Knowledge-based Scientific Discovery in Geological Databases *

Cen Li and **Gautam Biswas**
Department of Computer Science, Vanderbilt University
Box 1679, Station B, Nashville, TN 37235
Email: biswas, cenli@vuse.vanderbilt.edu

Abstract

A framework for knowledge-based scientific discovery in geological databases has been developed. The discovery process consists of two main steps: *context definition* and *equation derivation*. Context definition properly defines and formulates homogeneous regions, each of which is likely to produce a unique and meaningful analytic formula for the goal variable. Clustering techniques and a suite of visualization and interpretation routines make up a tool box that assists the context definition task. Within each context, multi-variable regression analysis is conducted to derive analytic equations between the goal variable and a set of relevant independent variables, starting with one or more of the initial base models. Domain knowledge, plus a heuristic search technique called *component plus residual plots* dynamically guide the equation refinement process. The methodology has been applied to derive porosity equations for data collected from oil fields in the Alaska Basin. Prelimina.y results demonstrate the effectiveness of this methodology.

Keywords:
knowledge discovery from databases, scientific discovery, clustering, regression analysis, component plus residual plots

1 Introduction

Like a number of other domains, database mining is becoming crucial in oil exploration and production. 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 ancillaries collecting large amounts of geophysical/geological data from production wells and exploration sites, and then organizing them into large databases. *Can this vast amount of history from previously explored fields be systematically utilized to evaluate new plays and prospects ?* As a first step, the oil industry has developed methodologies for finding fields and wells that are, in some sense, similar to a new prospect, and used the information in an ad hoc way to rank a set of new prospects[Allen and Allen, 1990].

Recent developments in database mining[Fayyad and Uthurusamy, 1994] and the advances in computer-based scientific discovery[Zytkow and Zembowicz, 1993] naturally lead to the following question *"can we derive more precise analytic relations between observed phenomena and parameters to make better quantitative estimates of oil and gas reserves ?"* 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. A large volume of porous sediments is crucial to finding good recoverable reserves, therefore, a primary task in determining hydrocarbon potential is to develop reliable and accurate methods for estimation of sediment porosities from the collected data.

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.

This paper develops a knowledge-based scientific discovery approach to derive analytic formulae for porosity as a function of relevant geological phenomena. The general rule of thumb is that porosity decreases quasi-exponentially with depth:

$$Porosity = K \cdot e^{-F(x_1, x_2, \ldots, x_m) \cdot Depth}. \quad (1)$$

But a number of other factors, such as rock types, structure, and cementation, appearing as the parameters of function F in equation 1, confound this relationship. This necessitates the definition of proper *contexts* in

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

which to attempt discovery of porosity formulae. In data analysis we have been conducting for geological experts, the feature "depth" is intentionally removed, so that other geological characteristics that affect hydrocarbon potential can be studied in some detail. Our goal is to derive the subset $x_1, x_2, ..., x_m$ from a larger set of geological features, and the functional relationship F that best defines the porosity function in a region.

Real exploration data collected from a region in the Alaska basin is analyzed using the methodology developed. The data is labeled by code numbers (the location or wells from which they were extracted) and stratigraphic unit numbers. Stratigraphic unit numbers describe sediment depositional sequences. These sequences are affected by subsidence, erosion, and compaction which mold them into characteristic geometries. The data is extracted from the database as a flat file of *objects*; each object is 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 during exploratory drilling processes.

Note that this is real data collected during real operations, therefore, we have almost no control on the nature and organization of data. By this we mean that there is no way in which variable values can be made to go up/down in fixed increments, and it is not possible to hold values of certain parameters constant, while others are varied. Techniques used in systems like BACON[Langley *et al.*, 1983; Langley and Zytkow, 1989], FAHRENHEIT[Langley and Zytkow, 1989], and ABACUS[Greene, 1988] cannot be directly applied to organize the search for relations among groups of variables in this system. On the other hand, we have access to human experts who have partial knowledge about the relations between parameters. Therefore, the methodology we have developed is tailored to exploit this partial knowledge and focus the search of a systematic discovery method to derive analytic relations for porosity in terms of observed geological parameters.

The framework for our knowledge-based discovery scheme is illustrated in Fig. 1. The first step in the discovery process, retrieval and preprocessing of data is not discussed in this paper. The next two steps: (i) clustering of the data into groups and interpreting the meaning of the clusters generated to define contexts, and (ii) equation discovery in each of these contexts, are the primary topics discussed in this paper. The next section reviews current work in this area.

2 Background

As discussed, early discovery systems like BACON[Langley and Zytkow, 1989], FAHRENHEIT[Langley and Zytkow, 1989], IDS[Langley and Zytkow, 1989], and COPER[Kokar, 1998], were designed to work in highly repeatable domains with well defined physical laws. For example, BACON assumed that the data required to derive equations could be acquired sequentially (say, by performing experiments) so that relations between pairs of variables could be examined while holding the other variables constant. Using systematic search processes, introducing ratio and product terms, and considering intrinsic properties, BACON correctly formulated equations involving varying degrees of polynomials. FAHRENHEIT augmented BACON's abilities to find and associate with each derived equation upper and lower boundary values. IDS employed Qualitative Process Theory (QPT)-like[Forbus, 1984] qualitative schema to embed numeric equations in a qualitative framework and thus constrain the search space. The qualitative framework makes it easier to understand the laws in context. Using dimension analysis[Bhaskar and Nigam, 1990], COPER eliminated irrelevant arguments and generated additional relevant argument descriptors in deriving functional formulae.

More recent systems like 49er[Zytkow and Zembowicz, 1993] and KEDS[Rao and Lu, 1992], are designed to work with real world data which, in addition to being fuzzy and noisy, is frequently associated with more than one context. It becomes important for the system to group data by context before attempting equation discovery. In such situations, discovery is characterized by a two-step process: preliminary search for contexts, followed by equation generation and model refinement.

In the preliminary search step, 49er organizes the data into contingency tables and performs slicing and projection operations to get the data into a form where strong regularity or functional relations can be detected. Subsets of data defined by slicing and projection define individual context, and the equation discovery process is applied separately to each context. In the KEDS system, the preliminary search step involves partitioning by model matching. The expected relations are expressed as polynomial equation templates. The search process uses sets of data points to compute the coefficient values of chosen templates, and then determine the probability of fit for each data point to the equation. This is used to define contiguous homogeneous regions, and each region forms a context in the domain of interest.

In the model refinement step, 49er invokes "Equation Finder"[Zembowicz and Zytkow, 1992] to uncover analytic relations between the goal variable and the control variable. Additional relations between the two variables can be explored by considering transformations like $log(x)$ and $\frac{1}{x}$, iteratively for the goal and control variables, enabling the derivation of complex, non linear forms. For the KEDS system, the matched equation templates are further refined by multi-variable regression analysis within each context to accurately estimate the polynomial coefficients.

A preliminary analysis of geological processes makes it clear that the empirical equations for porosity of sedimentary structures in a region are very dependent on the *context*, which can be expressed in terms of geological phenomena, such as geometry, lithology, compaction, and subsidence, associated with a region. It is also well known that the geological context changes from basin to basin(different geographical areas in the world) and also from region to region within a basin[Allen and Allen, 1990; Biswas *et al.*, 1995]. Furthermore, the underlying features of contexts may vary greatly. Simple model matching techniques, which work

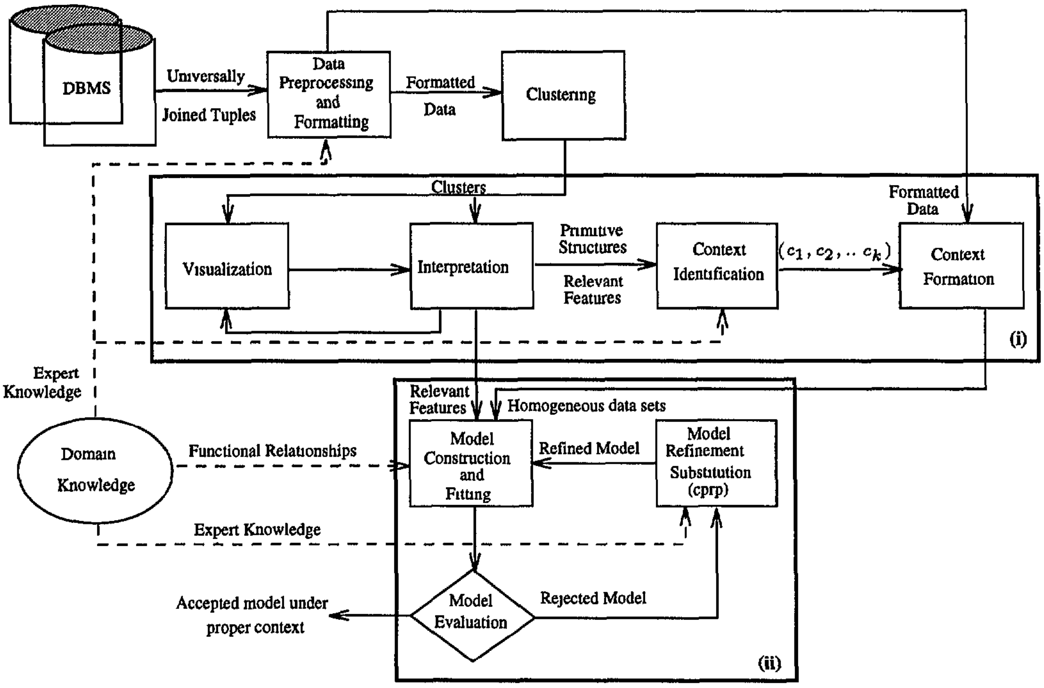


Figure 1: **Knowledge-based Scientific Discovery System Architecture**

in engineering domains where behavior is constrained by man-made systems and well-established laws of physics, may not apply in the hydrocarbon exploration domain. To address this, we use an unsupervised numeric clustering scheme, like the ABACUS system[Greene, 1988], to derive gross structural properties of the data, and map them onto relevant contexts for equation discovery.

3 Our Approach

Our approach to scientific discovery adapts the two-step methodology described in figure 2. It is assumed that each context is best defined by a unique porosity equation.

3.1 Context Definition

The context definition step identifies a set of contexts $\mathbf{C} = (C_1, C_2, ..., C_n)$, where each C_i is defined as a sequence of primitive geological structures. Primitive structures are identified using unsupervised clustering techniques. In previous work[Biswas *et al.*, 1995], the clustering task is defined as a three-step methodology: (i) feature selection, (ii) clustering, and (iii) interpretation. Feature selection deals with selection of object characteristics that are relevant to the study being conducted. In our experiments, this task has been primarily handled by domain experts, assisted by our visualization and interpretation tools.

The goal of clustering is to partition the data into groups such that objects in each group are more similar to each other than objects in different groups. In our data set, all feature-values are numeric, so we use a standard numeric partitional clustering program called

1. **Context Definition**
   - 1.1 discover *primitive structures* $(g_1, g_2, ..., g_m)$ by clustering,
   - 1.2 define *context*in terms of the relevant sequences of primitive structures, i.e., $C_i = g_{i1} \circ g_{i2} \circ, ..., \circ g_{ik}$,
   - 1.3 group data according to the context definition to form *homogeneous data groups*,
   - 1.4 for each relevant data group, determine the set of *relevant variables* $(x_1, x_2, ..., x_k)$ for porosity.
2. **Equation Derivation**
   - 2.1 select possible *base models* using domain theory,
   - 2.2 use the *least squares* method to generate coefficient values for each base model,
   - 2.3 use the *component plus residual plot* (cprp) heuristic to dynamically modify the equation model to better fit the data,
   - 2.4 construct a set of dimensionless terms $\pi = (\pi_1, \pi_2, ..., \pi_k)$ from the relevant set of features [Bhaskar and Nigam, 1990].

Figure 2: **Description of the Knowledge-based Scientific Discovery Process**

CLUSTER[Jain and Dubes, 1988] as the clustering tool. CLUSTER assumes each object to be a point in a multidimensional metric space, and uses the Euclidean distance 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 domain background knowledge (or by domain experts). For example, in these studies, our experts focused on the sediment characteristics to assign meaning to groups, a group characterized by high clay and siderite content but low in quartz was considered relevant and was consistent with a low porosity region. Experts often iterated through different feature subsets and changed feature descriptions to obtain meaningful and acceptable groupings.

A number of graphical and statistical tools have been developed to facilitate the interpretation task. For example, utilities help to cross-tabulate different clustering runs to study the similarities and differences between the groupings formed. Statistical tools identify feature value peaks in individual classes to help identify relevant features. Graphical plot routines also assist in performing this task visually.

The net result of this process is the identification of a set of homogeneous primitive geological structures $(g_1, g_2, ..., g_m)$. These primitives are then mapped onto the unit code versus stratigraphic unit map. Fig. 3 depicts a partial mapping for a set of wells and four primitive structures.

The next step in the discovery process identifies sections of wells 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: longer sequences are more useful than shorter ones, and sequences that occur more frequently are more useful than those that occur infrequently. Currently, the sequence selection job is done by hand, but in future work, tools, such as mechanisms for learning context-free grammars from string sequences, will be employed to assist experts in generating useful sequences. The reason for considering more frequently occurring sequences is that they are more likely to produce generally applicable porosity equations. From the partial mapping of Fig. 3, 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). After the contexts are defined, data points belonging to each context are grouped together to initiate equation derivation.

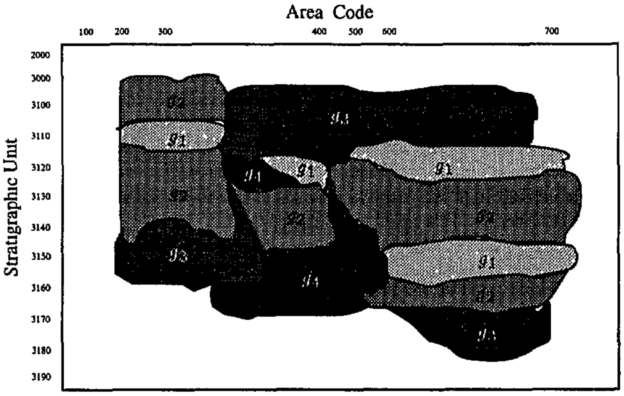


Figure 3: **Area Code versus Stratigraphic Unit Map for Part of the Studied Region**

3.2 Equation Derivation

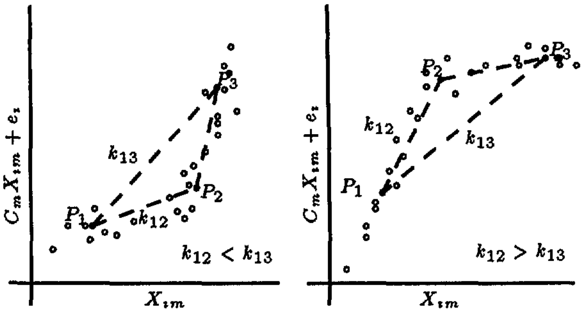
The methodology used for deriving equations that describe the goal variable as a function of the relevant independent variables, i.e., $y = f(x_1, x_2, ...x_k)$, is multivariable regression analysis[Sen and Srivastava, 1990]. Theoretically, the number of possible functional relationships that may exist among any set of variables are infinite. It would be computationally intractable to derive models for a given data set without constraining this search. Section 2 discussed why the simplistic search methods used in systems like BACON and ABACUS cannot be applied in this situation.

In general, the search space can be cut down by reducing the number of independent variables in the equation discovery process. This is achieved in the previous step by recording the relevant features associated with each class that make up a context. Further narrowing of the search space can be achieved by employing domain knowledge to select the approximate functional forms. This idea is exploited and it is assumed that pairwise functional relationships between the goal variable and each of the relevant independent variables can be derived from domain theory, or they are provided by the domain expert interactively. (Note that systems like BACON and 49er assume this can be derived). 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. One of the possible base models that the system then creates is the model $y = c_0 + c_1x_1 + c_2x_2^2 + c_3x_3^{-2}$. An alternate base model may be $y = c_0 + \frac{c_1x_1x_2^2}{c_2x_3^2+c_3}$. The standard least squares routine from the Minpack[1] statistical package is employed to derive equation coefficients.

The obvious next step is to evaluate the base models in terms of fit, and refine them to obtain better fitting models. This may require changing the equation form and dynamically adjusting model parameters to better fit the data. A heuristic method, the *component plus residual plots* [Sen and Srivastava, 1990], is used to analyze the error (residual) term in the manner described below.

First, convert a given nonlinear equation into a linear form. For example, the above base model 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 independent

[1] This is a free software package developed by B.S. Garbow, K.E. Hillstrom, J.J. Moore at Argonne National Labs.



(a) Convex (b) Concave

Figure 4: **Two Configurations**

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} is the *component plus residual plot*(cprp) (Fig. 4).

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.

1. If $k_{12} = k_{13}$, the data points describe a straight line and no transformation is needed.
2. If $k_{12} < k_{13}$, the line is convex, otherwise, the line is concave(see Fig. 4). In either case, the goal variable, y, or the independent variable, x_{im}, needs to be transformed using the *ladder of power transformations* shown in Fig. 5. The idea is to move up the 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.*

4 Experiments and Results

As discussed earlier, this method was applied to a data set of about 2600 objects corresponding to sample measurements collected from wells is the Alaskan Basin. Clustering this data set produced a seven group structure, and after the interpretation and context definition step, a set of 138 objects representing a context was picked for further analysis. The experts surmised that this context represents a low porosity region, and after studying the feature value ranges, picked variables to establish a relationship to porosity(P), the goal variable. Further the system was 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 non-linear relation to porosity. With this knowledge, three initial base models were set up as:

⋮		⋮
$-1/y^2$		⋮
$-1/y$	⇑	x^5
$-1/y^{\frac{1}{2}}$	*If Convex*	x^4
$ln(y)$	*Up the Ladder*	x^3
$y^{\frac{1}{2}}$		x^2
y	◁ *Current Position* ▷	x
y^2		$x^{\frac{1}{2}}$
y^3	*If Concave*	$ln(x)$
y^4	*Down the Ladder*	$-1/x^{\frac{1}{2}}$
y^5	⇓	$-1/x$
⋮		$-1/x^2$
⋮		$-1/x^3$
⋮		⋮

Figure 5: **Ladder of Power Transformations**

Model 1:$P = c_0 + c_1 M + c_2 S + \frac{c_3}{c_4 C^2 + c_5 L^2 + c_6 G^2}$
Model 2:$P = c_0 + c_1 M + c_2 S + \frac{c_3}{c_4 C^2 + c_5} + \frac{c_6}{c_7 L^2 + c_8} + \frac{c_9}{c_{10} G + c_{11}}$
Model 3:$P = c_0 + \frac{c_1 M S}{c_2 C^2 L^2 G^2 + c_3}$

where the $c_i's$ are the parameters to be estimated by multi-variable regression analysis. After the parameters are estimated, the equations listed below were derived:

$P = 9.719 + 0.43M + 0.033S + \frac{2.3*10^8}{-3.44*10^6 C^2 - 4\ 52*10^5 L^2 + 6.5*10^6 G^2}$

$P = 11.2 + 0.44M - 0.06S + \frac{1.2*10^3}{7.0*10^2 C^2 + 5\ 8*10^2} + \frac{7\ 23*10^5}{1.9*10^3 L^2 + 2.49*10^5} - \frac{7.5*10^2}{52G + 1\ 84*10^2}$

$P = 10.0 + \frac{1.7*10^5 MS - 7.5*10^3}{24.0*C^2 L^2 G^2 + 5.8*10^5}$

The Euclidean norm(Enorm) of the residuals for the three equations were 21.52, 16.06 and 23.97, respectively, indicating that model 2 was the best fit model. However, the high Enorms implied a poor fit, suggesting a change in the form of the dependent variable, using the left side of the ladder in Fig. 5. Just to be sure, however, a simpler transformation, consistent with the cprp process was tried for model 1: transform the form of variable S from linear to quadratic. This only brought the Enorm of the residuals down slightly from 21.52 to 20.47.

The cprp plots suggested moving up the ladder, so y was successively transformed to $y^{1/2}$ and then $ln(y)$. For the second transformation, the following equations were obtained:

$lnP = 2.26 + 0.037M - 0.0012S + \frac{2\ 7*10^5}{-1.7*10^6 C^2 - 2.9*10^5 L^2 + 3.7*10^6 G^2}$

	group 1	group 2	group 3	group 4
Model 2	16.06	20.07	56.69	45.065
Model 2'	1.59	1.61	5.2	3.45

Table 1: **The Enorms of Four Groups of Objects fitted by Model 2 and Model 2'**

$$lnP = 2.4 + 0.038M - 0.009S + \frac{1.46*10^2}{9.2*10^2C^2+8.8*10^2} + \frac{1.47*10^6}{3.9*10^4L^2+3.8*10^6} - \frac{4.95*10^2}{3.4*10^2G+1.1*10^3}$$

$$lnP = 2.3 + \frac{9.0*10^3SM-6.8*10^3}{-7.3C^2L^2G^2+1.5*10^6}$$

with Enorm of the residuals 2.20, 1.589, and 2.397, which is considerably less than the early residuals.

Model 2' was picked for further analysis, and the cprp plot suggested further improvements were possible. An incremental change was made in this case, with G being transformed to G^2. The resultant Enorm, 1.586, was slightly lower than that of model 2'. No further refinements improved the result, and the last equation derived was retained as the final model:

$$lnP = 2.3 + 0.0386M - 0.009S + \frac{2.77*10^2}{1.2*10^3C^2+1.5*10^3} + \frac{1.05*10^6}{2.4*10^4L^2+3.1*10^6} - \frac{1.0*10^2}{52\ 5G^2+3\ 0*10^2}$$

A comparison study is conducted to see the effect of context definition in equation derivation. In addition to the group of 138 objects(group 1) used in the previous experiment, three more groups of objects are formed as the following: group 2 with 142 objects was again derived through the context definition step, group 3 and 4 containing 140 and 210 objects respectively are not real contexts. Their objects were randomly picked from the original data set. The Enorm of the residuals for the best quadratic and exponential models are listed in table 1. One notes that groups 1 and 2 which define relevant contexts produce much more close fit models than groups 3 and 4 that are defined randomly. Therefore, deriving proper context by clustering is very important in fitting accurate analytic models to the data.

5 Conclusions

Our work on scientific discovery extends previous work on equation generation from data[Zytkow and Zembowicz, 1993]. Given complex real world data, clustering methodologies and a suite of graphical and statistical tools are used to define empirical contexts in which the set of independent variables that are relevant to the goal variable are first established. Empirical results indicating that the combination of multi-variable regression with the cprp technique is effective in cutting down the search for complex analytic relations between sets of variables.

Currently, we are looking at adopting approaches developed in MARS[Sekulic and Kowalski, 1992] to transform the chosen independent variables using the given relations, and then combine MARS's systematic search method to come up with the nonlinear base models. In future work, we hope to systematize the entire search procedure further, and develop a collection of tools that facilitates every aspect of the scientific discovery task(see Fig. 1).

Acknowledgments: We would like to thank Dr. Jim Hickey and Dr. Ron Day of Arco for their help as geological experts in this project.

References

[Allen and Allen, 1990] P.A. Allen and J.R. Allen. *Basin Analysis: Principles and Applications.* Blackwell Scientific Publications, 1990.

[Bhaskar and Nigam, 1990] R. Bhaskar and A. Nigam. Qualitative physics using dimensional analysis. *Artificial Intelligence*, 45:73–111, 1990.

[Biswas *et al.*, 1995] G. Biswas, J. Weinberg, and C. Li. *A Conceptual Clustering Method for Knowledge Discovery in Databases.* Editions Technip, 1995.

[Fayyad and Uthurusamy, 1994] U.M. Fayyad and R. Uthurusamy. Working notes: Knowledge discovery in databases. In *Twelfth AAAI.* 1994.

[Forbus, 1984] K.D. Forbus. Qualitative process theory. *Artificial Intelligence*, 24:85–168, 1984.

[Greene, 1988] G. Greene. Quantitative discovery: Using dependencies to discover non-linear terms. Master's thesis, Dept of Computer Science , University of Illinois, Urbana-Champaign, 1988.

[Jain and Dubes, 1988] A.K. Jain and R.C. Dubes. *Algorithms for Clustering Data.* Prentice Hall, Englewood Cliffs, 1988.

[Kokar, 1998] M.M. Kokar. *COPER: A Methodology for Learning Invariant Functional Descriptions.* 1998.

[Langley and Zytkow, 1989] P. Langley and J.M. Zytkow. Data-driven approaches to empirical discovery. *Artificial Intelligence*, 40:283–312, 1989.

[Langley *et al.*, 1983] P. Langley, J.M. Zytkow, J.A. Simon, and G.L. Bradshaw. *The Search for Regularity: Four Aspects of Scientific Discovery*, volume II. 1983.

[Rao and Lu, 1992] B.G. Rao and S.C-Y. Lu. Keds: A knowledge-based equation discovery system for engineering problems. In *Proceedings of the Eighth IEEE Conference on Artificial Intelligence for Applications*, pages 211–217, 1992.

[Sekulic and Kowalski, 1992] S. Sekulic and B.R. Kowalski. Mars: A tutorial. *Journal of Chemometrics*, 6:199–215, 1992.

[Sen and Srivastava, 1990] A. Sen and M. Srivastava. *Regression Analysis.* Springer-Verlag Publications, 1990.

[Zembowicz and Zytkow, 1992] R. Zembowicz and J.M. Zytkow. Discovery of equations: Experimental evaluation of convergence. In *Proceedings of the Tenth Conference on Artificial Intelligence*, 1992.

[Zytkow and Zembowicz, 1993] J.M. Zytkow and R. Zembowicz. Database exploration in search of regularities. In *Journal of Intelligent Information Systems*, volume 2, pages 39–81, 1993.