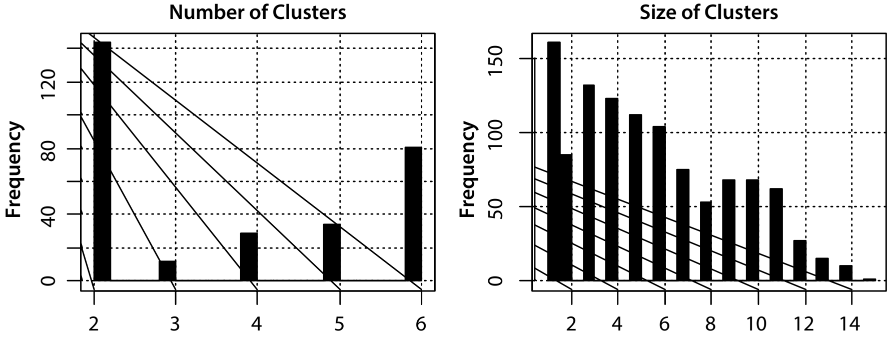

Figure 1 Histograms of the number of clusters and the number of records in individual clusters (cluster size) in the dataset containing records from brine-filled sands.

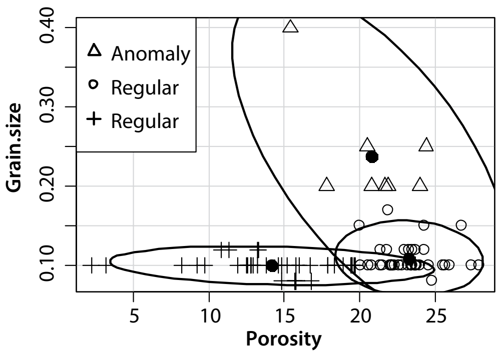

Figure 2 Clustering of a randomized test set that includes regular and anomalous records.

DOI: 10.7569/JSEE.2018.629516

3 Formation of Multiple Randomized Test Sets and Construction of the Clustering Assemblies

Algorithms presented in this paper rely on the use of the initial train set, that contains only regular records, labeled as "regular" and initial test set, that contains unlabeled records. Some of the unlabeled records may be regular others be anomalous. To obtain reliable identification of anomalous and regular records in the initial test set, the authors constructed multiple randomized test sets, built as the union of randomly formed subsets of the initial train and test sets. Randomization is done by Monte Carlo resampling of records of the train and test set. Clustering of multiple randomly-formed test sets results in the formation of an assembly of cluster sets *sCl(j)*, $1 \le j \le J$, that contains *J* cluster sets. Due to randomization, each of the records in the initial test set will appear multiple times in different cluster sets, so that multiple values of index *j* may be assigned to the same record. If the number of Monte Carlo runs is large enough, the number of appearances of individual records in the clustering assembly also will be large. This is illustrated by Figure 3. It shows values of the number of appearances of individual records in the clustering assembly build via repeated clustering of the randomized test set. Each randomized test set includes 15 randomly selected records from gas set and the same number of records from the brine set. Number of Monte Carlo runs is 300.

One can observe, that the number of appearances for any individual record exceeds 160. So large number of appearances of individual records in different clusters of the clustering assembly opens the way for building stable and reliable procedures for identification of anomalous records.

4 Irregularity Index of Individual Clusters in the Cluster Set

Further in this paper, the authors used the following notations: *m* is the index of the record in the union of the initial train and test sets, *j* is the index of the

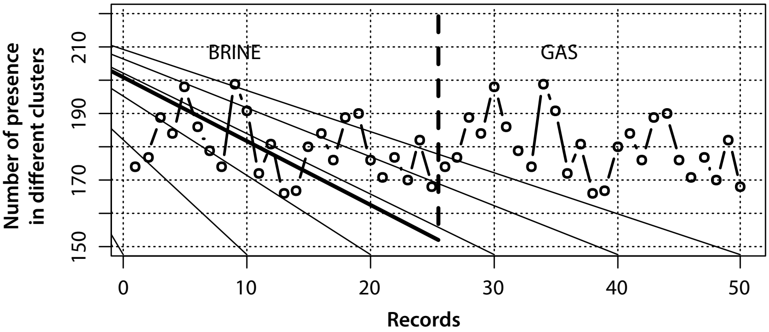

Figure 3 Number of appearances of individusl records in different clusters of the clustering assmbly.

DOI: 10.7569/JSEE.2018.629516

cluster set formed via clustering randomized test set, r is the index of the cluster in the cluster set with index *j*. Thus, each cluster is indexed by pair of indexes (*r, j*). Records in he initial trainset are labeled as regular, whereas records in the initial test set are unlabeled.

Key element of methodology of identification of anomalous and regular records, is the assumption, that due to the difference between properties of records of these two types, they will form different clusters. Based on this assumption, the authors introduce following parameters, that characterize possibility (not probability) that a given individual cluster and cluster set contain anomalous records:

a. Irregularity index, *Irreg*(*j*,*r*), of a cluster with index *r* in the cluster set with index *j*:

$$Irreg(j,r) = \frac{n.unlabeled(j,r)}{n.cluster(j,r)} \tag{1}$$

where *n.unlabeled*(*j*,*r*) and *n.cluster*(*j*,*r*) are the numbers of unlabeled records from initial test set and the total number of records in the cluster. According to Eq. 1, this parameter satisfies the following conditions:

$$0 \le Irreg(j,r) \le 1 \tag{2}$$

If irregularity index is close to 1, the majority of records in this cluster are unlabeled and a number of regular records is small. Consequently, this cluster may be identified as anomalous. On the other hand, if this index is close to zero, majority of its records are labeled as regular, and cluster should be identified as regular.

b. Index of anomaly presence in a cluster set *sCl*(*j*) is defined by Eq. 3:

$$AnClset(j) = \underset{r}{max}(Irreg(j,r)) - \underset{r}{min}(Irreg(j,r)) \tag{3}$$

According to Eq. 3, index of anomaly presence in the cluster set satisfies the following constraints:

$$0 \le AnClset(j) \le 1 \tag{4}$$

If irregularity index is the same for all clusters in the cluster set, then:

$$AnClset(j) = 0 \tag{5}$$

DOI: 10.7569/JSEE.2018.629516

On the other hand, if there is at least one cluster, that contains only labeled records from the train set and at least one cluster containing only unlabeled records, then:

$$AnClset(j)=1 \tag{6}$$

Values of anomaly index close to 1 are strong indication that there is anomalous cluster in the cluster set.

c. Anomaly index of an individual cluster is defined by Eq. 7:

$$AnCl(j, r) = Irreg(j, r)^{*}\ AnClset(j) \tag{7}$$

According to Eq. 7, anomaly index of a cluster is close to 1, if both its irregularity index and anomaly index of the cluster set are close to 1.

5 Anomaly Indexes of Individual Records and Clustering Assemblies

Specific feature of clustering assemblies, is that each record in the initial train and test sets will appear in a number of different clusters of the clustering assembly, and each cluster will be characterized by different values of irregularity index. This opens the way for construction of stable anomaly index for individual records in the studied dataset. In the following sections the authors test mean-aggregated anomaly index defined by Eq. 8:

$$rAnom(m) = \frac{1}{J(m)} \sum_{j=1}^{J(m)} Irreg(m, j, r) \tag{8}$$

where m is the index of the record, r is the index of the cluster containing record with the index m at Monte Carlo run with index j, $J(m)$ is the total number of appearances of record with index m in the clustering assembly.

Similarly, anomaly index of the whole clustering assembly is defined as:

$$anS = \frac{1}{J} \sum_{j=1}^{J} anClset(j) \tag{9}$$

Record with the index m will be assigned label "anomaly" or "regular" according to the following rule:

$$label(m) = \begin{cases} \text{“}anomaly\text{”};\ \ rAnom(\mathrm{m}) > tr \\ \text{“}regular\text{”};\ \ \ rAnom(\mathrm{m}) \le tr \end{cases} \tag{9}$$

where tr is the threshold, defined using prior false discovery rate (section 6).

6 Prior and posterior true and false discovery rates for anomalous and regular records

Concept of prior false discovery rate for individual records was introduced in [7]. Bellow, we expand this concept to prior estimates of false discovery rates of anomalous records, clusters, and anomalous cluster sets.

False discovery rates of anomalous records, clusters, and cluster sets in clustering assemblies are defined, respectively, as the fractions of individual records, clusters, and cluster sets, identified as anomalous, when all records in the clustering assembly are regular. Analysis of prior false discovery rates (FD) is done via repeated formation of a random test sets, built as the random subsets of the initial train set, and formation of the clustering assembly that contains only regular records. Identification of anomalous clusters, anomalous cluster sets, or anomalous records in this clustering assembly is the act of false discovery. Prior false discovery rate is used for selection thresholds in anomaly detection procedures to guarantee low posterior (actual) false discovery rate. Posterior false (FD) and true (TD) discovery rates are calculated using clustering assembly, built via repeated clustering of randomized test sets, that contain regular and anomalous records. Posterior (true) FD and TD are defined, respectively, as the fractions of regular and anomalous records identified as anomalous in this clustering assembly.

7 Estimates of prior false discovery rates for anomalous cluster sets, clusters, and individual records. Permeability dataset

In this section the authors present results of analyses of prior FD obtained using train set, that includes records with the regular permeability values smaller than 1000 mD. Figure 4 shows histograms of the three prior anomaly parameters, calculated using clustering assembly built via Monte Carlo resampling of this set. According to this figure, all three anomaly indexes vary in the range of 0.0 - 0.5 with maxima of histograms below 0.3. Low range of these parameters indicate absence of anomalous records in this clustering assembly.

Table 1 shows estimated prior false discovery rate of the cluster sets and clusters calculated as functions of threshold.

As shown in the Table 1, the prior false discovery rates for anomalous cluster sets and individual clusters in the cluster set drop below 0.01 at the threshold equal to 0.35.

8 Posterior analysis of efficiency of anomaly identification. High permeability anomaly

Assembly of individual cluster sets analyzed in this section and utilized for reliable anomaly identification includes 300 of individual cluster sets. Each individual

DOI: 10.7569/JSEE.2018.629516

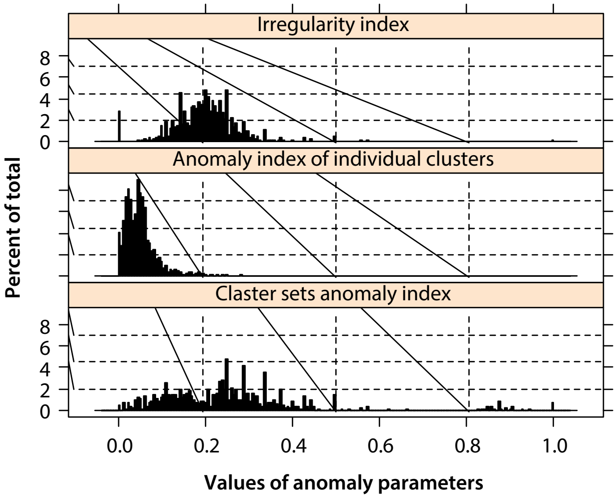

Figure 4 Histograms of three prior anomaly parameters.

Table 1 Prior rates of false discovery of anomalous clusters, and cluster sets

	Prior false discovery rate	
Threshold	**Cluster sets**	**Clusters**
0.05	0.864	0.079
0.1	0.717	0.033
0.15	0.626	0.012
0.2	0.426	0.005
0.25	0.273	0.002
0.3	0.162	0.001
0.35	0.098	0.001

cluster set is generated via clustering of a randomized test set constructed as the union of initial test set and randomly-generated subset of the training set. Initial test set contains 11 anomalous records and 10 regular ones. Figure 5 shows histograms of the distribution of irregularity index and anomaly indexes of cluster sets and individual clusters. Importantly, all 300 cluster set anomaly indexes are large and exceed 0.6. This is strong indication of the presence of anomaly in every randomized test set. Histogram of the anomaly index of individual clusters shows

DOI: 10.7569/JSEE.2018.629516

presence of two cluster groups – clusters with high anomaly index exceeding value of 0.5 that are potentially anomalous, and those with anomaly index as small as 0.1. Histogram of the irregularity index shows similar pattern with the presence of a group of irregularity indexes with values exceeding 0.8.

Table 2 shows parameters of clusters with posterior cluster anomaly index exceeding 0.5. According to this table, there is a wide range of sizes of clusters of this type, whereas three anomaly indexes are all in the narrow ranges. This

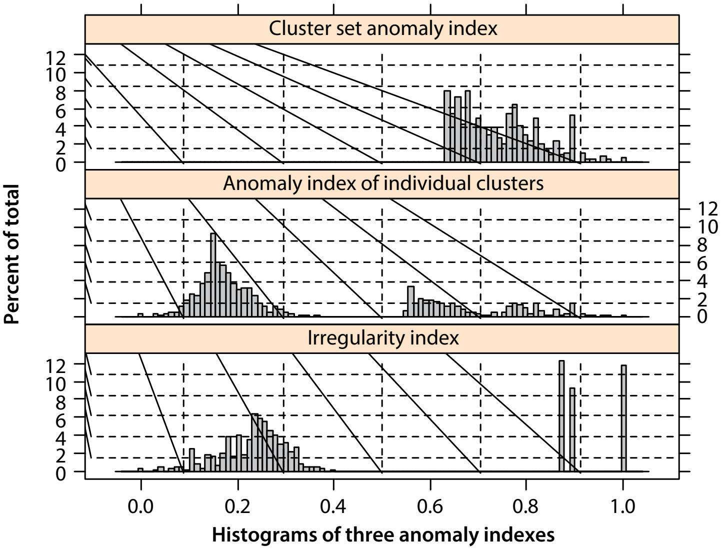

Figure 5 Histograms of three posterior parameters characterizing presence or absence of anomalous records in the randomized test sets.

Table 2 Parameters of clusters, labeled as anomalous with individual cluster anomaly index exceeding 0.5.

Cluster size	Number of clusters	Mean values		
		Irregularity index	Anomaly index of clusters	Anomaly index of cluster sets
1	9	1.000	0.814	0.814
3	1	1.000	0.833	0.833
7	215	1.000	0.813	0.813
8	368	0.900	0.646	0.714
9	395	0.908	0.659	0.723
10	1	1.000	0.923	0.923

DOI: 10.7569/JSEE.2018.629516

is another indication, that the size of the cluster alone is not reliable indicator of anomaly of records in this cluster.

According to the Table 2, majority of clusters with large anomaly indeces are not small and contain at least 7 records.

Efficiency and reliability of identification of anomalous records are illustrated in Figure 6, which shows plots of anomaly index of all records (regular and anomalous) in the permeability dataset. Horizontal dashed line is drawn at the threshold level equal to 0.4. There is only one falsely identified regular record with anomaly index as high as 0.8. Anomaly index of other regular records is smaller 0.4. Among eleven anomalous records there are eight records with anomaly index exceeding 0.4. These records correctly identified as anomalous. There are also three anomalous records with anomaly index smaller than 0.4. These records are falsely identified as regular. True discovery rate in this case is around 73%.

9 Identification of Records in the Gas Sand Dataset as Anomalous, using Brine Sand Dataset as Data with Regular Records

The authors present in this section results of analysis of accuracy of detection of gas sand anomaly, with initial test and train sets formed as gas and brine datasets. Randomized test sets were repeatedly built as the unions of the randomly formed subsets of 15 records from initial train and test sets. The main complicating factor of detecting gas-sand anomaly is inhomogeneous structure of gas and brine datasets, so that both train and test set contain several clusters.

Figures 7 and 8 show distributions of prior and posterior values of irregularity index and anomaly indexes of individual clusters and cluster sets. To calculate the prior indexes, randomized train and test sets were formed as not intersecting subsets of the brine dataset. Posterior indexes were calculated using brine and gas datasets as initial train and test sets. One can observe significant differences

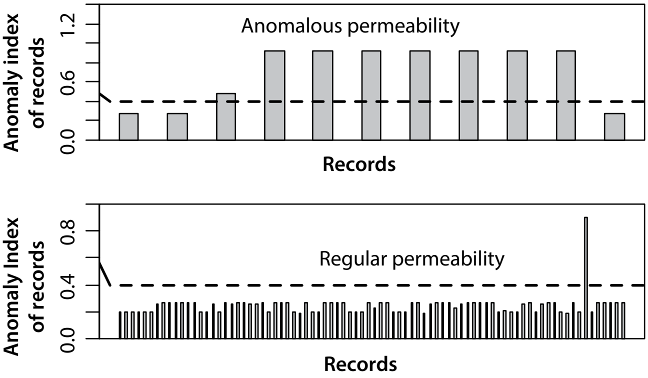

Figure 6 Anomaly indexes of individual records. High permeability anomaly.

DOI: 10.7569/JSEE.2018.629516

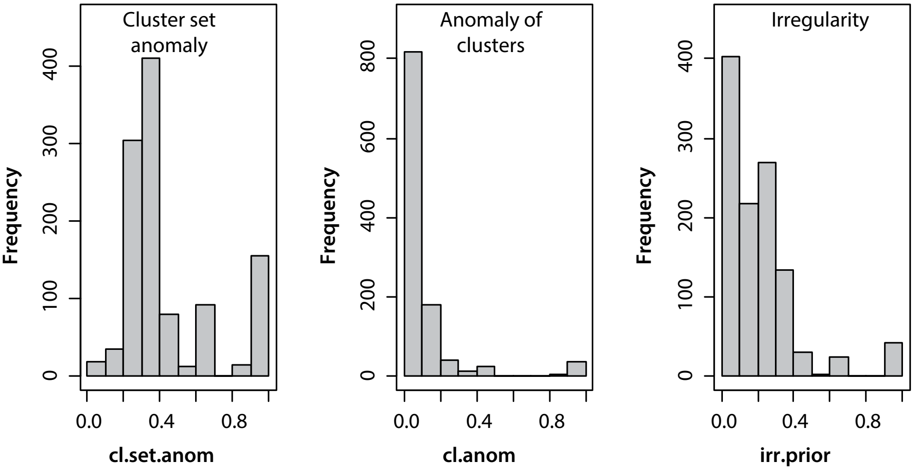

Figure 7 Histograms of three prior anomaly indexes.

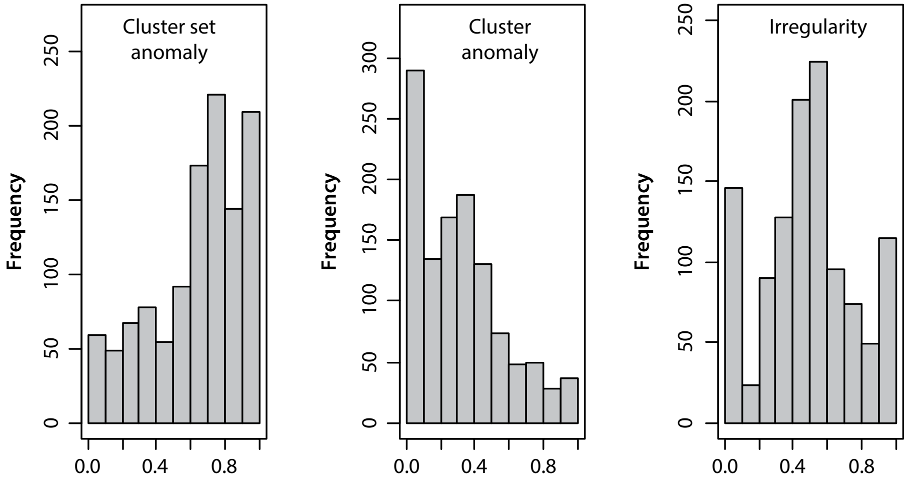

Figure 8 Histograms of three posterior anomaly indexes.

in distributions of prior and posterior indexes shown at these figures. According to Figure 7, absolute majority of all three prior indexes do not exceed threshold of 0.6. On the other hand, large number of values of posterior indexes, shown in Figure 8, is larger than this threshold. This indicates the possibility of detection of anomalous records, that form the initial test set.

Figure 9 illustrates efficiency of detection of gas-sand anomaly using clustering assemblies. It shows values of mean-aggregated anomaly index of individual

DOI: 10.7569/JSEE.2018.629516

records (Eq. 8) in brine-filled and gas-filled sand datasets. According to this figure, absolute majority of values of anomaly index of the records in the brine dataset is smaller than 0.31, whereas majority of the records in the gas dataset are characterized by values of this index exceeding this value. Accuracy and stability of detection of gas sand anomaly is illustrated by Table 3. It shows thresholds, true, and false discovery rates of anomalous records calculated using three independently generated clustering assemblies. Thresholds were calculated as quantile values of anomaly indexes for records in the brine dataset with the same quantile probabilities in all three clustering assemblies.

According to Table 3, values of true and false discovery rates obtained using three independently generated clustering assemblies show minor differences. For

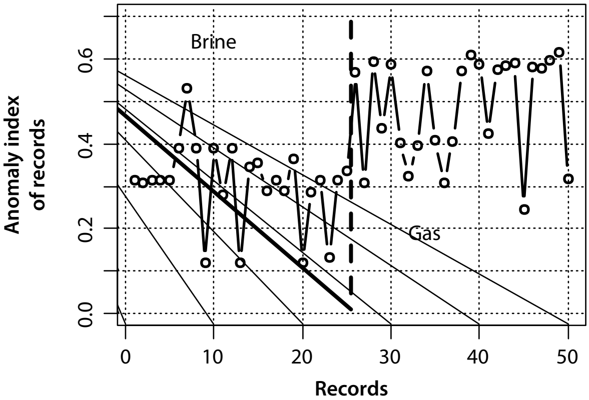

Figure 9 Values of anomaly index of individual records in gas-sand and brine-sand datasets.

Table 3 Thresholds, true, and false discovery rates of anomalous records using three independently generated clustering assemblies. 300 Monte Carlo runs.

Quantile probabilities	Thresholds			False discoveries			True discoveries		
	1	2	3	1	2	3	1	2	3
0.6	0.32	0.32	0.34	0.32	0.36	0.36	0.8	0.84	0.96
0.7	0.35	0.34	0.36	0.24	0.2	0.32	0.8	0.8	0.8
0.8	0.38	0.36	0.38	0.2	0.2	0.2	0.8	0.8	0.8
0.9	0.38	0.38	0.39	0.04	0.04	0.04	0.8	0.8	0.8
0.95	0.38	0.38	0.39	0.04	0.04	0.04	0.8	0.8	0.8

DOI: 10.7569/JSEE.2018.629516

all three clustering assemblies, true discovery rates not smaller 0.8, with false discovery rates as low as 0.04.

10 Notations

Gas set - dataset that contains records from gas-filled sands, brine set - dataset that contains records from brine-filled sands.

TD and FD – true and false discovery rates.

j is the index of the cluster set formed via clustering randomized test set, *r* is the index of the cluster in the cluster set with index *j*. Thus, each cluster is indexed by pair of indexes (*r*, *j*).

Irreg(*j*, *r*) - irregularity index of the cluster with index *r* within cluster set with the index *j*.

n.unlabeled(*j*, *r*) and *n.cluster*(*j*, *r*) are the numbers of unlabeled and the total number of records in the cluster.

AnClset(*j*) - index of anomaly presence in a cluster set *sCl*(*j*).

AnCl(*j*, *r*) - anomaly index of individual cluster.

rAnom(*m*) anomaly index of individual record.

J(*m*) is the total number of appearances of the record with index *m* in the clustering assembly.

TD(*tr*), *FD*(*tr*) - posterior true and false discovery rates, *tr*-threshold in anomaly detection rules.

11 Conclusions

- Authors present new algorithms for detection of geologic anomalies and results of their testing. Key element of the developed algorithms is construction of multiple randomized test sets, and construction of multiple cluster sets, that form clustering assembly. The clustering assembly is used for calculation of irregularity index of individual records and anomaly indeces for each cluster set, each cluster, and individual records in the test and train sets.
- The algorithms were tested on the data with anomalies of two types: (a) high permeability anomaly with regular records containing smaller permeability values, and (b) gas-filled sand anomaly with regular records containing data from brine-filled sands. Results of algorithms testing, presented in the paper, demonstrate high stability of anomaly detection with true discovery rate higher than 73% with false discovery rates equal or even lower than 0.2 for both anomaly types.

Reference

1. Aase, N., Bjorkum, P., and Nadeau, P., 1996, The effect of grain-coating microquartz on preservation of reservoir porosity. Bull. Am. Assoc. Petrol. Geologists 80, (1): 1654-1673.

DOI: 10.7569/JSEE.2018.629516

2. Aminzadeh, F. and Chatterjee, S., 1984, Applications of cluster analysis in Exploration Seismology, Geoexploration, 23: 147-159.
3. Charrad M, Ghazzali N., Boiteau V., Niknafs A., 2014. NbClust: An R Package for determining the relevant number of clusters in a data set. Journal of Statistical Software, 61 (6): 1-36.
4. Chilingar G., Mazzulo S., Rieke, H., 1992. Carbonate Reservoir Characterization: A Geologic-engineering Analysis. Elsevier, 639 pp.
5. Dvorkin J, Mavko G., Nur A., 1999, Overpressure detection from compressional and shear-wave data. Geophysical Research Letters, 26, (22): 3417–3420.
6. Gurevich A., Chilingar G, and Aminzadeh F., 1994. Origin of the formation fluid pressure distribution and ways to improving pressure prediction methods. J. Pet. Sci. Eng., 1: 67-77.
7. Katz S., Aminzadeh F., Chilingar G., Khilyuk L., 2015. Anomaly detection within homogeneous geologic area. J. of Sustainable Energy Engineering, 3, (2): 169-186.
8. Lian Duan, Lida Xu, Ying Liu., 2008. Cluster-based outlier detection. Annals of Operations Research, 168, (1): 151–168
9. Ramos, A. and Castagna, P., 2001. Useful approximations for converted-wave AVO. Geophysics, 66: 1721–1734.
10. Zimek, A., Campello, R., Sander, J.. 2014. Ensembles for unsupervised outlier detection. ACM SIGKDD Explorations Newsletter. **15**: 11–22.
11. Zengyou He, Xiaofei Xu, Shengchun Deng. 2003. Discovering cluster-based local outliers. Pattern Recognition Letters, 24(9-10):1641-1650.

DOI: 10.7569/JSEE.2018.629516