Data Mining Methodologies to Reduce the Uncertainty of Reservoir Selection*

Jin Fei[1], Jeffrey Yarus[1], and Richard Chambers[1]

Search and Discovery Article #41814 (2016)**
Posted September 6, 2016

*Adapted from extended abstract prepared in relation to an oral presentation at AAPG 2016 Annual Convention and Exhibition, Calgary, Alberta, Canada, June 16-22, 2016

[1]Halliburton, Houston, TX (jin.fei@gmail.com)

Abstract

Geostatistical-based earth modeling can create hundreds of reservoir property realizations. The challenge is to select a few optimal models from these realizations for further analysis. Realizations are often ranked according to various criteria based on pore volume, connected volume, or drained volume. Traditional workflows then select candidates from ranked distributions, representing quantiles (e.g., P10, P50, and P90) of the reservoir volumetrics, which assume contiguous reservoir volumes. However, such methods lack any physical geometrical information. Geostatistical reservoir modeling could result in realizations that have unconnected pore volumes equaling pore volumes of other realizations with connected pores.

This article discusses two novel ways to choose simulation models from multiple realizations. Before the application of data mining techniques, a three-dimensional (3-D) realization is collapsed into a one-dimensional array so that each element of the array maps to one single [*i, j, k*]-indexed cell. This array carries the spatial property information. A matrix is then created from multiple realizations. Each column represents an individual simulation, and each row represents a single 3-D grid cell.

The first method introduced is a two-way clustering of both columns (R-mode) and rows (Q-mode). It is a new method that quantitatively evaluates a grid cell by cell throughout the realizations. During the cluster analysis, the spatial-property location of the cell is carried in the distance calculation. R-mode clustering shows the similarity of different simulations. Conversely, Q-mode clustering shows the grouping of different cells. The location and connectivity of cells can be mapped, in addition to evaluating their statistical ranking.

The second method uses classification and regression trees to select the reservoir simulation that is strongly associated with other petrophysical properties or seismic attributes. It can be associated with both numerical and categorical properties.

Both methods were applied to a West Texas Permian basin reservoir. A statistical analysis of the connected reservoirs was performed. Through the above novel quantitative evaluations, geologists and reservoir engineers can select optimistic to pessimistic realizations to aid in the

economic assessment of the reservoir. These realizations help to reduce the uncertainty when applying volumetric histogram cutoffs, particularly in shale plays where the reservoirs may not be contiguous.

Introduction

Data mining has been widely used in the oil and gas industry to automatically characterize data through pattern recognition techniques. In this study, we focused on hierarchical clustering and classification and regression tree (CART) methodologies and the application for selecting optimal reservoir models. Both approaches are intuitive, easy-to-use methods for geologists and engineers to communicate.

Applications of an unsupervised hierarchical clustering algorithm can be found in the oil and gas industry. This algorithm generates natural groupings in petrophysical, geological, geomechanical, or seismic data, such that members of one group are similar to each other but dissimilar to members of other groups. It is mainly used to predict lithofacies from well logs (e.g., facies, porosity, saturation, and permeability) (Adoghe et al., 2011). To overcome the increasing computational complexity, a cluster analysis of logging data from stratigraphic zonation was introduced to merge adjacent zones into larger clusters (Gill et al., 1993).

A decision tree is another supervised data mining technique. CART is a typical approach, with a more generic tree-growing process, and is defined as a classification tree or regression tree depending on if the response variable is categorical or continuous, respectively. CART was brought into general use in the medical field for the outcome prediction of triage heart patients (Breiman et al., 1984). This technique is ideally suited for the analysis of the complex datasets used in the petroleum industry. Recently, permeability was predicted using well logs and classified into homogeneous clusters based on a statistical regression tree (Perez et al., 2005). Literature has investigated the traditional permeability prediction method on the basis of electrofacies, lithofacies, and hydraulic flow units. The advantage of the approach was to handle missing log data, which is a common occurrence in field acquisition. Yarus et al. (2006) predicted the production of a well from several variables, such as producer, acid volume, and strength. CART was also applied to predict job pause time in well treatment programs to find patterns between subsequent pumping (Marko et al., 2015). The prediction was improved with normal score transform to normalize data and k-means clustering to reduce data.

To evaluate the hydrocarbons of a reservoir, geologists and reservoir engineers typically choose optimal reservoir models out of hundreds of geostatistical property realizations. Properties, like porosity and permeability, are ordered by rank. Figure 1 shows the cumulative probability distribution curve and the six reservoir realizations that are close to the P50 quantile distribution. This quantile distribution of volumetrics is related to contiguous reservoir pore volume. P10 represents an optimistic result with a large connected volume, and P90 is a conservative result with a smaller connected volume. P50 is typically considered the optimal reservoir simulation. However, this method only considers probability distribution but lacks the physical geometrical information. Among the geostatistical realizations, a realization of unconnected pores, with neither horizontal nor vertical permeability, may have equaling pore volumes to one another with connected pores. However, the similarity of the two realizations is low. Selecting the optimal reservoir simulation using this approach increases the uncertainty, because the sweet spots could differ geometrically.

In another scenario, domain experts select the realizations that correspond to other seismic attributes or petrophysical properties. Adding lithological constraints or co-located co-Kriging with other properties (e.g., permeability or acoustic impedance) can reduce the uncertainty. However, there are more properties that cannot be taken into consideration when adding such constraints. CART is an analytical tool that is capable of finding such relationships between realizations and associated properties. In the following sections, we describe how to preprocess data for the analysis. Clustering and CART analyses are then introduced. A case study performed in West Texas is also presented. From the clustering analysis, geobodies with similarities across realizations are automatically extracted. A statistical analysis of the connected geobodies identifies the optimal realization for reservoir simulation and well placement. From the CART analysis, realizations that correspond well with other properties are ranked. The interaction between the resulting map and tree and the geocellular grid assists in distinguishingdifferent realizations so as to reduce the uncertainty when selecting the candidate reservoir model. The paper concludes with discussions on the feasibility of future applications.

Methodology

A realization is represented by a 3-D grid of total *n* cells, denoted by ***Si***, where *i* is the *i*th realization and $i \in [1, m]$ (Figure 2). In the Euclidean coordinate, each cell is indexed by [*i, j, k*] and carries its geometrical location of the data; the property is color-coded and mapped onto the grid. If the size of a grid is 6×7×5, when reshaped as a one-dimensional (1-D) array, the index of a cell is ×6×7+ ×7+ . For example, the cell at [0, 6, 3] will be the 168th element of the 1-D array. Each realization is reshaped into a 1-D array with 6×7×5 elements before further calculations.

After all the geological property realizations are organized into arrays, they are placed into a matrix. Each column represents a different realization, and each row indicates a single cell of the 3-D grid. Clustering and CART analyses are then applied.

Clustering Analysis

Hierarchical clustering creates a hierarchy of the given set of data. A matrix of similarities or distances of the attributes is computed. A clustering algorithm is then iteratively applied to the similarity matrix. The pairs of data with the greatest similarities are merged, and the matrix is recomputed. Ultimately, all data will be linked together in a hierarchy, which is commonly illustrated as a dendrogram tree.

There are many methods to merge clusters, but experience seems to suggest that the weighted-pair group (WPGM) method tends to be superior to single linkage or unweighted average methods, as it produces a higher cophenetic correlation coefficient, leading to minimal distortion of the dendrogram. The essential steps of the WPGM method with arithmetic averaging are as follows (Davis 2003):

- Calculate the correlation coefficient matrix from realizations (R-mode),
- Calculate the Euclidean distance matrix from grid cells (Q-mode),
- Link pairs of data or clusters with the greatest similarities,
- Connect two objects that demonstrate the greatest similarity to each other,
- Average similarity coefficients with all other objects after two datasets or clusters are combined.

Figure 3 depicts the clustering of petrophysical property simulations (e.g., porosity). Though the realizations are all close to the P50 quantile distribution, the geological locations with pore could be dramatically different. Each realization, ***Si***, is placed on the column with an array of grid cells. The geological location of each cell is mapped to one row across all the realizations. The results of R-mode clustering are represented at the top of the hierarchical dendrogram tree. From the resulting clusters, geologists can choose the proper candidates following the hierarchies. In this example, [***S5, S1, S8***] is grouped and linked with the [***S2, S4***] group and then connected with the [***S3, S6, S7***] group. The three clusters are marked as green, red, and blue, respectively. The clusters from Q-mode clustering are shown on the right of the hierarchical dendrogram tree. Given the proper cutoff, two clusters form; the top cluster includes two red groups, which links with the bottom cluster of two blue groups. The clusters are geometrically mapped on the geological grid. If red represents higher porosity, then the top-cluster grid cells more than likely indicate the pay zone. Further studies of the geometrical distribution and connected geobody are necessary to evaluate the sweet-spot mapping of all different simulations.

Classification and Regression Trees (CART)

CART is an important machine learning technique in the oil and gas industry. It is intuitive, by nature, to classify a pattern by asking a series of questions. This technique presents rules for predicting the response of both categorical variables, such as lithofacies, and continuous variables, such as permeability. CART is graphically presented as a classification or regression tree, which explains the variation of a single response variable by repeatedly splitting the data into homogeneous groups. The main difference between the classification tree and the regression tree is their dependent variable. For the classification tree, the dependent variables are categorical, while the regression tree has numerical dependent variables. The decision trees are:
(1) flexible enough to handle a broad range of response types, including numeric and categorical data;
(2) easy to construct and interpret; and
(3) able to handle missing values in both response and explanatory variables.

Thus, CART represents an alternative to many traditional statistical techniques (e.g., variance and linear discriminant analyses). The tree structure also makes it easy to determine the relative importance of different realizations during data classification and to account for missing data.

CART recursively partitions datasets into subgroups. The process of partitioning is also called splitting. The root of the binary tree consists of all data, which are then split into two offspring nodes. The process continues until certain criteria are met. After the splitting process stops, it can be pruned by eliminating offspring nodes. An optimal tree is selected using the above process. Categorical and continuous variables follow different splitting rules:

- The common splitting rules for categorical variables involve the Gini or entropy indices. Assuming p is the percentage of each category,
  - The Gini index is measured as $1 - \Sigma\, p2$,
  - The entropy index is measured as $-\Sigma\, p \ln(p)$
- Continuous variables are split using the residual sums of squares in the form of $(\Sigma[y\text{-}E(y)]^2)$, where y is the response variable.

These splitting rules are normally referred to as the impurity of a node, N, $i(N)$. Ideally, we expect $i(N)$ to be 0 if the subgroup of the node bears the same category label. Heuristically, to continue splitting is to decrease the impurity as much as possible. The function for the drop in impurity criteria is defined by:

$$\Delta i(N) = i(N) - P_L i(N_L) - (1 - P_L)\, i(N_R)$$

Where, N_L and N_R are the left and right offspring nodes, $\boldsymbol{i}(N_L)$ and $\boldsymbol{i}(N_R)$ are their impurities, and P_L is the percentage of the node, N, that will go to N_L. The best variable to split is to maximize the above function, though we may have to perform an extensive or exhaustive search of the subsets to find the rule for maximizing it.

Case Study

Data mining techniques are used to select optimal reservoir models. Such techniques were recently applied to an interbedded carbonate-elastic reservoir in the West Texas Permian Basin (Figure 4). Structurally, the reservoir is a north-south-trending asymmetrical anticline, gently dipping eastward into the Midland Basin. Lithologically, the Permian formation is composed of alternating dolomites and siltstones for a total thickness of about 140 m. The formation is divided into five dolomite intervals (DI–D5), each separated by a siltstone interval (Sl–S3), except for the D2/D3 dolomite intervals, which are separated by a 1–3-m thick shale layer. Within the study area, the second siltstone (S2) is the most productive interval. Dolomites tend to be non-productive. S1 and S3 intervals have only a minor contribution to the total production in this part of the field. Therefore, this study focuses on the S2 interval.

A structural framework was constructed out of a total of 32 vertical wells from a 2×3-km field. The target interval is the S2 siltstone. Fifty realizations of porosity properties were simulated using Turning Band (Matheron, 1972; Chilès et al., 2005). Volumetrics are then calculated and ranked. A cumulative probability distribution representing quantiles of the reservoir volumetrics is presented in Figure 1. Cluster and CART analyses were performed on all the realizations.

A two-way clustering algorithm can be used to distinguish between different realizations in terms of the geometrical locations. Figure 5 only shows clustering results of the six realizations closest to P50. Porosity property realizations are displayed as a color-coded map in plot (a). A group-linkage dendrogram of different realizations is illustrated at the top; the three clusters from the geocelluar grid are displayed on the right, and the corresponding geobodies of each cluster are visualized in plot (b). Cluster 2 apparently contained higher porosity, while Cluster 3 contained lower porosity; the remaining cells belonged to Cluster 1. In plot (c), the profile curves delineated statistical distributions of each cluster. The Cluster 2 profile peaked around 0.18 (18%), which most likely indicates high-production zones or sweet spots.

A further study of connectivity can help in selecting the connected reservoirs, consequently optimizing reservoir simulation and well placement. After searching the neighborhood for connected cells, connected geobodies are identified. From the Q-mode cluster analysis, cells with similar properties across each of the realizations are grouped together; their geometrical locations form the geobodies, which are subregions of the grid, and their shapes are consistent across each of the realizations. By interactively selecting the cells from the hierarchy of the dendrogram tree, the geobody can be refined. Traditionally, a threshold cutoff of a property is specified; however, the subregional geobodies are different from one realization to another, or a subregion is selected, but the property distributions mapped on the goebodies may

not be similar. The new method provides an alternative means for studying each of the realizations of consistent geobodies and similar property maps.

Figure 6(a) is an example to illustrate that the four largest geobodies were identified. Geobody #1, with 1,890 cells, exhibited the largest volume, and geobody #2 was more spread out, indicating a thinner reservoir. In plot (b), porosity properties of each of realizations were mapped onto an identical geobody. The similarity of property maps was greater if two realizations were from the same R-mode realization clusters (realizations 13 and 16). As shown in Figure 7, if the traditional porosity cutoff [25%, 30%] was specified, the geobodies from realizations 10 and 15 were significantly inconsistent.

A statistics analysis of properties mapped onto the geobodies was performed to quantitatively study the uncertainty when selecting the optimal reservoir candidate models. The porosity means (μ) and standard deviations (σ) of the six realizations closest to P50 were calculated (Table 1). It turned out that realization 10 had the highest μ (0.2000) but the lowest σ (0.0399), which represented the best candidate; on the other hand, realization 22 represented the worst candidate, having the lowest μ (0.1822) but highest σ (0.0506). In fact, although the closest realization to the P50 quantile is realization 22, it could potentially be the worst candidate. Therefore, simply selecting the candidate realization from P50 may not be the best option.

To quantify the significance of the difference in the means, a student *t*-test is analyzed. The value of *t* is 1.962 at the significant level $\alpha=0.05$. The pair-by-pair calculations of *t* values were all higher than 1.962; therefore, the analysis rejected the null hypothesis that all of the realizations possessed the same average porosity. In particular, realization 10 is different from the rest at a significance level of 0.05, but represents the best candidate among the six realizations.

In many cases, domain experts search for a reservoir model that strongly corresponds to the petrophysical and geomechanical properties and seismic attributes. CART is a supervised learning technique and is applicable to both continuous and categorical variables. In our study, the realizations are the input variables and the associated variables (e.g., lithofacies) are treated as the predictor. A decision tree binarily and recursively splits the geocellular grid into homogeneous groups. As the process continues, the splitting criteria are a series of ranked property realizations that can best reduce the variation of the response variable. Such realizations, which form a path from the root to the leaf of the tree, are candidates for purifying the subset data.

In the study, each of the realizations (***Si***) was used to predict lithology facies. Figure 8 shows an optimal classification tree from the above process. The interpretation is straightforward, with the information at each node presenting the percentage of each category, as well as the percentage of samples of the total dataset. For example, there were three lithology types: dolomite, limestone, and shale at the tree root, each accounting for 16%, 28%, and 56% of the whole data (i.e., all the grid cells). Gini-splitting rules were chosen. The tree began splitting at the root to find the best possible realizations to classify the lithofacies. The first split was based on the porosity value at a 7.4% cutoff. If the porosity value of ***S23*** was greater or equal to 7.4%, it was classified as shale of the right-most offspring node that accounted for 75% of all the data, which was primarily composed of shale (72%). Therefore, ***S23*** could best distinguish shale. The left tree continued to split. ***S23*** was again selected to classify limestone (76%), where the porosity was less than 6.8%. If another realization follows the splitting path, it is also highly possible that it will be ranked as a candidate to correspond with certain facies.

If nodes from the tree are selected, the related geobodies are highlighted in the grid. In Figure 8, the right-most node of the tree was most likely mapped to associated shale grid cells. A connectivity and statistical analysis of the geobodies can be further performed in the same manner as the cluster analysis.

Conclusions

The traditional process for selecting a candidate realization from representative probability distribution quantiles is problematic. In addition to assuming contiguous reservoir volumes, this approach lacks physical geometrical information and does not guarantee unique connectivity or similar distribution of rock properties.

This article discusses two data mining methodologies for geologists and reservoir engineers to easily choose optimal models from multiple realizations. A two-way clustering analysis is a new method that quantitatively evaluates a grid cell by cell throughout each of the realizations. R-Mode clusters groups of realizations with similar property distributions; Q-Mode analysis classifies cells across the realizations to identify geobodies with identical geometries. Unlike the unsupervised clustering technique, CART is supervised, but is simple and easy to understand. This makes it a very effective tool for data interpretation and communication between engineers and geoscientists. CART allows for seeking realizations that correlate with any existing continuous or categorical properties. By traversing the tree, different classes map each geobody of the geocellular grid.

The connectivity analysis ranks the volume of connected geobodies, which are identical among all the realizations. The statistical analysis quantitatively identifies the optimal candidate for each quantile (P10, P50, and P90). The clustering map or the CART tree can act interactively with the grid to quality check the selections.

Both the two-way cluster analysis and CART are valuable techniques for selecting the candidate reservoir and reducing uncertainty. The methods are applicable to an unstructured grid once indexed and reshaped into one dimension. K-means may be an alternative supervised clustering technique to offer similar results but improved performance. Through such quantitative evaluations, geologists and reservoir engineers can select optimistic to pessimistic realizations to aid in the economic assessment of the reservoir, thereby reducing uncertainty when applying volumetric histogram cutoffs.

Acknowledgments

The authors thank Halliburton for permission to publish this work. We would also like to thank Genbao Shi who assisted in the research.

Selected References

Adoghe, L.I., O.S. Aniekwe, and C. Nwosu, 2011, Improving electrofacies modeling using multivariate analysis techniques: A deepwater turbidite case study:. Paper SPE 150776, presented at the Nigeria Annual International Conference and Exhibition, Abuja, Nigeria, 30 July–3 August.

Breiman, L., J.H. Friedman, R.A. Olshen, and C.G. Stone, 1984, Classification and Regression Trees: CRC Press, Wadsworth International Group, Belmont, CA.

Chambers, R., M.A. Zinger, and M.C. Kelly, 1994, Constraining geostatistical reservoir descriptions with 3-D seismic data to reduce uncertainty, *in* J.M. Yarus and R.L. Chambers, editors, Stochastic Modeling and Geostatistics: AAPG Computer Applications in Geology, no. 3, p. 143-158.

Chilès, J-P., and C. Lantuéjoul, 2005, Prediction by conditional simulation: Models and algorithms, *in* M. Bilodeau, F. Meyer, and M. Schmitt, editors, Space, Structure, and Randomness: Lecture Notes in Statistics, v. 183,p. 39-68.

Davis, J.C., 2003, Statistics and Data Analysis in Geology: John Wiley and Sons, Inc., New York, NY, chapter 6, p. 461–594.

Doveton, J., 1994, Geological Log Interpretation Using Computer Methods. AAPG Computer Applications in Geology, no. 2, 256p.

Gill, D., A. Shomrony, and H. Fligelman, 1993, Numerical zonation of log suites by adjacency-constrained multivariate clustering: AAPG Bulletin, v. **77**/10, p. 1781–1791.

Matheron, G., 1972, The turning bands: A method for simulating random functions *in* IRn. Technical Report N-303: Centre de Morphologie Mathèmatique, Ecole des Mines de Paris.

Maucec, M., A. Singh, S. Bhattacharya, J. Yarus, D. Fulton, and J.M. Orth, 2015, Multivariate analysis and data mining of well-stimulation data by use of classification-and-regression tree with enhanced interpretation and prediction capabilities: SPE Economics & Management, v. **7**/02, SPE 166472.

Perez, H.H., A. Datta-Gupta, and S. Mishra, 2005, The role of electrofacies, lithofacies, and hydraulic flow units in permeability predictions from well logs: A comparative analysis using classification trees: SPEREE, v. 8/02, p. 143–155.

Yarus, J., M. Srivastava, and R. Chambers, 2006, Geologic success but economic failure: Uncovering hidden problems using recursive partitioning (abstract): AAPG Annual Convention and Exhibition, Houston, Texas, 9–12 April, Search and Discovery Article #90052 (2006). Website accessed August 28, 2016, http://www.searchanddiscovery.com/documents/2006/06088houston_abs/abstracts/yarus.htm.

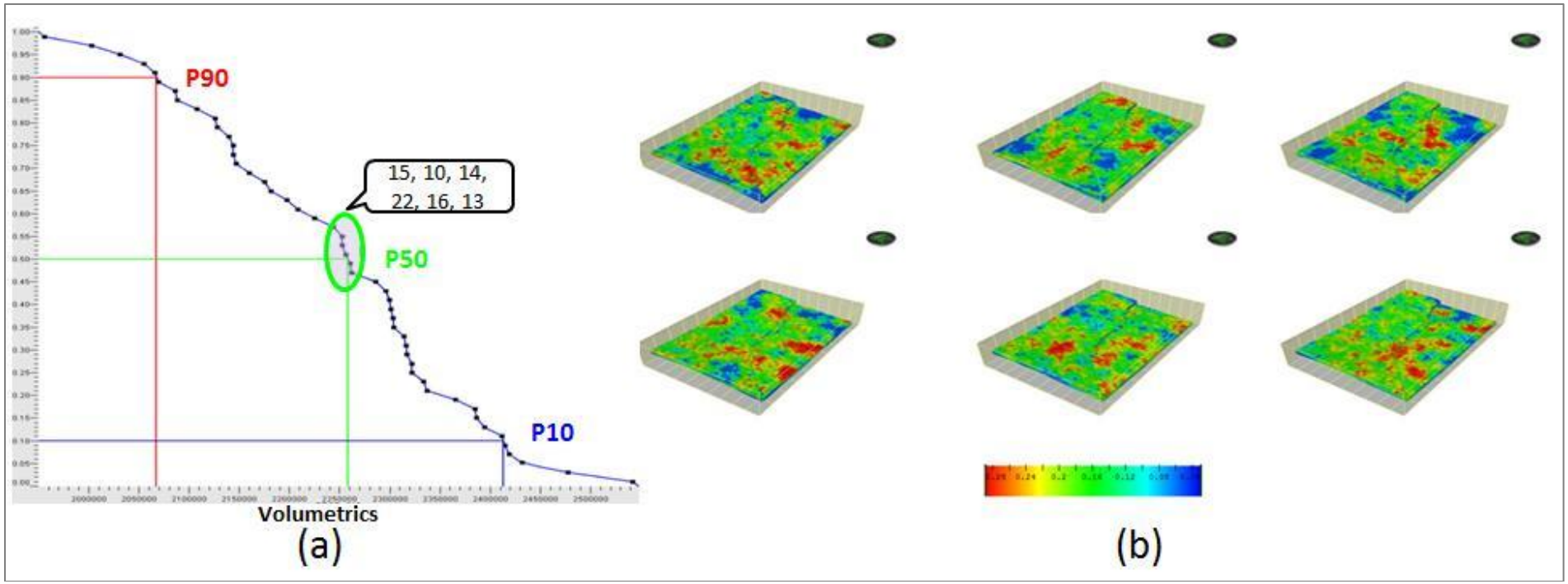

Figure 1. 3-D reservoir simulations at quantile P50. (a) Cumulative probability distribution curve; (b) Six simulations close to P50.

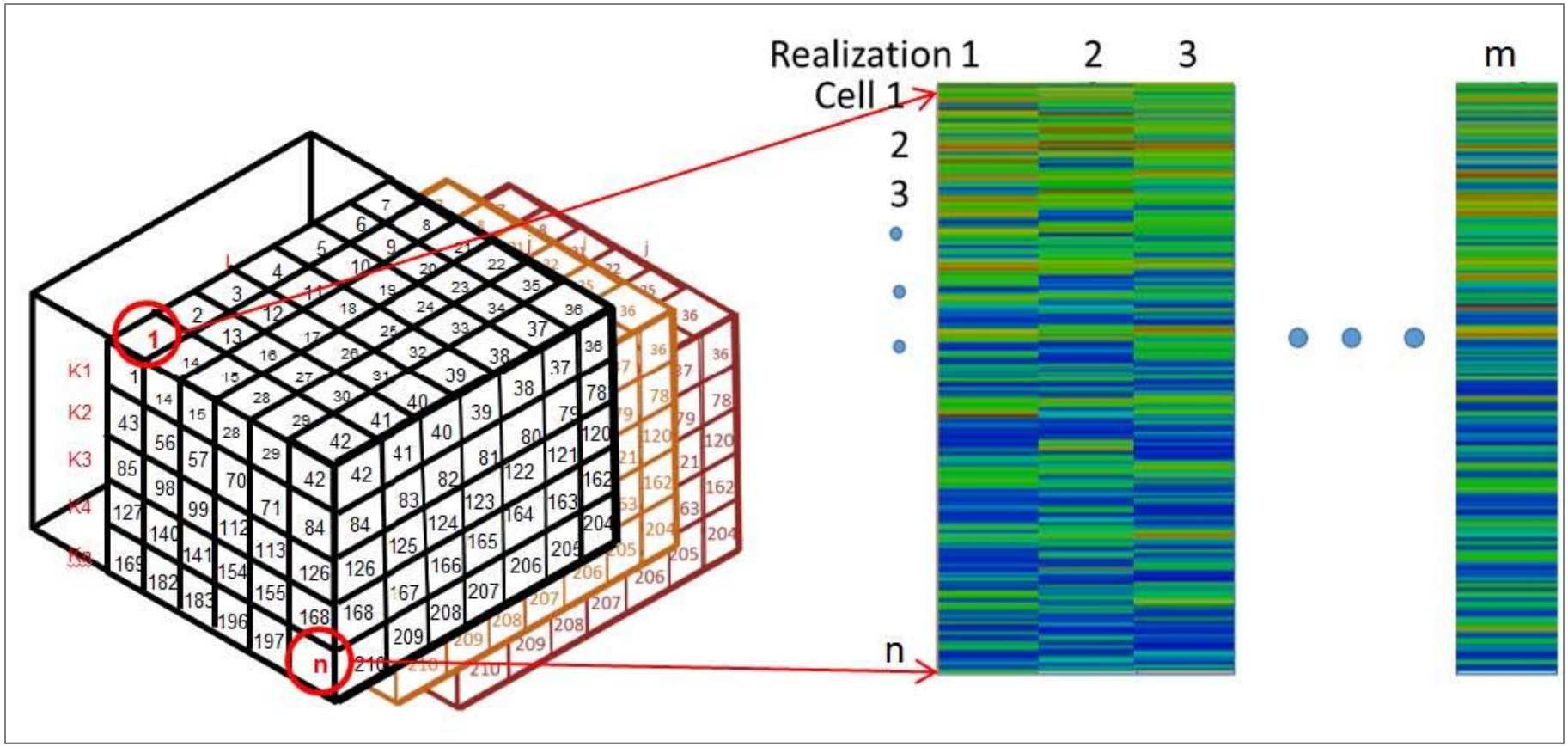

Figure 2. Organized 3-D realizations of rock properties into arrays.

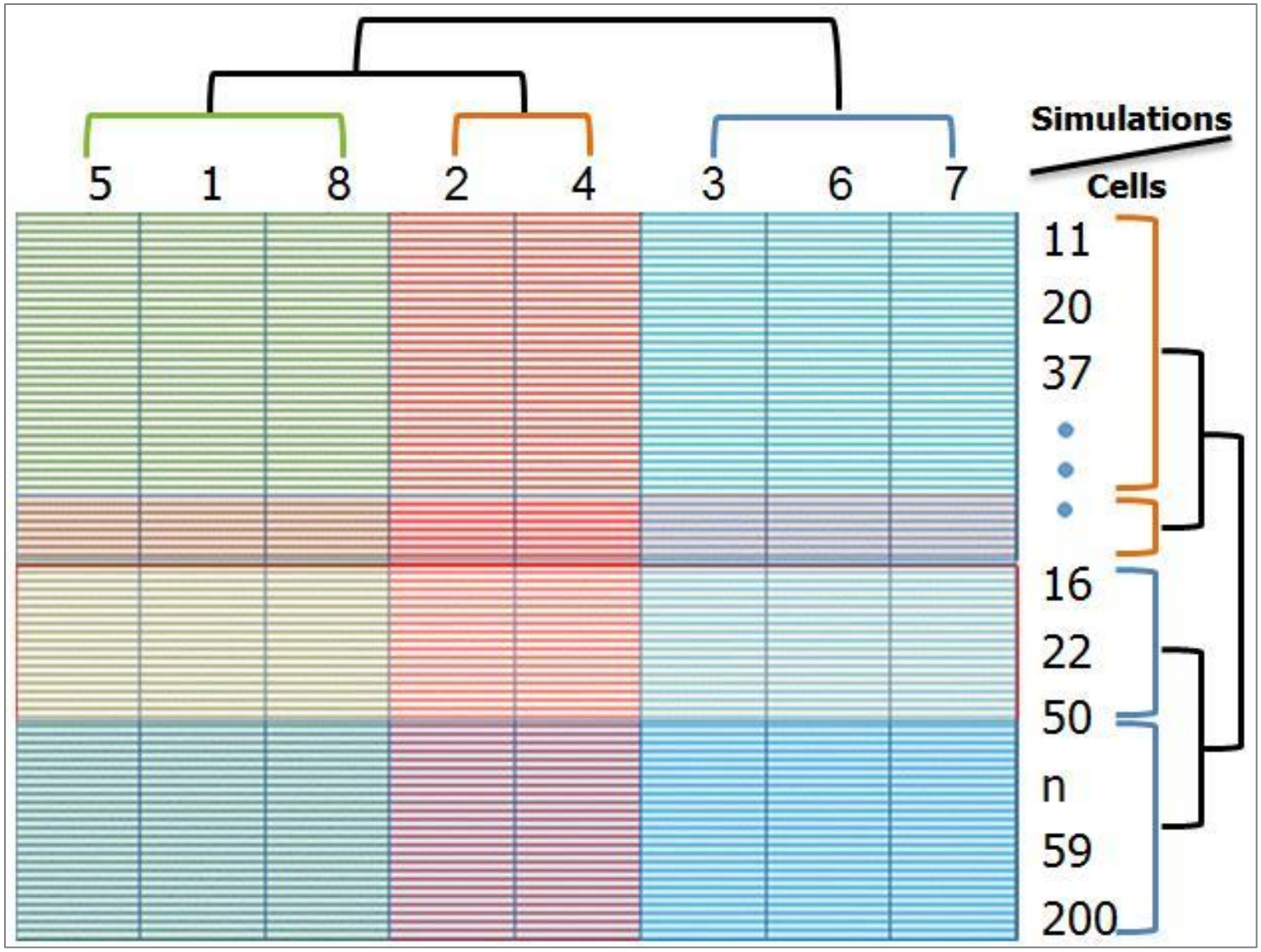

Figure 3. Two-way cluster analysis of multiple realizations.

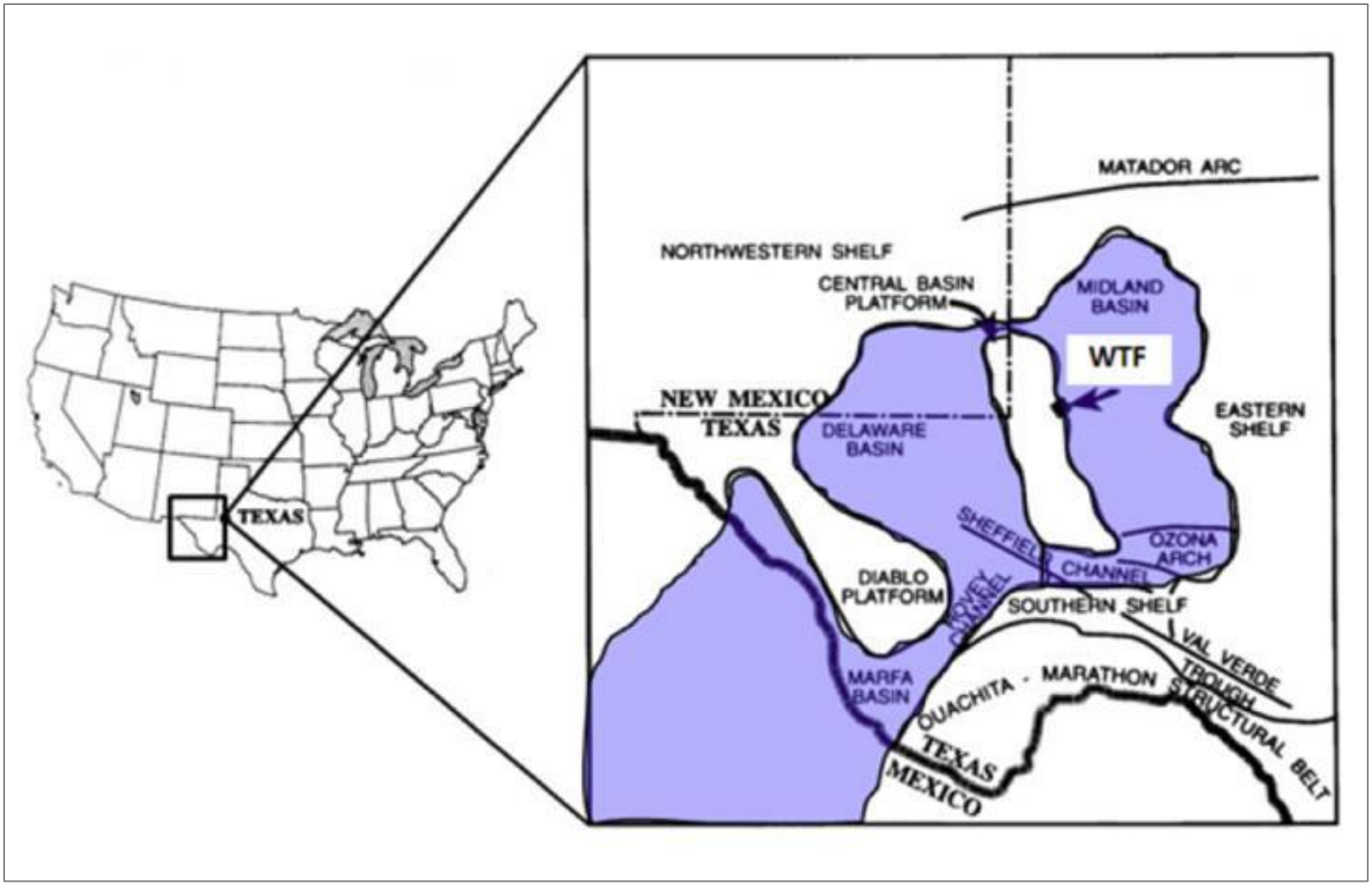

Figure 4. Study area.

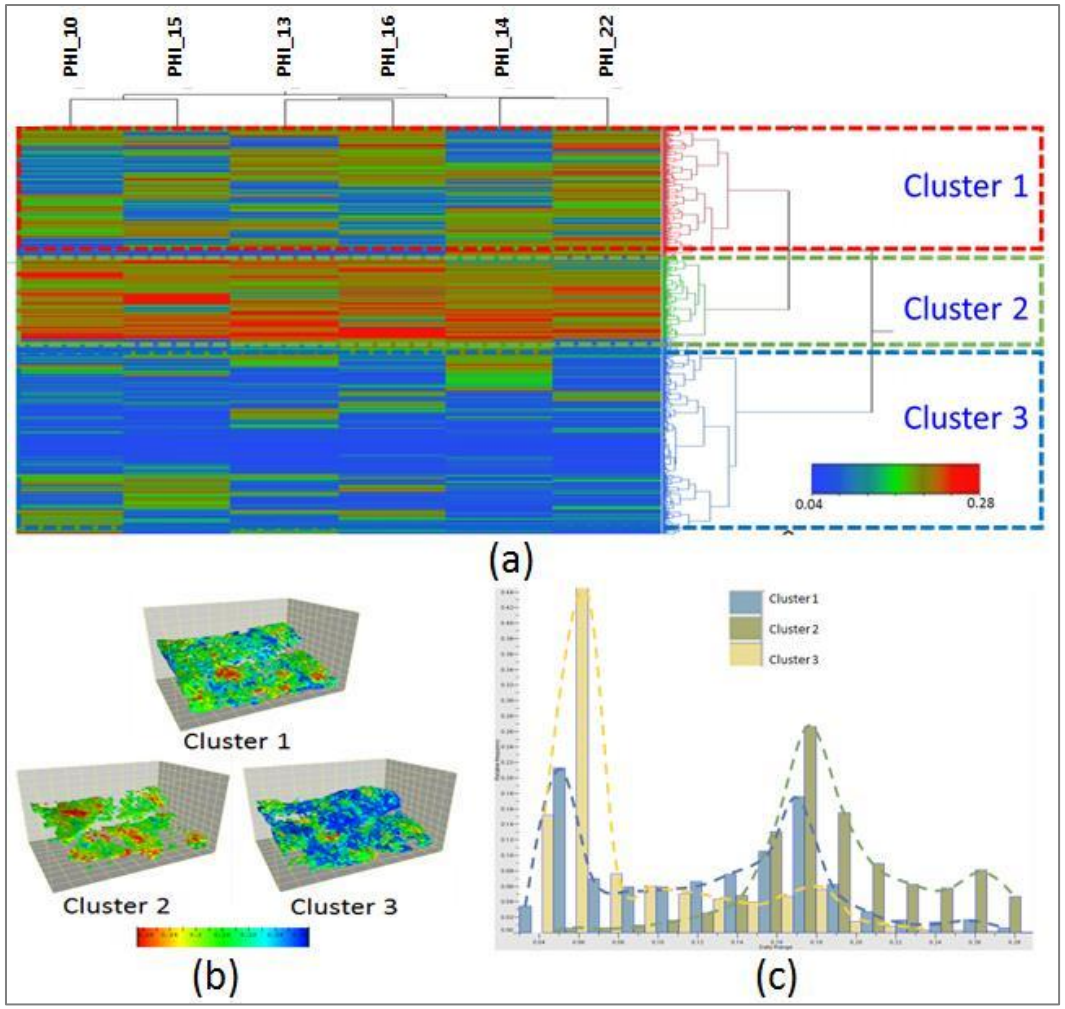

Figure 5. Results from the two-way cluster analysis. (a) Color-coded porosity map, together with both the top and right dendrograms, to show the realizations and cell clusters. (b) Clusters of cells mapped to the 3-D grid. (c) Porosity statistics of each clusters.

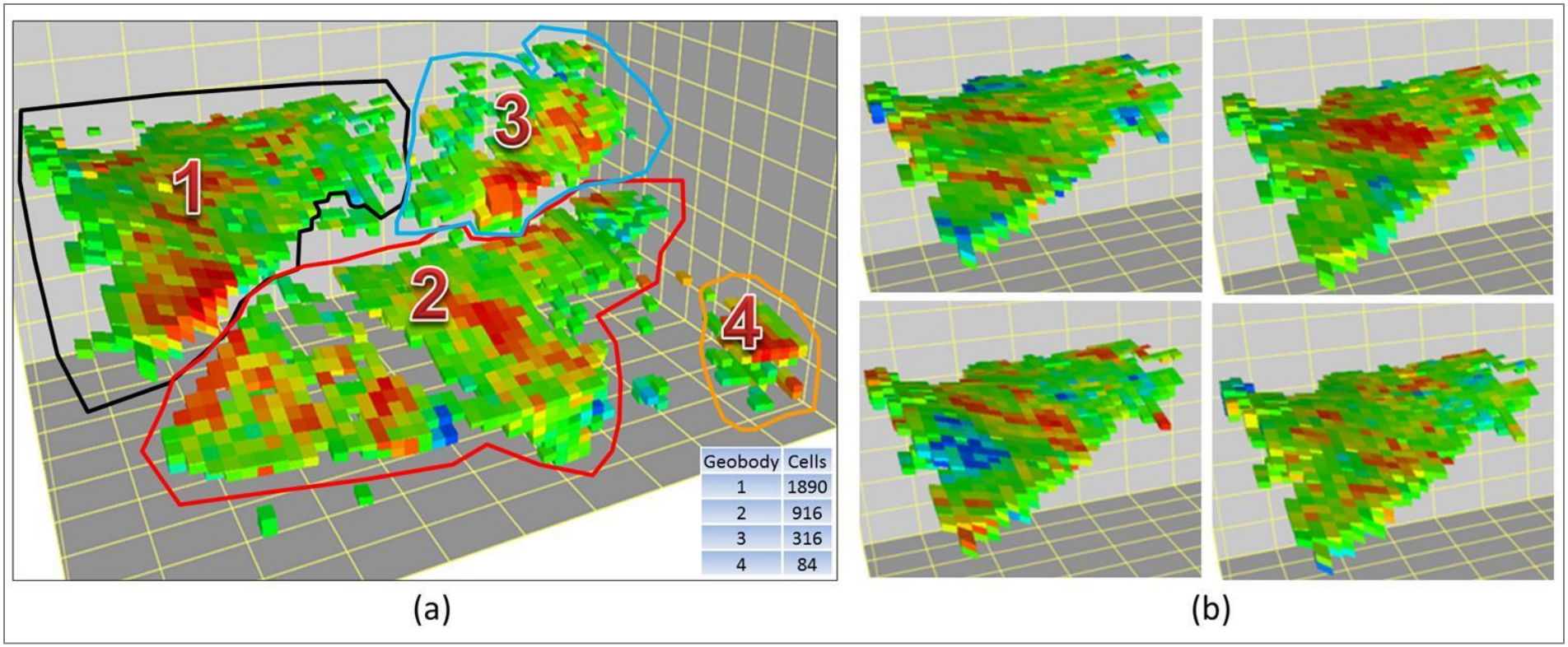

Figure 6. Connected geobodies of Cluster 2. (a) The four largest geobodies with volumes of 1,890, 916, 316, and 84 cells, respectively. (b) From top to bottom and left to right, the porosity maps of geobody #1 from realizations 13, 16, 15, and 22.

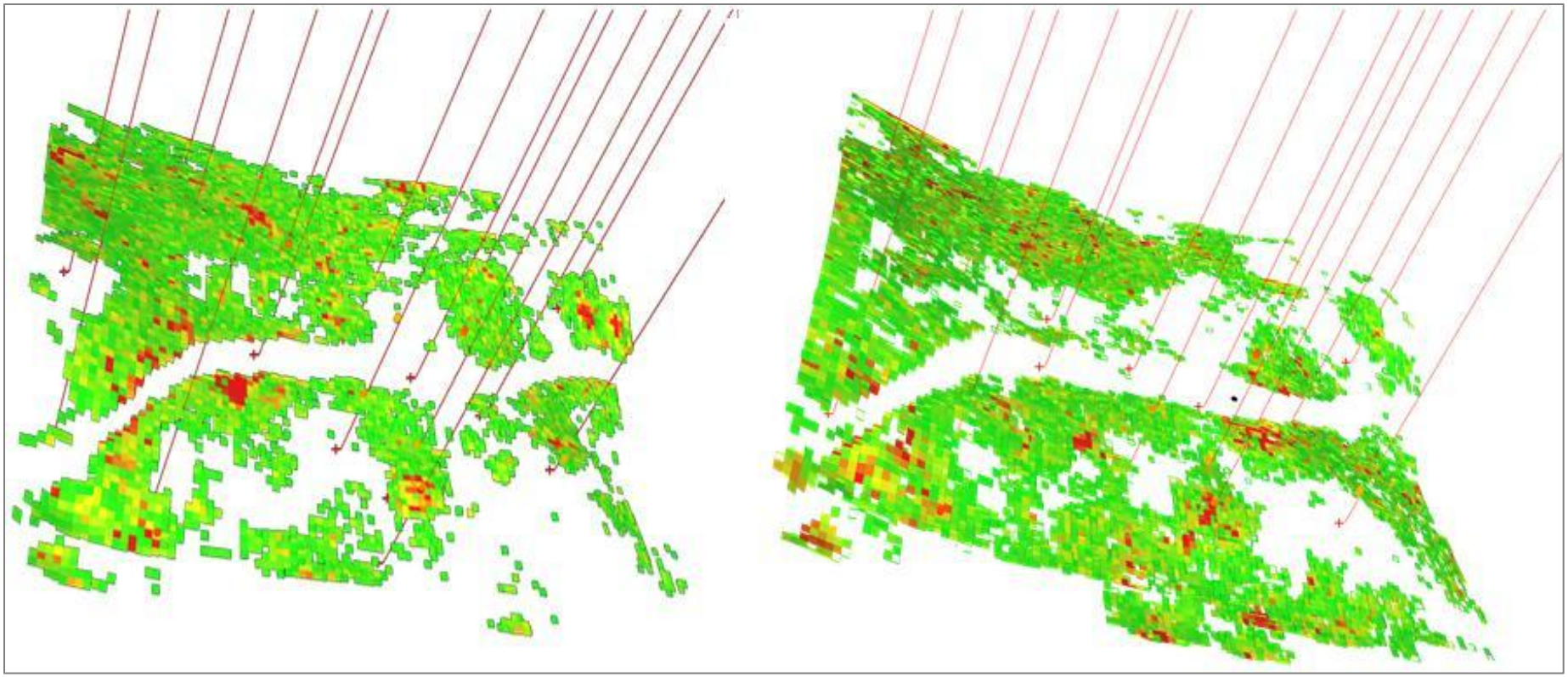

Figure 7. Geobodies of realizations 10 and 15 with porosity.

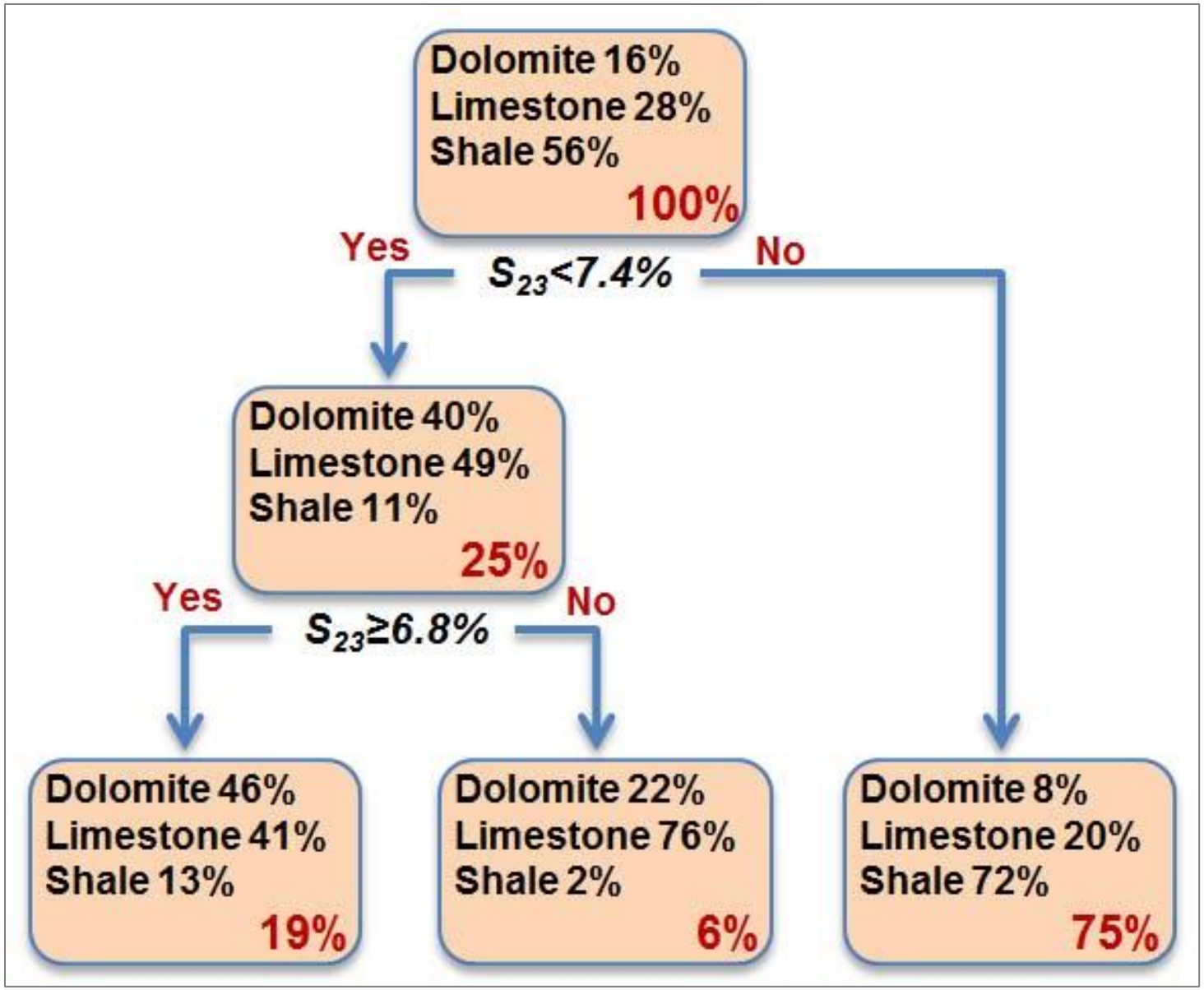

Figure 8. CART to correlate multiple realizations with lithofacies.

Realization	Mean	Std	t-test with Realization 10
10	0.2000	0.0399	
13	0.1966	0.0442	2.48
14	0.1880	0.0474	8.42
15	0.1846	0.0503	10.43
16	0.1973	0.0432	2.00
22	0.1822	0.0506	12.01

Table 1: Statistics of porosity realization on the largest geobody.