

A Framework for Scientific Discovery in Geological Databases*

Cen Li and Gautam Biswas
 Department of Computer Science
 Vanderbilt University
 Box 1679, Station B
 Nashville, TN 37235
 Tel: (615)-343-6204
 Email: cenli, biswas@vuse.vanderbilt.edu

The Problem

It is common knowledge in the oil industry that the typical cost of drilling a new offshore well is in the range of \$30-40 million, but the chance of that site being an economic success is 1 in 10. Recent advances in drilling technology and data collection methods have led to oil companies and their ancillary companies collecting large amounts of geophysical/geological data from production wells and exploration sites. This information is being organized into large company databases and the question is *can this vast amount of history from previously explored fields be systematically utilized to evaluate new plays and prospects?* A possible solution is to develop the capability for retrieving analog wells and fields for the new prospects and employing Monte Carlo methods with risk analysis techniques for computing the distributions of possible hydrocarbon volumes for these prospects. This may form the basis for more accurate and objective prospect evaluation and ranking schemes.

However, with the development of more sophisticated methods for computer-based scientific discovery[6], the primary question becomes, *can we derive more precise analytic relations between observed phenomena and parameters that directly contribute to computation of the amount of oil and gas reserves.* For oil prospects, geologists compute potential recoverable reserves using the *pore-volume equation*[1]

$$\frac{\text{Recoverable Reserves in STB}}{BRV} = \frac{N}{G} * \phi * Shc * RF * 6.29$$

where

BRV = Bulk Rock Volume in m^3

N/G = Net/Gross ratio of the reservoir rock body

*This research is supported by grants from Arco Research Labs, Plano, TX.

making up the BRV
 ϕ = average reservoir porosity(pore volume)
 Shc = average hydrocarbon saturation
 RF = Recovery Factor(the fraction of the in-place petroleum expected to be recovered to surface)
 6.29 = factor converting m^3 to barrels
 FVF = Formation Volume Factor of oil(the amount that the oil volume shrinks on moving from reservoir to surface)
 STB = Stock Tank Barrels, i.e. barrels at standard conditions of 60°F and 14.7 psia.

In qualitative terms, good recoverable reserves have high hydrocarbon saturation, are trapped by highly porous sediments(reservoir porosity), and surrounded by hard bulk rocks that prevent the hydrocarbon from leaking away. Having a large volume of porous sediments is crucial to finding good recoverable reserves, and therefore, a primary emphasis of this study is to determine the porosity values from collected data in new prospect regions. We focus on scientific discovery methods to derive empirical equations for computing porosity values in regions of interest.

Determination of the porosity or pore volume of a prospect depends upon multiple geological phenomena in a region. Some of the information, such as pore geometries, grain size, packing, and sorting, is *microscopic*, and some, such as rock types, formation, depositional setting, stratigraphic zones, and unconformities (compaction, deformation, and cementation) is *macroscopic*. These phenomena are attributed to millions of years of geophysical and geochemical evolution, and, therefore, hard to formalize and quantify. On the other hand, large amounts of geological data that directly influence hydrocarbon volume, such as porosity and permeability measurements, grain character, lithologies, formations and geometry are available from previously explored regions.

The goal of the study is to use *computer-assisted*

analysis and scientific discovery methods to derive general analytic formulations for porosity as a function of relevant geological phenomena. The general rule of thumb is that porosity decreases quasi-exponentially with depth, but a number of other factors, such as rock types, structure, and cementation confound this relationship. This necessitates the definition of proper *contexts* in which to attempt the discovery of porosity formula. The next section outlines our approach to this methodology.

The Approach

From the above description, it is clear that the empirical equations for porosity of sedimentary structures in a region are very dependent on the *context* associated with this region. The term context includes geological phenomena that govern the formation of the structures and the lithology of the region, therefore, define a set of observable and measurable geological parameters from which values of porosity can be computed. It is well known that the geological context can change drastically from basin to basin (different geographical areas in the world), and also from region to region within a basin [1, 3].

With this background, we develop a formal two-step scientific discovery process for determining empirical equations for porosity from geological data. To test our methodology, we use data from a region in the Alaskan basin. This data is labeled by code numbers (the location or wells from which they were extracted) and the stratigraphic unit numbers. The stratigraphic unit numbers are related to sediment depositional sequences that are affected by subsidence, erosion, and compaction resulting in characteristic geometries. Each data object is then described in terms of 37 geological features, such as porosity, permeability, grain size, density, and sorting, amount of different mineral fragments (e.g., quartz, chert, feldspar) present, nature of the rock fragments, pore characteristics, and cementation. All these feature-values are numeric measurements made on samples obtained from well-logs.

A more formal description of the two-step discovery process follows:

1. Context Definition

The first step in the discovery task is to identify a set of contexts $\mathbf{C} = (C_1, C_2, \dots, C_n)$. Each one of which will likely produce a unique porosity equation. Each $C_i \in \mathbf{C}$ is defined as a sequence of primitive geological structures, $C_i = g_1 \circ g_2 \circ \dots \circ g_k$ (primitive structures may appear more than once

in a sequence). The set of primitive geological structures (g_1, g_2, \dots, g_m) are extracted by a clustering process. The context definition task itself is further divided into the following subtasks:

- discovering the primitive structures (g_1, g_2, \dots, g_m) ,
- identifying relevant sequences of such primitive structures, i.e., $C_i = g_{i1} \circ g_{i2} \circ \dots \circ g_{ik}$,
- grouping the data that belong to the same sequence to form the context, and determining the relevant set of features (x_1, x_2, \dots, x_k) that will be used to derive the porosity equation for that context.

2. Equation Derivation

This step involves using statistical techniques, such as multivariate regression methods [5], to derive a form of porosity equation $\phi = f(x_1, x_2, \dots, x_k)$ for each context defined in step 1. The task is further divided into the following three subtasks:

- construct the base model from domain theory and estimate the parameters of the model using least square methods,
- for each independent variable in the model, construct and examine the *component plus residual plot* (cprp) for that variable, and transform its form, if required.
- Construct a set of dimensionless terms $\pi = (\pi_1, \pi_2, \dots, \pi_k)$ from the relevant set of features [2]. Incorporate the π_i 's into the model in a way that reduces the residual of the model.

The first step in the context definition task is to identify the set of primitive structures using a clustering methodology. In previous work [3], we have defined a three-step methodology that governs this process: (i) feature selection, (ii) clustering, and (iii) interpretation. Feature selection deals with issues for selecting object characteristics that are relevant to the study. In our experiments, this task has been primarily handled by domain experts. Clustering deals with the process of grouping the data objects based on similarities of properties among the objects. The goal is to partition the data into groups such that objects in each group are more homogeneous than objects in different groups. For our data set, in which all features are numeric-valued, we use a partitional numeric clustering program called CLUSTER [4] as the clustering tool. CLUSTER assumes each object to be a point in multidimensional space and uses the Euclidean metric as a measure of (dis)similarity between objects. Its criterion function is based on minimizing the mean

square-error within each cluster. The goal of interpretation is to determine whether the generated groups represent useful concepts in the problem solving domain. In more detail, this is often performed by looking at the intentional definition of a class, i.e., the feature-value descriptions that characterize this class, and see if they can be explained by the experts' background knowledge in the domain. For example, in our studies, experts interpreted groupings in terms of the sediment characteristics of the group. For example, if a group is characterized by clay and siderite features having high values, the expert considers this relevant because it indicates a low porosity region.

Often the expert has to iterate through different feature subsets, or express feature definitions in a more refined manner to obtain meaningful and acceptable groupings. In these studies, the experts did this by running clustering studies that define the data from four viewpoints: (i) Depositional setting, (ii) Reservoir quality, (iii) Provenance, and (iv) Stratigraphic zones. Each viewpoint entailed a different subset of features. A brute force clustering run with all features provided a gross estimate of sediment characteristics.

A number of graphical and statistical tools have been developed to facilitate the expert's comparison tasks. For example, as part of the clustering and interpretation package, we have developed software that allow users to cross-tabulate different clustering runs to study the similarities and differences in the groupings. Besides, a number of graphical tools have been created to allow the expert to compare feature-value definitions across various groups.

The net result of this process provides the primitive set (g_1, g_2, \dots, g_m) of step 1 of the discovery task. These primitives are then mapped onto the unit code versus stratigraphic unit map. Fig. 1 depicts a partial mapping for a set of wells and four primitive structures. In the actual experiments, our experts initially identified about 8-10 primitive structures, but further experiments are being conducted to validate these results.

The next step in the discovery process is to identify sections of wells and regions that are made up of the same sequence of geological primitives. Every such sequence defines a context C_i . Some criterion employed in identifying sequences are that longer sequences are more useful than shorter ones, and sequences that occur more frequently are likely to define better contexts than those that occur infrequently. Currently, this sequence selection job is done by hand, but in future work, we wish to employ tools, such as mechanisms for learning context-free grammars from string sequences to assist experts in generating useful sequences. A reason for considering sequences that occur more fre-

quently is that they will produce more generally applicable porosity equations than ones from infrequent sequences.

After the contexts are defined, data points belonging to each context can be grouped to derive useful formulae. From the partial mapping of Fig. 1, the context $C_1 = g_2 \circ g_1 \circ g_2 \circ g_3$ was identified in two well regions (the 300 and 600 series).

Step 2 of the discovery process uses equation discovery techniques to derive porosity equations for each context. Theoretically, the possible functional relationships that may exist among any set of variables are infinite. It would be computationally intractable to derive models given a data set without constraining the search for possible functional relations. One way to cut down on the search is to reduce the number of independent variables involved. Step 1 achieves because the cluster derivation process also identifies the essential and relevant features that define each class. A second effective method for reducing the search space is to use domain knowledge to define approximate functional relations between the dependent variable and each of the independent variables. We exploit this and assume a basic model suggested by domain theory is provided to the system to start the equation discovery process. Parameter estimation in the basic model is done using a standard least squares method from the Minpack¹ statistical package.

Our application requires that we be able to derive linear and nonlinear relationships between the goal variable and the set of independent variables, not being bound to just the initial model suggested by domain theory. The discovery process should be capable of dynamically adjusting model parameters to better fit the data. Once the basic equation model is established, the model fit is improved by applying transformations using a graphical method, *component plus residual plots* (cprp)[5].

Note that domain theory suggests individual relations between independent variables and the dependent one. For example, given that $y = f(x_1, x_2, x_3)$, domain theory may indicate that, x_1 is linearly related, x_2 is quadratically related, and x_3 is inverse quadratically related to the dependent variable y . Our methodology starts off with an equation form, say $y = c_0 + c_1x_1 + c_2x_2^2 + c_3x_3^{-2}$, estimates of the coefficients of this model using the least squares method. Depending on the error (residual) term, the equation is dynamically adjusted to obtain a better fit. This is described in some detail next.

The first step in the cprp method is to convert a

¹This is a free software package developed by Burton S. Garbow, Kenneth E. Hillstom, Jorge J. Moore at Argonne National Laboratories, IL.

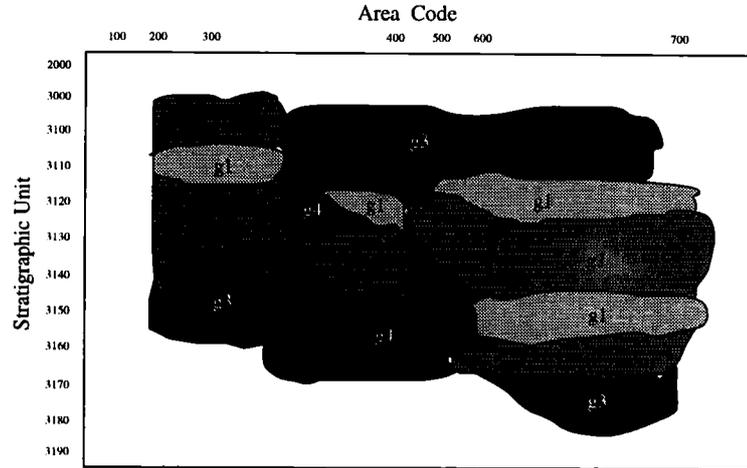


Figure 1: Area Code versus Stratigraphic Unit Map for Part of the Studied Region

given nonlinear equation into a linear form. In this case, the above equation $y = c_0 + c_1x_1 + c_2x_2 + c_3x_3^{-2}$ would be transformed into

$$y_i = c_0 + c_1x_{i1} + c_2x_{i2} + c_3x_{i3} + e_i,$$

where $x_{i1} = x_1$, $x_{i2} = x_2^2$, and $x_{i3} = x_3^{-2}$, and e_i is the residual. The *component plus residual* for an independent variable, x_{im} , is defined as

$$c_m x_{im} + e_i = y_i - c_0 - \sum_{j=1:j \neq m}^k c_j x_{ij},$$

since $c_m x_{im}$ can be viewed as a component of \hat{y}_i , the predicted values of the goal variable. Here, $c_m x_{im} + e_i$ is essentially y_i with the linear effects of the other variables removed. The plots of $c_m x_{im} + e_i$ against x_{im} are called *component plus residual plots* (Fig. 2).

The cprp for an independent variable x_{im} determines whether a transformation needs to be applied to that variable. The plot is analyzed in the following manner. First, the set of points in the plot is partitioned into three groups along the x_{im} value, such that each group has approximately the same number of points ($k \simeq n/3$). The most "representative" point of each group is calculated as $(\frac{\sum_k x_{im}}{k}, \frac{\sum_k (c_m x_{im} + e_i)}{k})$. Next, the slopes, k_{12} , for the line joining the first two points and k_{13} for the line joining the first and the last point is calculated. Compare the two slopes: if $k_{12} = k_{13}$, the data points should be described as a straight line which implies that no transformation is needed; if $k_{12} < k_{13}$, the line is convex, otherwise, the line is concave (see Fig. 2). In either case, the goal variable, y , or the independent variable, x_{im} , need to be transformed using the *ladder* of power transformations shown in Fig. 3. The idea is to move up the

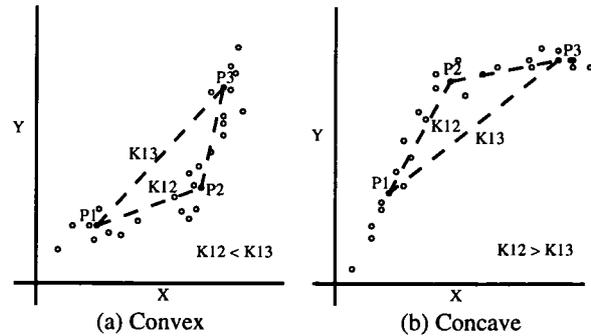


Figure 2: Two Configurations

ladder if the three points are in a convex configuration, and move down the ladder when they are in a concave configuration. Coefficients are again derived for the new form of the equation, and if the residuals decrease, this new form is accepted and the cprp process is repeated. Otherwise, the original form of the equation is retained. This cycle continues till the slopes become equal or the line changes from convex to concave, or *vice versa*.

As discussed earlier, we have applied this method to the Alaskan data set which contains about 2600 objects corresponding to wells, and each object is described in terms of 37 geological features. Clustering this dataset produced a seven group structure, from which group 7 was picked for further preliminary analysis. Characteristic features of this group indicate that it is a low porosity group, and our domain experts picked 5 variables to establish a relationship to porosity, the goal variable. We were also told that two variables, macroporosity (M) and siderite (S) are linearly related to porosity, and the other three, clay

matrix(C), laminations(L) and glauconite(G) have an inverse square relation to porosity. With this knowledge, the initial base model was set up as:

$$P(\text{porosity}) = c_0 + c_1M + c_2S + \frac{c_3}{c_4C^2 + c_5L^2 + c_6G^2}$$

where, c_0, \dots, c_6 are the parameters to be estimated by regression analysis. After the parameters are estimated, the model is derived as:

$$P = 9.719 + 0.43M + 0.033S + \frac{2.3 \cdot 10^8}{-3.44 \cdot 10^6 C^2 - 4.52 \cdot 10^5 L^2 + 6.5 \cdot 10^6 G^2}$$

To study the methodology further, the cprp for variable S suggested that it be transformed to a higher order term. Therefore, S , was replaced by S^2 in the model and the coefficients were rederived.

$$P = 9.8 + 0.468M - 0.004S^2 + \frac{1.2 \cdot 10^7}{-4.7 \cdot 10^6 C^2 - 7.8 \cdot 10^5 L^2 + 2.2 \cdot 10^7 G^2}$$

The residual for the new model, 20.47, was smaller than that of the original model (21.52). This illustrates how the transformation process can be systematically employed to improve the formula.

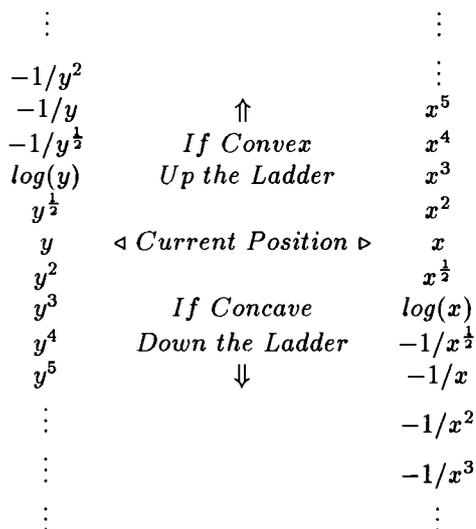


Figure 3: Ladder of Power Transformations

Note that the current method applies to transformations carried out one at a time on variables, and may not apply in situations where terms are multiplicative or involved in exponential relations, e.g., $\frac{x_1^{-1/2} x_2}{x_3^3}$. In such situations, one of two things can be done: (i) use logarithm forms to transform multiplicative terms to additive ones, and (ii) derive appropriate π terms (identified in step 1) to replace existing model components with the proper π terms that better fit the equation model to the data.

Summary

Our work on scientific discovery extends previous work on equation generation from data[6]. Clustering methodologies and a suite of graphical and statistical tools are used to define empirical contexts in which porosity equations can be generated. In our work to date, we have put together a set of techniques that address individual steps in our discovery algorithm. We have also demonstrated that they produce interesting and useful results.

Currently, we are working on refining and making more systematic context generation techniques, and are coupling regression analysis methods with the heuristic model refinement techniques for equation generation. Encouraging results have been obtained. This work shows how unsupervised clustering techniques can be combined with equation finding methods to derive empirically the analytical models in domains where strong theories do not exist.

Acknowledgements

The authors wish to thank Dr. Jim Hickey and Dr. Ron Day of Arco for their help as geological experts in this project.

References

- [1] P.A. Allen and J.R. Allen. Basin Analysis: Principles & Applications. Blackwell Scientific Publications, 1990.
- [2] R. Bhaskar and A. Nigam. Qualitative Physics using Dimensional Analysis. *Artificial Intelligence*, vol. 45, pp. 73-111, 1990.
- [3] G. Biswas, J. Weinberg, and C. Li. ITERATE: A Conceptual Clustering Method for Knowledge Discovery in Databases. *Innovative Applications of Artificial Intelligence in the Oil and Gas Industry*, B.Braunschweig and R. Day, Editors. To appear, Editions Technip, 1995.
- [4] A.K. Jain and R.C. Dubes. "Algorithms for clustering data," Prentice Hall, Englewood Cliffs, 1988.
- [5] A. Sen and M. Srivastava. Regression Analysis. Springer-Verlag Publications, 1990.
- [6] J.M. Zytkow and R. Zembowicz. Database Exploration in Search of Regularities. *Journal of Intelligent Information Systems*, 2:39-81, 1993.