

Data warehousing and mining technologies for adaptability in turbulent resources business environments

Shastri L. Nimmagadda* and Heinz Dreher

School of Information Systems,
Curtin Business School,
Curtin University, Perth, WA, Australia
E-mail: snimmagadda@bogota.oilfield.slb.com
E-mail: h.dreher@curtin.edu.au
*Corresponding author

Abstract: Resources businesses often undergo turbulent and volatile periods, due to rapid increase of resource demand and poorly organised resources data volumes. This volatile industry operates multifaceted business units that manage heterogeneous data sources. Data integration and interactive business processes, distributed across complex business environments, need attention. Historical resources data, geographically (spatial dimension) archived for decades (periodic dimension), are source of analysing past business data dimensions and predicting their future turbulences. Periodic data, modelled in an integrated and robust warehouse environment, are explored using data mining methodologies. The data models presented, will optimise future inputs in the turbulent resources business environments.

Keywords: resources business data; data warehousing; data mining.

Reference to this paper should be made as follows: Nimmagadda, S.L. and Dreher, H. (2011) 'Data warehousing and mining technologies for adaptability in turbulent resources business environments', *Int. J. Business Intelligence and Data Mining*, Vol.

Biographical notes: Shastri L. Nimmagadda is presently a Senior Geophysicist with the Schlumberger Company in Bogota, Colombia. He worked for several petroleum companies in India, Australia, Uganda, Kuwait, Abu Dhabi, Egypt and Colombia. He did his MTech and PhD in Exploration Geophysics from the Indian Institute of Technology, Kharagpur, India. He is currently engaged with an offshore research project on "Ontology based multidimensional warehousing of heterogeneous data", at Curtin Business School, Curtin University of Technology, Australia. He already published and presented more than 40 research and technical papers in various international journals and conference proceedings.

Heinz Dreher is Professor of Informatics in the School of Information Systems at Curtin University, Perth, Western Australia. His PhD was awarded for the thesis on "Empowering Human Cognitive Activity with Hypertext Technology". His current research interests are focused on deriving meaning from vast repositories of data through data mining and modelling, pattern searching, and semantic analysis of text, including for applications such as conceptual search, and automated essay grading, for business and education.

1 Introduction

Decision-based forecast systems are often required in many aspects of resources business operations. Resources industry that involves exploration, production and development of mineral, oil and natural gas handles of volumes of time-varying data (Hoffer et al., 2005; Pujari, 2002; Roiger and Geatz, 2003). Crucial decisions are made at different managerial levels by means of forecast models done at different periodic intervals, so that business operations carried out at different geographic (lateral dimension) locations and periodic (longitudinal dimension) times are assessed for better resources optimisation.

In this paper, an attempt has been made to review and document IT solutions to problems associated with management of resources industries. *Exploration*, which is one of the major key operations of resources industry, needs to address issues of forecasting of resources, required to operate the exploration business in oil and gas industry. *Drilling* is the next operation, in which again resources are forecast to perform the drilling tasks both on land and on offshore drill sites. Production is an ultimate upstream operation of resources industry, in which forecast of resources needed to exploit mineral, and oil and gas deposits in different basins is addressed. Rigorous expenditure analyses, namely cost vs. production analysis and cost vs. benefit analysis, are carried out to ensure growth and profitability of the business. Human resources are also required to make the resources industry's business run more efficiently and thus future forecast of manpower resources in the resources industry is a much needed task.

Ontology-based data warehousing and mining approaches (Flahive et al., 2004; Bhatt et al., 2004; Gornic, 2000; Mattison, 1996; Shanks et al., 2003) are proposed to adapt in these business environments and address these issues and challenges. Problems associated with the existing data-structuring approaches and the remedies have been discussed. Methodologies in organising heterogeneous data and data integration through data-warehousing approach are briefly discussed. Data mining and forecasting procedures for optimum utilisation of resources, analysis of results and discussions including interpretation of results in a knowledge domain that is understood by multiple domain champions are highlighted. Certain conclusions and recommendations discussed in the section are useful for project managers. The deduced forecast models provide inputs to corporate management on technological changes, economic conditions, predicting future company growth and industry. In making budgetary proposals for exploration and development of oil and gas fields, several predictions are necessary, because of varied overhead costs and multiple contractual and sub-contractual tasks and their costs. Literature survey on statistical data-mining studies and their shortcomings in analysing oil and company's data are highlighted.

2 Problem statement, issues and solutions

Forecast of precise and accurate models is crucial for sustainable technical and financial inputs and for up running the resources business. Nimmagadda and Dreher (2008a) demonstrate the adaptability of data warehousing and mining approaches addressing mapping of multidimensional data structures and domain knowledge modelling. Data integration and mining of specific knowledge domains facilitate interpretation of business knowledge in turbulent environments. Fine-grained data structuring and

denormalising data relationships among multiple dimensions of resources business are key ideas for demonstrating data warehousing and mining technologies. Lack of proper semantic knowledge and limited sharing of domain knowledge have held back in developing database logics in resources industries. Poor data integration and lack of understanding of operational and business knowledge are also causative to poor understanding of data warehouse and mining approaches. Data repositories are represented mostly in relational structures with several constraints in terms of applying business rules, with the result, inconsistency; less flexibility in data structuring and inability to adapt to fast changing company situations are reported. Sharing knowledge and common understanding of data structuring, reusing of domain knowledge, making domain assumptions more explicit, separating domain knowledge from the operational knowledge, analysing and interpreting domain knowledge are key remedies of the current problem issues.

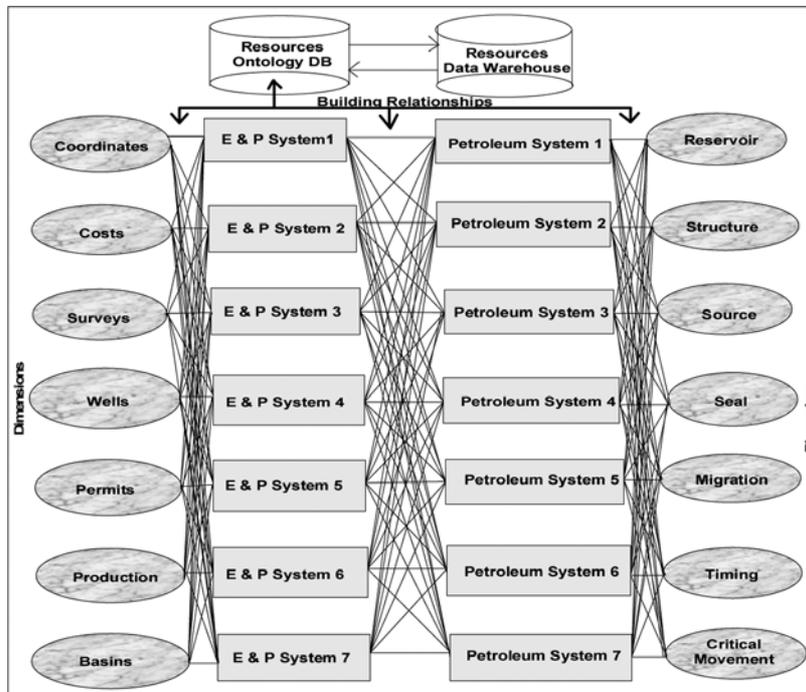
3 Review of literature

Rusu et al. (2005) described a methodology for building XML data warehouses. Nimmagadda and Dreher (2004, 2008a) and Nimmagadda et al. (2005) demonstrated feasibility and applicability of ontology-based multidimensional data warehousing and mining technologies in different application domains. Heterogeneous data structures are prevalent in industries and their organisation in the resources businesses and industries are documented. Aczel (1993) discussed concepts of business statistics and applications of statistical methods in the industrial scenario. Taniar et al. (2008) discussed the rule-mining procedures and exceptions in their usages in different categories. Tjioe and Taniar (2005) demonstrated associative mining rules in the context of data warehouses. Hair et al. (1984) described multivariate data analysis along with an application of time-varying industrial data. Dunham (2003) demonstrated several data-mining algorithms that handle complex data types. Gregersen and Jensen (2002) provided various conceptual time domain database models, in which implementation of models is discussed. Berenson and Levine (1992) provided concepts and applications of statistical methods in the business environments. Nimmagadda et al. (2006) demonstrated multidimensional database modelling in the Australian resources industry scenarios. Rudra and Nimmagadda (2005) discussed roles of multidimensional fine-grained data-structuring methodologies that affected the data-warehousing and data-mining implementations in resources business. Periodic historical resources data have been analysed and interpreted for business knowledge from correlations, trends and patterns of data views drawn from volume of databases. Use and misuse of statistical methods in information systems research have been discussed in Graham and Desmond (1992). Gupta (1990) provided an exhaustive literature on statistical modelling for business situations and demonstrated with industrial case studies. Roiger and Geatz (2003) demonstrate data mining algorithms with practical examples. These are classical mining methods, yet to be widely tested in resources business applications. Multidimensional modelling approaches (Nimmagadda and Dreher, 2007a, 2007b, 2008a, 2008b), used for warehousing resources data, are described in Section 4.

4 Data warehousing and mining methodologies

Flahive et al. (2004) and Bhatt et al. (2004) describe ontologies in distributed environments. Resources data belonging to multiple operational units are typically exploration, drilling, production and marketing. These business units may be operating both locally and globally, in a distributed system environment. Data sources and requirements for conceptualising and building data models are discussed exhaustively in Nimmagadda and Dreher (2008a). For integrating the data from multiple sources, they need to be cleaned, reformatted, logically and physically organised in an intelligent storage environment. As demonstrated in Figure 1, data from operational units gathered and stored in an integrated warehouse from which several data views are extracted for data-mining workflows.

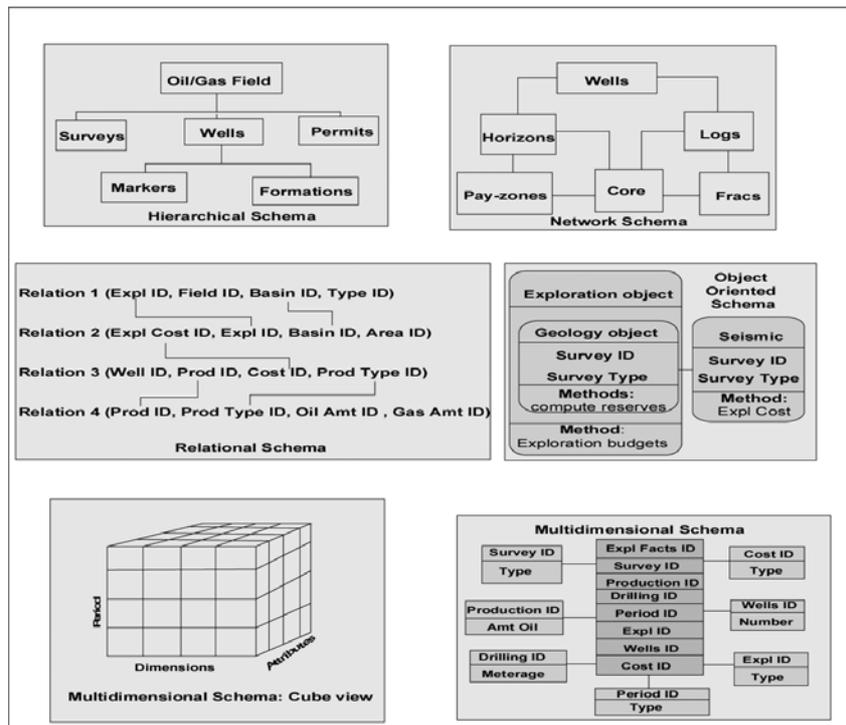
Figure 1 Managing several petroleum systems in an integrated business environment



Exploration in a resources business is a key generalised dimension, in which hundreds of dimensions and attributes are described (Nimmagadda and Dreher, 2007a, 2007b, 2008b). Several other factors, based on processes in the exploration business from which these data dimensions conceptualised, are narrated in Figure 1. During data integration process, data from several operational dimensions are gathered and inter-connected through clustering. Periodic dimension, which has been the subject of any multidimensional data-modelling approach, is addressed exhaustively in our methodologies. Resources data are organised in different data schema approaches, as narrated in Figure 2. Nimmagadda and Dreher (2008a) discuss a data-warehousing environment using multidimensional modelling. The star schema data models shown in Figure 2 represent how multidimensional data associate with *period* data attribute.

Petroleum system is an information system, which has inherent natural data entities and their associated attributes, still they need to be conceptualised and contextualised, if connectivity and or relationships need to be established among multiple system entities and attributes (as narrated in Figure 1). Analogous to entity, dimensions are used in our present data modelling schemas. For this purpose, an integrated framework is addressed. But dimension names and their contexts are semantically interpreted, which are not inherent, such as the ones' in any project management, *surveys*, *wells*, *permits*. Similar many other entities associated with *drilling*, *production* and *marketing* business units are interpreted. Major composite dimensions, such as oil and gas fields (mineral provinces in case of mining farms) in which *wells*, *boreholes*, *surveys*, *geological markers*, *seismic horizons* and multiple *log-profiles* (including *vertical seismic profiling*) *sub-type* dimensions are described, interact (communicate) logically between surface (known seismic data instances) and sub-surface unknown geology (drilled well data instances). There are many users in different operational centers, share these multi disciplinary datasets in different contexts.

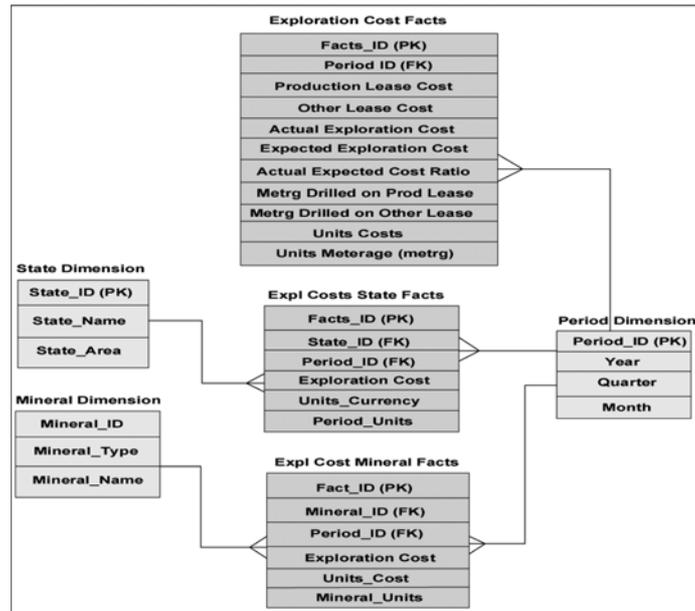
Figure 2 Different schema types considered in the resources data management



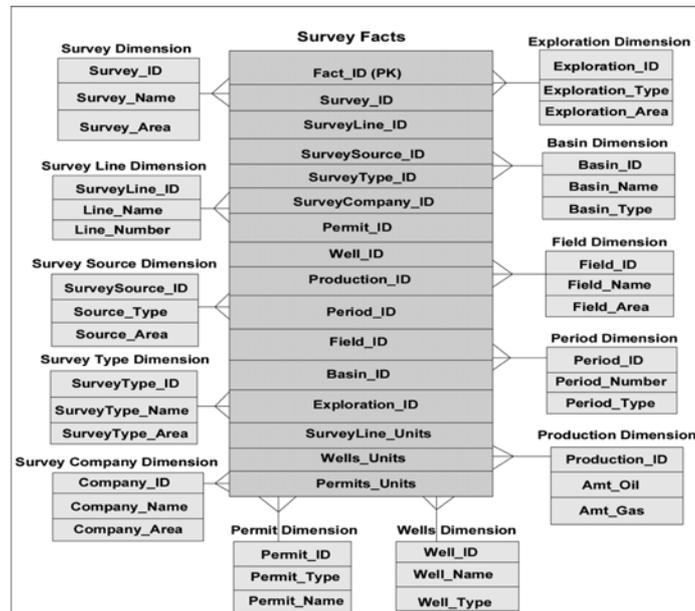
Under shared project management, various users handle their individual projects with several functions and activities. Areas Of Interest (AOI) sub-projects are also created to handle a specialised data, such as *well* and *seismic* data accessed by geologists and geophysicists working in a shared project under *exploration* entity or dimension. Other important functions of *exploration* operational unit are procurement of necessary financial budgets for conducting exploration operations. The authors describe them as external entities, in which each entity is narrated by a different data dimension and its

corresponding fact data tables as described in Figures 2 and 3. Total schemas considered are as narrated in Figure 2. Various concepts, forecasting methods and their practical applications in industries have been discussed in this section.

Figure 3 Multidimensional star schema data models for (a) *exploration costs* database and (b) *surveys* database



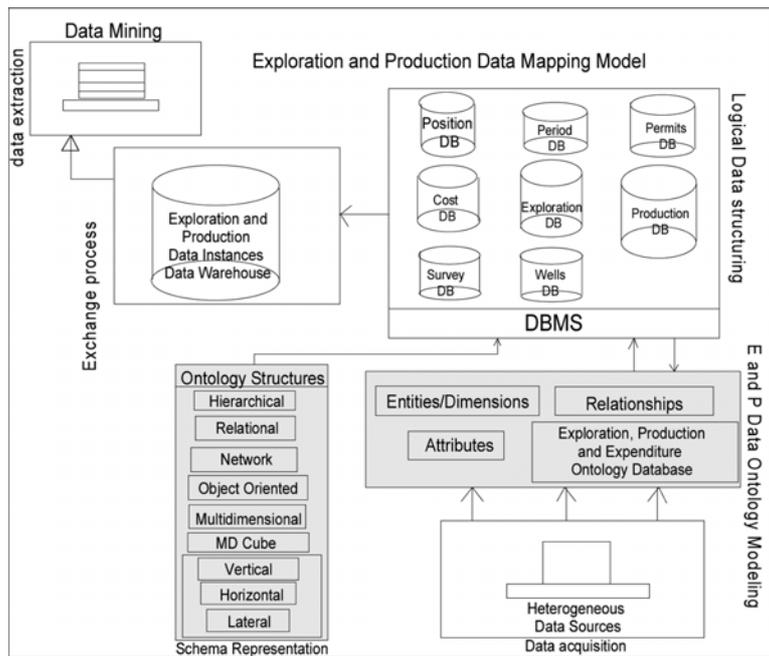
(a)



(b)

Surveys, wells and *permits* are key data sets used in our study for generating conceptual data models described in star schemas, as demonstrated in Figure 3. The integrated framework for warehousing resources data has been given in Figure 4, in which all the workflows are integrated including data modelling, warehouse and other analytic procedures needed to compute and interpret data views are described. Besides technical data, financial data have also been interpreted with mineral and petroleum exploration costs that affect industry turbulent business environments.

Figure 4 An integrated framework methodology, used for managing resources data



Source: Modified after Nimmagadda and Dreher (2008a)

As narrated previously, *exploration, drilling, production* and *marketing* are important sub-systems within a major oil company. Data needed in each sub-system have been identified and with the best use of data acquisition and processed data (information) repositories, manager of each sub-system confidently utilises this information to perform organisational functions. Database Management System (DBMS), which is a relational logical and physical data structure, has been incorporated with multidimensional designs in the data schemas. Tables with rows and columns have been relationally organised by common data attributes and their instances. This is an integrated model, combining the data warehouse with the data-mining utilities.

Daily, volumes of dimension and fact tables with multiple rows and columns are handled in petroleum business operations. Data modelling makes use of the concepts of similarity and associativity to inter-relate several data attributes among several dimension and fact tables. As shown in Figure 3, data models represent a link to multiple fact tables of a dimension named, *petro2-surveys* database. *Petro2-surveys* database consists of multiple fact and dimension tables, from which only representative dimensions have been denoted in Figure 3(a) and (b) for demonstration purposes. Overall integrated framework

as described in Figure 4 is used for integrating warehoused multidimensional data structures with data mining and other forecast procedures. Issues related to the forecasting and decision-making in the resources industry have been discussed in the following sections.

4.1 Forecasting vs. decision making

For decision making, *exploration* data and information analysis, personal judgements, evaluating alternative actions in terms of probable exploration costs and pay-offs are used. Impacts on entrepreneurship, while making company's operational decisions, have been addressed. Forecasting is used to simply predict future – it is not a decision making tool, but a key input to a decision model, therefore, forecast is a guiding tool.

4.2 Data vs. information

Data here refer to any number of facts that may be available in the resources industry. Information refers to that portion of data that is relevant to what happens in the future. One of the most difficult tasks in forecasting is separating informational content of data and conversely identifying these data, which has particular informational content in the mining or petroleum industry.

4.3 Basic elements of forecasting

There is wide diversity in presenting forecasting techniques; all have three basic elements in common:

- Which deal with situations in the future, i.e., every forecast must be made for a specific point of time, referred to as the time frame of the forecast
- All forecasts deal with some level of uncertainty; therefore, it is necessary to make assumptions, judgements, or hypothesis about relevant conditions and interactions. Some error in forecasting must be expected
- All forecasts must, to some extent, rely on information that is contained in historical data.

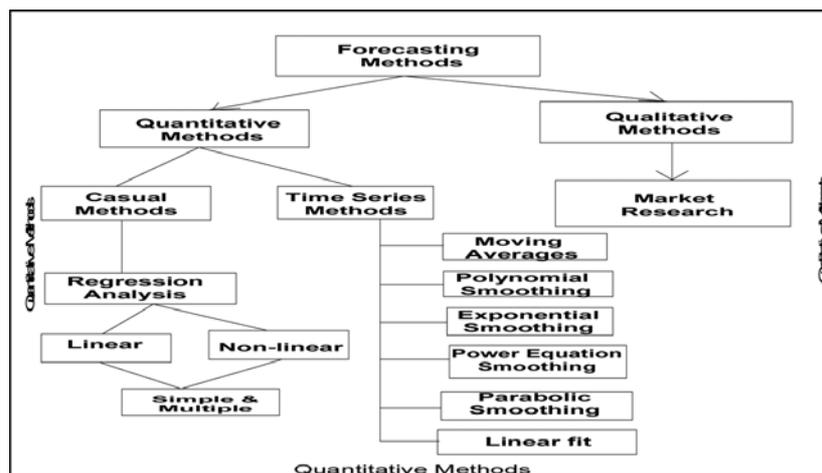
The forecasting methods discussed in the following sections also refer to various statistical techniques (as classic data-mining methodologies). Data that have been accessed in the form of data views from warehouse have been analysed and generated statistical models with varied attributes of the resources data. Time series analysis is used for quantitative data interpretation.

5 Forecasting procedures

The classification of the forecasting methods is discussed in Figure 5. Irrespective of any forecasting method used, the development of a particular forecast technique will require the completion of the following steps:

- determine the objective of forecast – what is it that we are trying to predict and for what purpose
- determine the variables to be forecast – such as *exploration costs*, *production* performance and other resources industry's performance indicators
- determine the time horizon (or period dimension) for the forecast – weekly, monthly, quarterly, half-yearly and annually
- decide on which forecasting method best suits – such as moving average and regression
- collect the required data – past data or records
- make the forecast
- implement the results and review the forecasting process through an effective feedback system.

Figure 5 Classification of forecasting methods



5.1 Time series forecasting

In time series models, the variables that do predictions are analysed with respect to historical periods and thus the data patterns are modelled. Such models assume that the time sequential patterns that have occurred in the past will also occur in the future. As such, historical quantitative data provide the basis of the time series analysis. For example, exploration costs or production data acquired are over a period of 50 years. These historical time series data interpret patterns and trends. These patterns are used to assist in developing forecasts as basis for future decision making in the resources industry. Time series analysis uses guidelines that are set for the future, but are designed around events that have taken place in the past. More observations or data instances can make the data mining procedures more effective and further realistic forecasts.

For time series analysis, forecasting methods take into account the following criteria:

- identification of the underlying trend line
- measurement of past data patterns and the assumption that these patterns will be repeated in future time periods
- forecast of the future trends.

6 Four main components of a time series

The scrutiny of historical set of quantitative data series identifies four key movements that are appropriate to the resources business data:

- secular trend
- cyclical movement
- seasonal movement
- irregular movement.

6.1 Secular trend

A secular trend identifies the underlying trend (direction) of the data – increasing, decreasing or remaining constant. It is a long-term direction of the data, usually described by the “line of best fit”. As an example, the increase in petroleum production is linearly proportional to its consumption. Increase in *exploration* costs or dollar fluctuations and market values of essential commodities will affect the discovery index of mineral or petroleum deposits.

6.2 Cyclical movement

This reflects the level of business activity and economic movement over time by fluctuating patterns. These variations measure periods of expansion and contraction in industry and the economy, in general. Their regularity and intensity are not predictable, however certain economic indicators contribute to their existence – level of investment, confidence, and confidence in the economy, GNP, trade indexes and government policy. Cyclical movements in the economy – for example, inflationary pressures – affect the time series trends. Any regular pattern of sequences of observations above and below the trend line, which lasts more than one year, is a result of the cyclical component of the time series. For meeting demand and supply of resources, economic measures taken in the resources industries that boost the mineral and petroleum production are interpreted to have associated with the cyclical movements of the periodic historical data.

6.3 Seasonal movement

Seasonal movement refers to regular periodic fluctuations that occur in each time period – usually yearly, monthly and daily variations in the resources data. Some examples

are special periods during discoveries of petroleum resources, monthly petroleum loads, and production decline and oil-field sickness. At times, seasonal variations can also be documented in the quarterly exploration costs data. Seasonal variations greatly impact on the outcomes of recorded data and often belie the underlying trend. Businesses need to identify the following seasonal impacts:

- so that a measurable (index) can be used to adjust the expected outcome
- to recognise the direction of the underlying trend.

6.4 Irregular movements

These patterns refer to the uncontrollable, random variations that impact greatly on the level of resources business activity. Some examples are extreme weather patterns (flood, fire, cyclone), extreme business variation (stock market crash, drop in Aus \$\$), political climate (sudden elections, wars, death of a leader) and industry changes (pilot strikes, waterside strikes). The resulting patterns will exert a great pressure on the predicted underlying trends and such eventualities must be accounted for, when planning for the future. All these data patterns, described so far, have been identified and well documented in the resources data. Interpretation of these patterns is carried out for business intelligence analysis in the upcoming sections. Forecasting methodologies have been presented briefly in the next sections.

7 Forecasting: general methodology

In general, forecasting focuses on the decomposition of historical data into each of the above-mentioned components, estimating each pattern separately, and then combining the projected impact of each component in the future, to produce the final forecast. For instance, the determination of whether a drop in sales is due to seasonal, random, or trend variations (or how much can be attributed to each) can be vitally important to any level of management in evaluating current policies and indicating the corrective action required. The methodologies considered in the present data computations are more classic and derived from basic statistical techniques. Before interpreting the quantitative models, a brief discussion of computational methods is given in the next section.

7.1 Computational methods

The moving average, multivariate regression, polynomial regression, construction of exponential, power and linear equations using the actual resources data have been addressed. These have been presented in the following sections.

7.1.1 Forecasting using smoothing methods

Smoothing techniques are appropriate for forecasting purposes in those situations where the time series is fairly stable in which there is no significant trend, cyclical or seasonal effects. In these situations, the objective of the forecasting method is to 'smooth' out the irregular component of the time series through some type of averaging process. The methods to be covered here include:

- moving average
- weighted moving average
- exponential smoothing.

7.1.1.1 *Moving average*

The moving average method consists of computing an average of the most recent n data values in the time series. Average is then used as the forecast for the next period. Mathematically, the moving average calculation is made from the following:

$$\text{Moving average} = \frac{\text{sum of most recent } n \text{ data values}}{N}.$$

7.1.1.2 *Weighted moving average*

The moving average method provides equal weights to actual data observed in each period. This is one of the reasons for its slow reaction to variations in the most recent observations in the time series of resources data. A modified version, using the weighted moving averages, is used to assign a greater weight to the more current data.

For example, weightings may be applied as follows:

- 3 to the most recent observations $t - 1$
- 2 to the second most recent observations $t - 2$
- 1 to the third most recent observations $t - 3$.

[the sum of the weights is equal to 6, i.e. $(3 + 2 + 1)$], and the forecast may be calculated as follows:

$$\text{Weighted moving average} = \frac{(3t - 1 + 2t - 2 + 1t - 3)}{6}.$$

A major problem, in our case, is the determination of the weightings to be assigned. If greater weight is assigned to the most recent observations, the weighted average may over-react to an irregular movement. However, if the weight assigned to the most recent observations is not too much greater than that is assigned to earlier observations, the meaning of weighted average may be lost.

7.1.1.3 *Exponential smoothing*

The exponential smoothing forecast method attempts to predict the time series in the next period based on the moving average of the current method. This method also weights on the most recent observations, more heavily than older data. Consequently, the most recent changes are strongly reflected in the forecast. Exponential smoothing uses a single weighting factor called *alpha* symbolised as α . The exponential smoothing formula is as follows:

$$E_t = \alpha A_{t-1} + (1 - \alpha) E_{t-1}$$

or

$$E_t = E_{t-1} + \alpha (A_{t-1} - E_{t-1}).$$

Both formulae provide the same result, where

E_t : Forecast of the time series

A_{t-1} : Actual or observed time series value for the most recent period $t - 1$

E_{t-1} : Forecast of the time series for the most recent period $t - 1$ (old forecast)

α : Smoothing factor, which has a positive value between 0 and 1

$1 - \alpha$: Damping factor (if $\alpha = 0.3$, then damping factor = 0.7).

When starting with exponential smoothing calculations, the first actual result becomes the forecast for the second period, and from then on, the exponential smoothing formula will provide the forecast for the next data event. These forecasting methods are more basic, but have been used in the present computations to test the validity of these techniques. In our studies, basic and more advanced data-mining techniques are combined to confirm the forecast. Each one of these mining techniques has been described in the following sections.

7.1.1.4 Regression analysis

Statistical regression (Graham and Desmond, 1992) is a supervised learning technique that generalises a set of numeric data by creating a mathematical equation relating one or more input attributes to a single numeric output attribute. Regression analysis is used for the purpose of prediction. In this study, a statistical model is developed through regression analysis. This model is used to predict the values of a dependent or response variable based on the values of at least one independent variable. Regression analysis or curve fitting is a procedure (Gupta, 1990) for estimation of average of a variable (Y) corresponding to a given value of X . This is called regression of Y on X . If the average value of X corresponding to a given value of Y is estimated, then it is known of X on Y . Depending on the fancy of the work, several regressions are straight lines to a given set of points. However, regardless of the type of curve fitted, there exists a relationship between two variables, which can be defined with the help of a correlation coefficient.

The correlation coefficient cannot exceed one and can be less than -1 . A value of 1 denotes perfect functional relationship between Y and X , an increasing X being associated with an increasing Y . A value of -1 indicates perfect functional relationship, though now an increasing X is associated with a decreasing Y . The graphs show different configurations of forecast at plotted points. Configurations of plotted points may or may not indicate trends or relationships in the data. But quite often, such inherent trends or data relationships are measured when they actually exist. Curve fitting is the solution for a given problem, in which period a set of points is fitted along a particular trend.

For example

$Y = a + bX$ is a linear equation in which points (X_i, Y_i) lie on a line.

From this equation, it is significant to derive and analyse the values of a and b from the constructed line. For every unit of increase in X , there is corresponding change in Y , thus b measuring the steepness of the line and it is termed as regression coefficient.

When the value of b is positive, the line ascends from left to right. When b is negative, the line descends from left to right.

7.1.1.5 Fitting straight line by method of least squares

This implies finding the values of the parameters a and b of straight line that fits with actual observed data. The “least squares” method assumes the best-fitting line in which the sum of the squares of the vertical distances of the points (X_i, Y_i) from the line is minimal. The vertical distance E_i from the line of any point P_i with coordinates (X_i, Y_i) is

$$E_i = Y_i - (a + b X_i).$$

The best-fitting line is that line for which the sum of squares $\sum E_i^2$ is minimum.

7.1.1.6 Method of computing multivariable regression

The simplest kind of relation between two statistical variables X and Y , a least square line, is made up of a line called regression line. For a change in the value of variables X , there is corresponding variation in the variable Y . A trend or data relationship is defined by constructing equations between two variables X and Y . The strength of regression between two different attribute variables is interpreted by means of correlation coefficient and association between variables.

7.1.1.7 Computation of correlation coefficients

It is the degree of similarity, both in direction and in magnitude, of variations in corresponding pairs of observations of two variables. Simple correlation implies finding out degree of association between pairs of observations.

In this study, the following Karl Pearson’s Method of finding correlation coefficient has been used:

$$R = \frac{\sum XY}{N \sigma_x \sigma_y}.$$

Alternative formula is

$$R = \frac{\sum XY}{\left(\sum X^2 \sum Y^2 \right)^{1/2}}$$

where X = difference between the actual period value and the average of its values, and Y = the difference between the actual cost and its average values.

When actual means are taken in fraction, the use of the above-mentioned formula becomes time consuming. The following assumed method is used for computing the correlation coefficient when the cost values are in fractions:

$$R = \frac{\left(\sum dx dy - \left(\sum dx \sum dy \right) / N \right) / \left[\left(\sum dx^2 - \left(\sum dx \right)^2 / N \right)^{1/2} \right]}{\left(\sum dy^2 - \left(\sum dy \right)^2 / N \right)^{1/2}}.$$

At places, probable and standard errors of Correlation Coefficients have been computed using the following formulae:

Probable error of $R = 0.6745 ((1 - (R)^2)/N)$ where $N =$ Total observations

Standard error of $R = ((1-(R)^2)/N)$.

The following pairs of attribute instances have been considered for computing correlation coefficients (for different mineral exploration costs):

Mineral exploration costs

- 1st Quarter/2nd Quarter of Coal Exploration Cost: $R = 0.753$
- 1st Quarter/Yearly of Coal Exploration Cost: $R = 0.848$
- 2nd Quarter/Yearly of Coal Exploration Cost: $R = 0.969$
- 3rd Quarter/Yearly of Coal Exploration Cost: $R = 0.96$
- 4th Quarter/Yearly of Coal Exploration Cost: $R = 0.965$
- 1st Quarter/Yearly of Diamond Exploration Cost: $R = 0.80$
- 2nd Quarter/Yearly of Diamond Exploration Cost: $R = 0.869$
- 3rd Quarter/Yearly of Diamond Exploration Cost: $R = 0.78$
- 4th Quarter/Yearly of Diamond Exploration Cost: $R = 0.80$
- Nickel Actual Exploration Cost/Computed Cost: $R = 0.68$
- Gold Actual Exploration Cost/Computed Cost: $R = 0.90$
- Base metal Actual Exploration Cost/Computed Cost: $R = 0.870$
- WA Actual Mineral Exploration Cost/Computed Cost: $R = 0.94$
- QLD Actual Mineral Exploration Cost/Computed Cost: $R = 0.94$
- NT Actual Mineral Exploration Cost/Computed: $R = 0.850$
- Rest Australia Actual Mineral Exploration Cost/Computed Cost: $R = 0.86$.

Estimated errors in correlation

- Probable Error of R for Base Metal Exploration Cost/Computed Cost = 0.026
- Standard Error of R for Base Metal Exploration Cost/Computed Cost = 0.039
- Probable Error of R for Gold Exploration Cost/Computed Cost = 0.021
- Standard Error of R for Gold Exploration Cost/Computed Cost = 0.031
- Probable Error of R for Nickel Exploration Cost/Computed Cost = 0.059
- Standard Error of R for Nickel Exploration Cost/Computed Cost = 0.088.

Petroleum exploration costs

- Actual/Expected Offshore Petroleum Exploration Cost: $R = 0.92$
- Actual Offshore Exploration Cost/Computed Cost: $R = 0.91$
- Onshore Actual Petroleum Exploration/Expected Cost: $R = 0.658$

Basin-wise petroleum production

- Gippsland/Total: $R = 0.22$
- Gippsland/Eromanga: $R = 0.70$
- Gippsland/Carnarvon Barrow: $R = 0.899$
- Gippsland/Perth: $R = -0.306$
- Eromanga/Carnarvon Barrow: $R = 0.410$
- Carnarvon Barrow/Perth: $R = -0.360$.

Other petroleum industry performance indicators

- Number of Surveys/Number of Wells Drilled: $R = 0.680$
- Number of days surveyed/Number of days wells drilled: $R = 0.26$
- Total Survey Line Kilometres/Total Depth Metres Drilled: $R = 0.615$
- Number of Surveys/Number of Hydrocarbon-Producing Wells: $R = 0.470$
- Number of Surveys Conducted/Number of Structures Interpreted: $R = 0.655$
- Number of Structures/Number of Hydrocarbon-Producing Wells: $R = 0.932$
- Number of Wells Drilled/Number of Hydrocarbon-Producing Wells: $R = 0.88$.

Essential difference between correlation and regression

Correlation analysis, in contrast to regression, is used to measure the strength of the association between quantitative variables. For example, number of surveys conducted may have a correlation to number of wells drilled (attribute) in an oil-field area. The cost of exploration may be proportional to discoveries made (a dimension, in our multidimensional modelling analysis) and the production forecast. In a pure correlation problem, a sample of pairs of observations is chosen from a bivariate. A trend is defined by correlating these two pairs. Here, the functional relationship that exists is reversible from the statistical standpoint. In a pure regression problem, there is an independent or casual variable X and a dependent variable Y . The values of X assumed to be selected in advance and held fixed, and then the corresponding values of Y are monitored.

7.1.2 Two lines of regression

In a correlation problem, it is sometimes useful to consider two lines of regression, that of Y on X and that of X on Y . In the former case, a least squares line that is minimised, sum-up the squares of the vertical or Y distances of the points from the line that is used. In the latter case, the sum of the squares of the horizontal or X distances of the points

turns the line. After having evaluated the plots between actual exploration cost values and the period, parabola type of curve has been computed. Regression analysis has been carried out to further evaluate this parabola curve.

The following equation is used for fitting the second-order parabola:

$$Y = a + bX + cX^2.$$

Calculating the values of a , b , c , is crucial for fitting the parabola curve. Since the deviations of X and Y series can be taken from their means, the normal equations are reduced to:

$$\sum Y = na + c\sum X^2$$

$$\sum XY = b\sum X^2$$

$$\sum X^2Y = b\sum X^3 + c\sum X^4.$$

Using these equations, the parabolic trend values have been computed for the following actual mineral exploration cost values (for Australian situations):

- Queensland Actual Mineral Exploration Cost

$$Y = 6590.87 + 1816.58X + 104.05X^2$$
- Western Australian Actual Mineral Exploration Cost

$$Y = 15634.35 + 4557.2X + 258.23X^2$$
- Rest of Australia Actual Mineral Exploration Cost

$$Y = 24240.14 + 2268.08X + 38.32X^2$$
- Northern Territory Actual Mineral Exploration Cost

$$Y = -1143.14 + 393.53X + 36.34X^2$$

where X = period and Y = exploration cost.

Similar parabolic trend values have also been computed for the following exploration cost values of

- Base metal Minerals

$$Y = 44607.94 + 3694.43X + 33.36X^2$$
 (Figure 12)
- Gold Mineral

$$Y = -5343.57 + 4859.74X + 394.23X^2$$
 (Figure 13)
- Nickel

$$Y = 6057.9 + 481.22X + 9.35X^2.$$

where X = period; Y = Exploration Cost.

Parabola curve fitting method has also been used for evaluating the offshore petroleum exploration costs and trend values have been computed for the following:

- Offshore petroleum other exploration cost

$$Y = 124.5 + 10.3X + 0.28X^2$$

- Other Petroleum Lease Costs

$$Y = 555.03 + 12.62X + 0.24X^2$$

- Offshore Actual Exploration Cost

$$Y = 519.77 + 17.79X + 0.106X^2.$$

where Y = Exploration Cost; X = Period.

7.1.3 Types of regressive models and computation of regression equations

The nature of the relationship can be in many forms, ranging from simple mathematical functions to extremely complicated ones. The simplest one is a linear or straight-line relationship in which for each increase in one value, there is corresponding increase in the other. But some relations may be curvilinear, in which with increase in value, there is corresponding decrease in the other value.

$$X \text{ on } Y: x - X = r\sigma_x/\sigma_y (y - Y)$$

$$Y \text{ on } X: y - Y = r\sigma_y/\sigma_x (x - X)$$

$$r\sigma_x/\sigma_y = \left(\frac{\sum d_x d_y}{\sum d_x^2} \right) / \left(\frac{\sum d_x^2}{N} \right)$$

where $X = Y =$ average of the observed or actual values; d_x and d_y are the difference of the average and actual values. Using these formulae, regression equations have been computed for the following pairs of Industry Performance indicators:

- Number of Surveys Conducted (x)/Number of Wells Drilled (y):

$$x = 0.605y + 17.22; y = 0.773x + 2.11.$$

- Number of Surveys (x)/Petroleum-Producing Wells (y)

$$x = 0.506y + 26.99; y = 0.437x + 0.11.$$

- Number of Wells Drilled (x)/Number of Petroleum-Producing Wells (y):

$$x = 1.0675y + 12.63; y = 0.722x + 5.61.$$

- Number of Surveys (x)/Number of Structures interpreted (y):

$$x = 0.591y + 20.06; y = 0.726x - 0.4.$$

- Number of Structures (x)/Number of Petroleum-Producing Wells (y):

$$x = 1.11y + 7.84; y = 0.78x - 4.06.$$

Some of the regression models have also been represented in the form of polynomial regressions and regression with exponential and power equations. A brief description of the different regressions has been given here.

7.1.3.1 Polynomial regression

Statistical analysis carried out in the business organisations with current and past data (also in time domain) has been well demonstrated in Aczel (1993) and Berenson and Levine (1992). Often, the relationship between dependent variable, Y , and one or more of the independent X variable is not a straight-line relationship but, rather, has some curvature to fit. In our present analysis, in each of the situations shown, a straight line provides a poor fit to the data. Instead, polynomials of the order higher than 1, i.e., functions of higher powers of X , such as X^2 , X^3 , provide much better fit to the data. Such polynomials in the X variable, or several X_i variables are still considered linear regression models. The multiple linear regression models thus cover situations of fitting data to the polynomial functions. The general form of a polynomial regression model in one variable X is:

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \dots + \beta_m X^m + \varepsilon.$$

where M is the degree of polynomial. The degree of the polynomial is the order of the model. The polynomial regression fit displays a curve based on the above-mentioned equation. The polynomial degree can be set from 0 to 10. A polynomial degree of zero is the average Y value, degree one is a linear fit, degree two is a quadratic fit, degree three is cubic fit and degree four is a quadric fit. Polynomial equations have been constructed for some of the situations of the data explored from *petro-2 surveys* database, as discussed in Figures 1, 2, 3, 8–25 and 31–35.

7.1.3.2 Exponential regression

One method often useful in forecasting time series (Aczel, 1993) is exponential regression. One of such methods is simple exponential smoothing, a useful method for forecasting time series that have no pronounced trend or seasonality. The concept is an extension of a moving average. In exponential smoothing, more recent values of the time series are allowed to have greater influence on the forecasts of future values than the more distant observations. Exponential smoothing is based on a weighted average of current and past series values. The largest weight is given to the present observation, less weight to the immediately preceding observation, even less weight to the observation before that, and so on. The weights decline geometrically as one goes back in time. The exponential smoothing model:

$$Z_{t+1} = w(Z_t) + (1 - w)(Z_t);$$

where Z_t is the actual known series value at time t , and Z_t is the forecast value for time t .

In our forecast studies, the following exponential equation fits with actual data:

$$\ln Y = bX + a \quad \text{or} \quad Y = a e^{bX}.$$

Using these concepts and formulae, exponential equations have been constructed using the queried data, extracted as shown in Figures 8–35. Power fit has also been used in the present studies for some situations. The following power fit displays a power fit through the data:

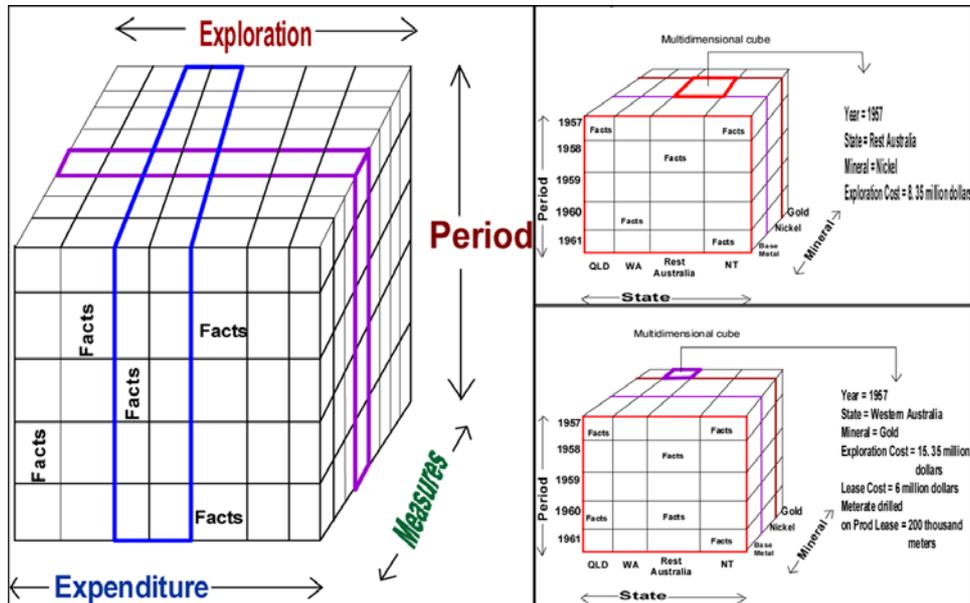
$$\ln Y = b(\ln X) + a \quad \text{or} \quad Y = a X^b.$$

Using this relation, power equations have been computed for some situations of queries generated from Mineral and Petroleum Exploration as shown in Figures 8–35.

8 Analysis of results and discussions: forecast reliability

Many queries and sub-queries have been generated and plotted using grapher and an intelligent data analyser solution. The statistical and graphic techniques have been used for visualisation of the data. Several data views drawn from 3D cubes (Figure 6) are visually represented and interpreted them in terms of predictions and forecasts. It is observed visually from the characteristics of time series data plots that the construction expenditures have shown a tendency to increase in a curvilinear over a period of 30–50 years of time. In the present studies, data have been considered from 1953 to 2000 years. This overall long-term tendency or impression is known as trend. For particular periods of time, the observed values are dipping below the trend curve. They are representing the peaks of their respective business cycles. Any observed data that do not follow the smooth fitted trend curve modified by the aforementioned cyclical movements are indicative of the irregular or random factors of influence. When data are recorded monthly rather than annually, an additional factor has an effect on the time series data. It could be due to seasonal component. At certain periods of time, the trends appear to be irregular or random and seasonal at other periods of time.

Figure 6 Data views from multidimensional data cubes (see online version for colours)



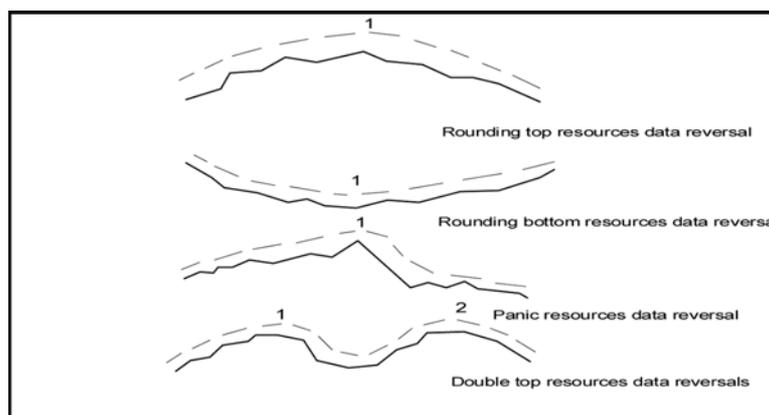
As described earlier, in addition to various time series components, Fayyad et al. (1996) discuss several other patterns of time series stock market data and detect temporal variations from which stock market knowledge has been discovered. They use dynamic programming approach for interpreting time series patterns. However, as shown

in Figure 7, similar shaped patterns have been observed in our historical resources data, which have been interpreted for forecast knowledge. Rounding top resources data reversal indicates more demand of resources at the period indicated '1', where economic boom or growth may be interpreted. In the rounding bottom reversal case, probably, at a period indicated at '1', there may be economic recession. In case of panic reversals, high trading of resources may be observed. There are two top reversals, which may be cyclic, indicating inflation pressures. All these patterns measure expansion and contraction of periodic dimension as envisaged in the resources database of resources industry, which forecast future economic growth.

8.1 Interpretation of the explored data

The computational data presented have been searched for data correlations, trends and patterns and thus for interpretation. Mineral exploration and discovery and petroleum exploration and production data dimensions have been analysed considering attributes and attribute instances that vary with time and space. The trends in the data in particular, when the data vary with period, have significance while forecasting for future resources. Four main types' trends are distinguished from mineral and petroleum data views. First one is secular trend, when attribute values are interpreted in increasing or decreasing direction. Second trend, called cyclical movement or variation, is due to expansion or contraction of economy. At times, inflationary pressures affect these cyclical movements. Third one is seasonal, interpreted in regular periodic fluctuations, for example monthly or quarterly. The last one is irregular variation or trend that refers to uncontrollable and random variations impacting greatly on the level of business activity. Some examples of this type are: extreme weather patterns (cyclone, floods, fires, etc...), extreme business variations (oil price down, stock market crash, drop in Australian dollar value) or extreme political situations.

Figure 7 Patterns analysis used in the resources data

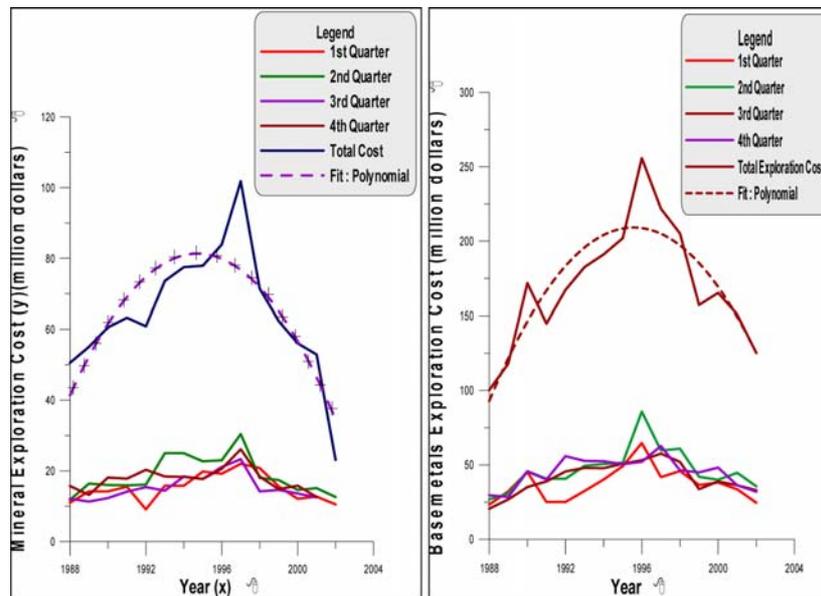


Source: After Fayyad et al. (1996)

As shown in Figure 7, data reversal implies similarities and coherencies at early and latter periods. Quarterly data have been examined and investigated for similarity trends in the exploration costs data except in the second quarter during 1994. An economic peak has

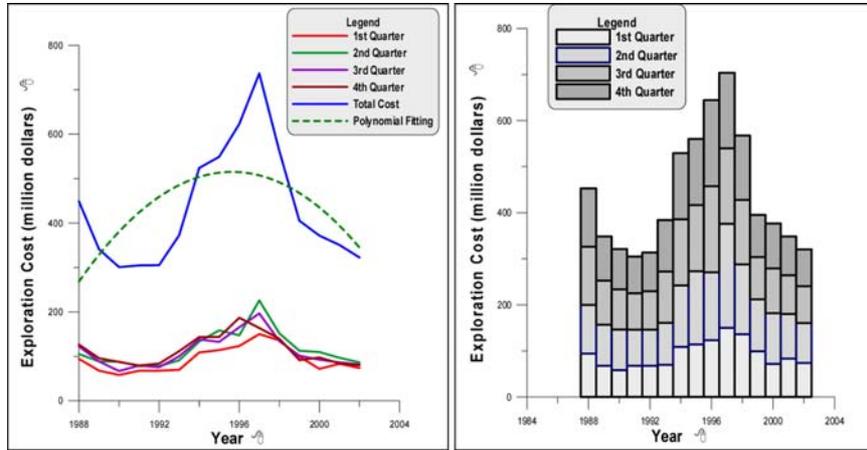
been interpreted for all quarters, indicating a rigorous mineral exploration and mining activity during this period. A polynomial equation has been fitted with actual cost data, with a good correlation. As discussed in Figure 8, rounding top resources data reversals (Figure 7) have been observed. The variations observed in the actual data during years 1992 and 1997 are cyclic or seasonal (see Figure 8). There is more investment on mineral exploration, thereby indicating more demand of minerals, in particular with state of New South Wales (NSW).

Figure 8 Quarterly presentation of NSW state's mineral (base metals) exploration cost data (see online version for colours)



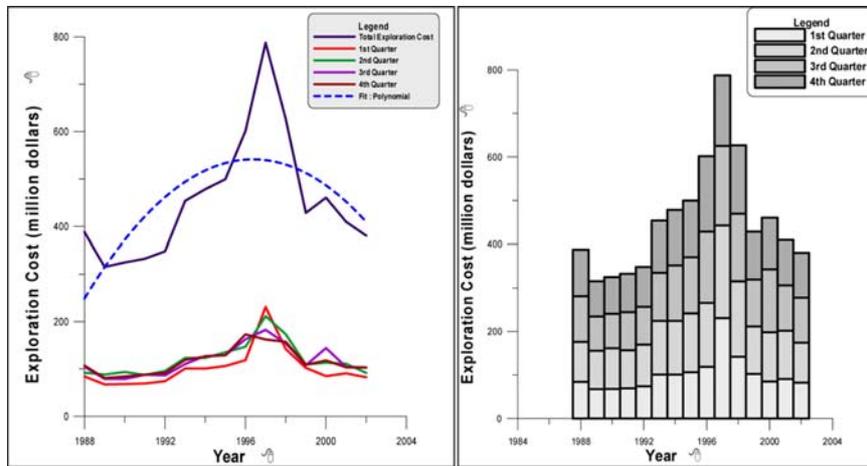
Quarterly exploration costs data have been (Figure 8) plotted for base metal minerals. There are periodic fluctuations especially during the years 1990, 1996 and 1999, which could be due to seasonal as illustrated in Figure 8. Maximum exploration cost is interpreted during 1996, with a decreasing trend and increasing period, which appears to be rounding top resources reversal trends (as patterns of models envisaged in Figure 7). A polynomial equation has been constructed and thus used for future exploration cost prediction. Quarterly exploration cost for gold has been plotted (see Figure 9) for detecting any trends in the cost data. Maximum and minimum exploration costs have, respectively, been interpreted during the years 1997 and 1991. In all these quarters, similar exploration cost patterns have been interpreted. But, the computed trend, as detailed in Figure 9, does not fit well with the actual cost data, because the actual data contain unexpected exploration costs, and indirectly affecting rise of prices and market demand. Panic reversals of data patterns have been observed (as patterns of models envisaged in Figure 7), indicating high volume of gold mineral trading during 1997. As shown in Figure 9, bar chart is another way of presenting the gold exploration cost data.

Figure 9 Quarterly presentations of gold exploration cost data (see online version for colours)



In general, as narrated in Figure 10, the quarterly exploration expenditure patterns are similar in all quarters in the Western Australian state. Maximum expenditure is reported in 1997 and minimum in 1990. The computed trend provides a poor match with actual exploration cost, with lot of fluctuations in exploration costs data. Here, panicked reversal patterns (Figure 7) indicate high volume of minerals trading in the WA state during the year 1997.

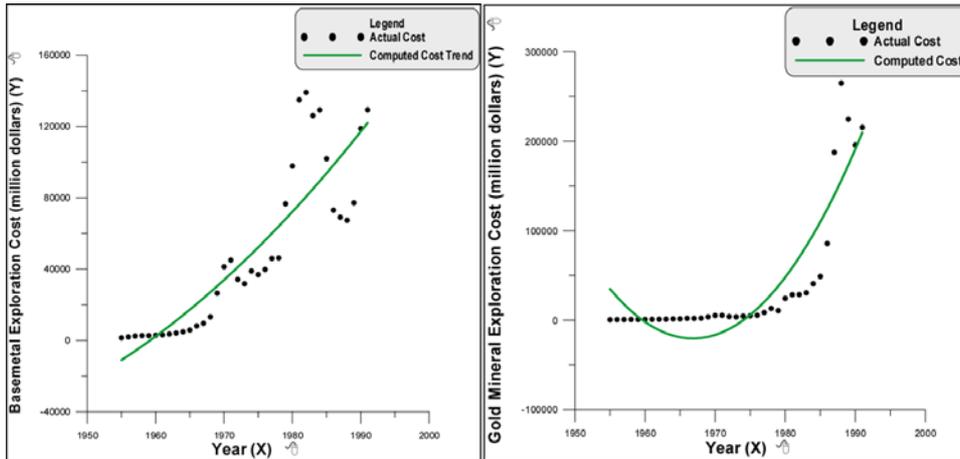
Figure 10 Quarterly presentation of WA state mineral exploration cost data (see online version for colours)



Upward and downward periodic movements of actual exploration cost around the computed trend are cyclical and irregular at some periods as illustrated in Figure 11. Seasonal variations affect these trends in the actual data. In general, base metal exploration cost is in increasing trend with period. A parabolic curve has been fitted with the actual data and thus an equation is constructed. Large periodic fluctuations observed between 1980 and 1990 are worth mentioning for base metals exploration. The variations are smaller in between 1955 and 1980. Peaks and troughs reported during

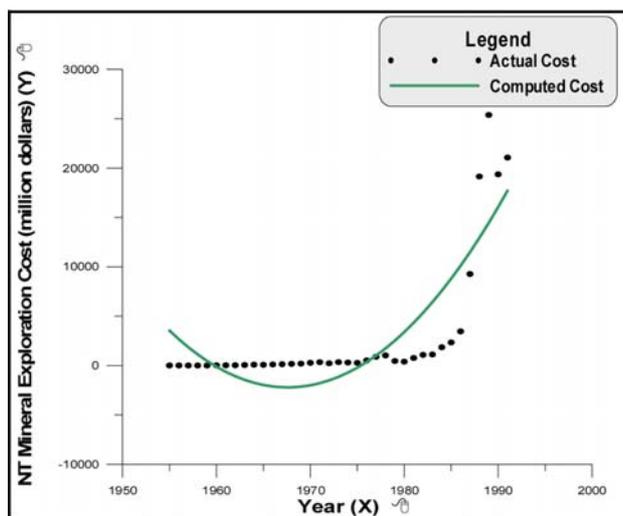
1965–1975 have been interpreted in the business point of view (see Figure 11). Any observed data that do not follow the smooth fitted curve without cyclical movements are indicative of random factors of influence. Secular and regular periodic fluctuation trends are again due to seasonal effects.

Figure 11 Construction of parabolic equations for base metal and gold actual exploration cost (see online version for colours)



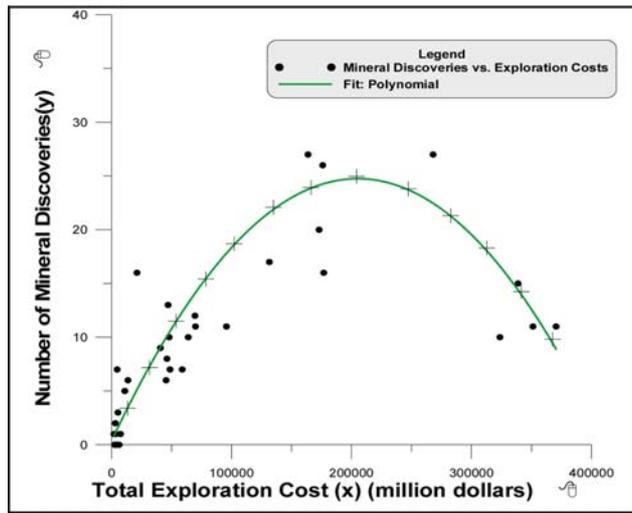
As illustrated in Figure 11, a parabolic curve has been fitted with the data of actual exploration cost of gold, which matches well. This equation may be used for predicting the future exploration costs of gold mineral in Australia. Again another parabolic curve match fit (Figure 12) has been carried out with the actual exploration cost data of Northern Territory (NT). Fairly good match has been observed and may be used to compute future exploration costs in northern Australia.

Figure 12 Construction of parabolic equation for NT mineral exploration cost (see online version for colours)



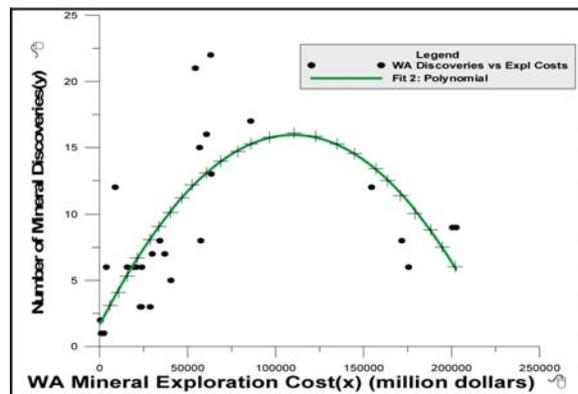
Total mineral exploration cost and number of mineral discoveries (both viewed as dimensions in the database) made in Australia have been plotted (Figure 13) for establishing any trends or correlations. In general, the actual data are not user friendly with the computed trend. There are many random fluctuations in the actual data and the constructed polynomial equation interprets an increase in exploration cost trend with corresponding decreasing trend in the number of mineral discoveries. As shown in Figure 13, computed trend indicates that even after increase in exploration cost, the number of mineral discoveries substantially has fallen down.

Figure 13 Construction of polynomial equation between total exploration cost and number of mineral discoveries made in Australia (see online version for colours)



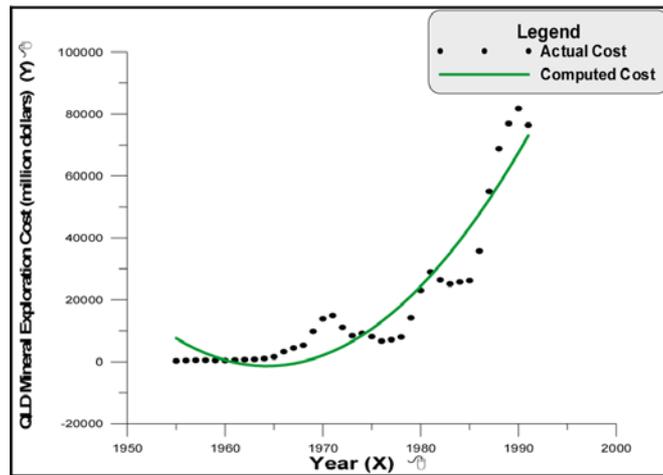
Similar analysis has been made in the case of Western Australian mineral discoveries vs. mineral exploration cost data (Figure 14). The number of mineral discoveries has fallen down in WA even with the increase in mineral exploration costs. Future trends can be predicted from the polynomial curve fit provided in Figure 14.

Figure 14 Construction of polynomial equation between (Western Australia) WA mineral exploration cost and mineral discoveries made (see online version for colours)



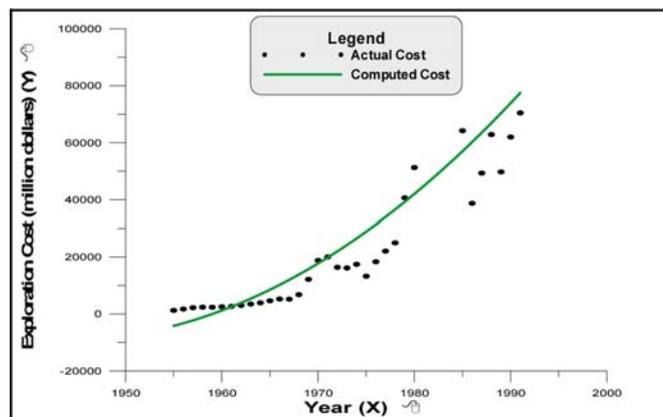
Similar analysis has been done using Queensland (QLD) State’s mineral exploration cost data (Figure 15) and the number of mineral discoveries made. In general, the parabolic curve match fits with the actual data. There are periodic fluctuations in the exploration costs particularly in the years 1970, 1980 and 1985 with peaks and troughs, which are interpreted as seasonal. The computed trend again can guide the future predictions in the QLD state. A minimum exploration cost has been reported in the year 1965, as shown in Figure 15.

Figure 15 Construction of parabolic equation for Queensland (QLD) state actual mineral exploration cost data (see online version for colours)



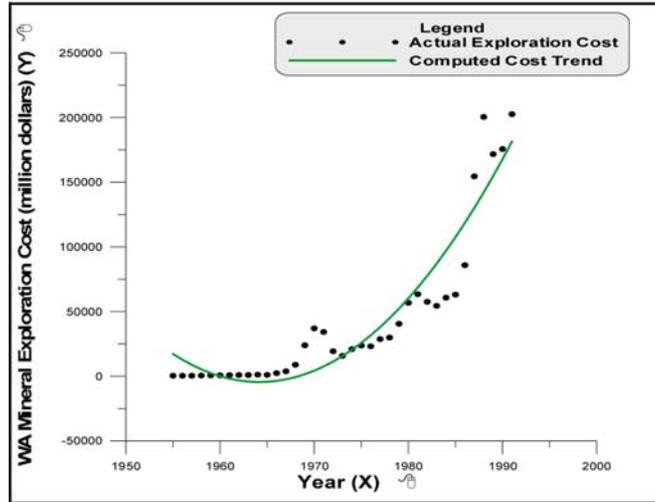
As illustrated in Figure 16, mineral exploration costs in the rest of Australia (excluding WA, NT and QLD states) seem to be an increasing trend with period (secular variation), but could have been affected due to seasonal variations especially during 1970 and 1988. The parabolic computed trend matches (Figure 16) with actual data, but with minor fluctuations in the exploration costs around the computed trend. However, the computed trend may be used to predict the future exploration costs.

Figure 16 Construction of parabolic equation for mineral exploration costs data (see online version for colours)



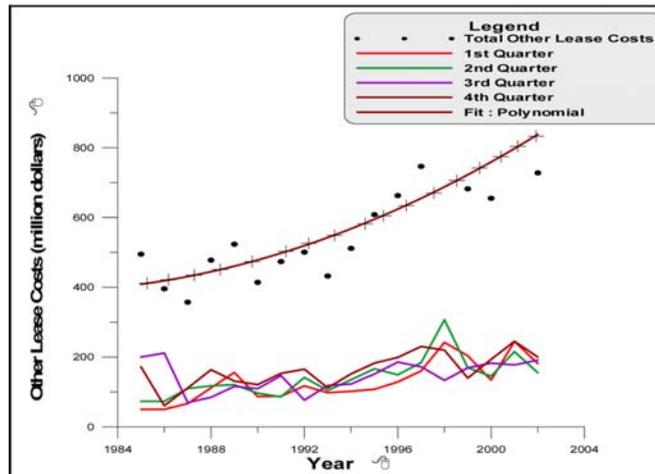
Western Australia’s actual exploration costs (as shown in Figure 17) data have been fitted well with the parabolic trend, with minor fluctuations (irregular trends) around the computed trend. The correlation coefficient between these costs and period appears to be good.

Figure 17 Construction of parabolic equation for WA mineral exploration cost (see online version for colours)



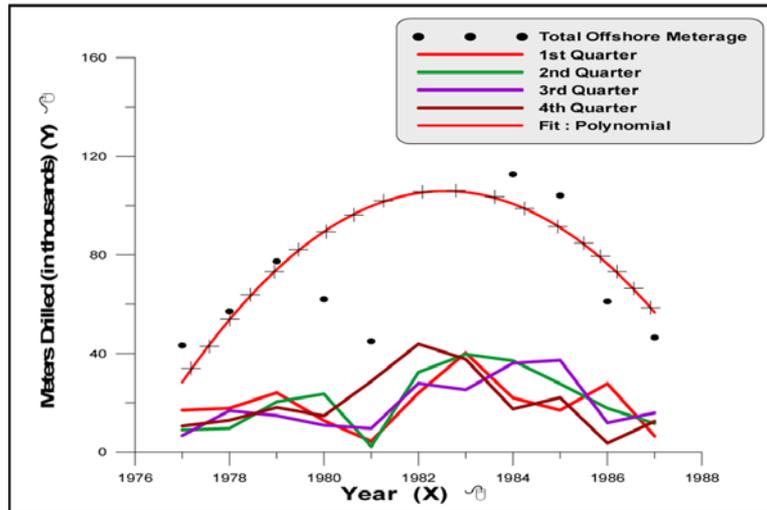
So far, mineral exploration costs data have been interpreted. Exploration of petroleum resources is also active in Australia, especially in the state of WA. Quarterly costs of petroleum lease exploration areas do not match each other as shown in Figure 18. But, there is general increase in the trend (secular variation) of other lease costs with period. Fluctuations in the actual lease costs are irregular, but minor. Coefficient of determinations is fairly good as shown in Figure 18.

Figure 18 Correlation among quarterly data of other petroleum lease costs data (see online version for colours)



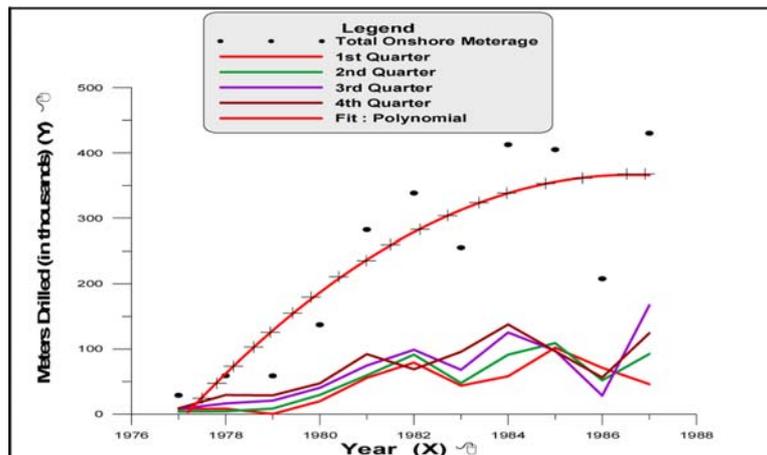
The combined quarterly offshore meterage data, in general, do match data patterns (see Figure 19) with computed trend. But, data are not correlatable to each quarter. Maximum meterage drilled in the year 1983 have two peaks, interpreted one in 1983 and the other in 1979, separated by a trough in 1981. This is another top reversal resources data model pattern (as envisaged in Figure 7).

Figure 19 Analysis of offshore meterage drilled data (see online version for colours)



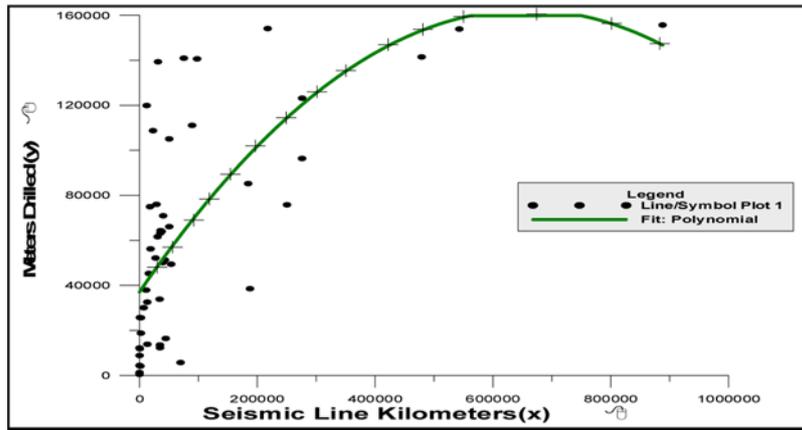
In general, patterns (Figure 20) of quarterly onshore petroleum meterage drilled data responses do match. Maximum meterage drilled is in the year 1984. Peaks are in 1982 and 1984 and troughs are in 1983 and 1986. In general, there is periodic increase in the data. This is another panicked top reversal data pattern shape as narrated in Figure 7. This explicitly indicates that more wells have been drilled in recent years, with a pursuit of more oil production and meet demand. Global oil pricing and more demand of crude are also contributing factors for this type of data response patterns.

Figure 20 Analysis of onshore meterage drilled data (see online version for colours)



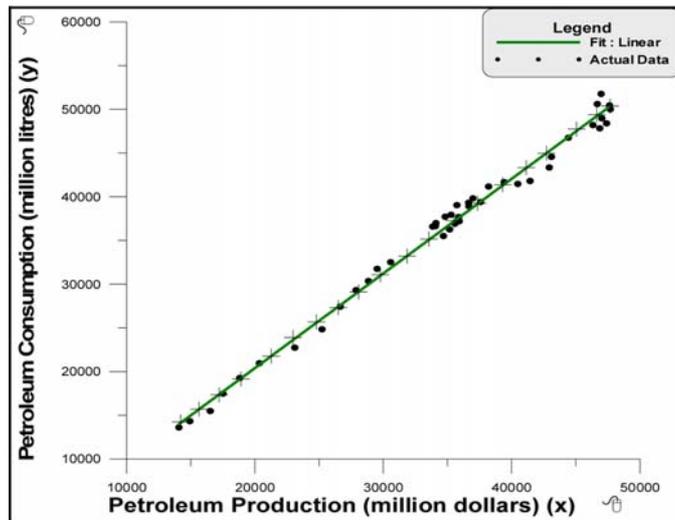
An exponential equation (as drawn in Figure 21) drawn between two attribute variables, seismic (surface) line kilometres and metres drilled (sub-surface), has poor correlation and is not user friendly. But, one is independent of the other variable. But, the data trend indicates an exponential relation between these two variables as illustrated in Figure 21. However, these computed trends may help the seismic field parties to have an advanced fact of the metres to be drilled in an under investigation.

Figure 21 Construction of polynomial equation between seismic line kilometres and meterage drilled data (see online version for colours)



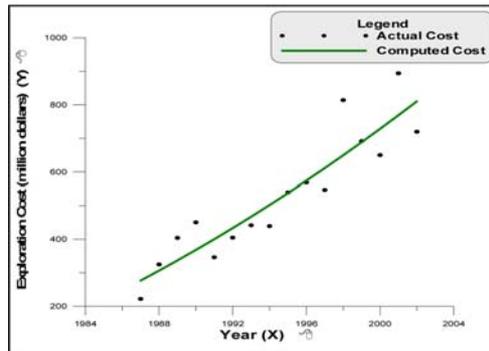
A linear equation (Figure 22) has been established between the petroleum production and petroleum consumption. The actual data perfectly matches with the computed trend. With increase in production, there is corresponding effect on its consumption as demonstrated in Figure 22.

Figure 22 Construction of linear equation between petroleum production and petroleum consumption data (see online version for colours)



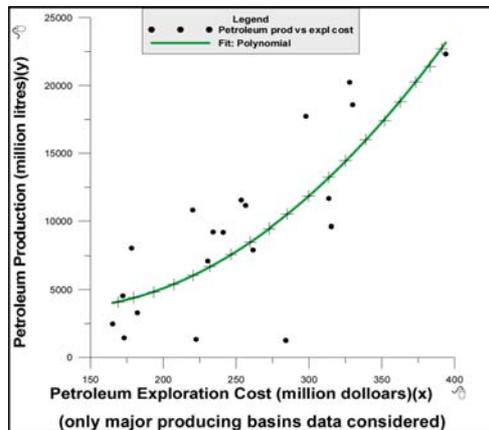
The actual offshore exploration costs data have been fitted with the computed parabolic trend as shown in Figure 23. There is good fit, because of good correlation coefficient. The fluctuations in the offshore exploration costs data are due to seasonal.

Figure 23 Construction of parabolic equation for actual offshore exploration data (see online version for colours)



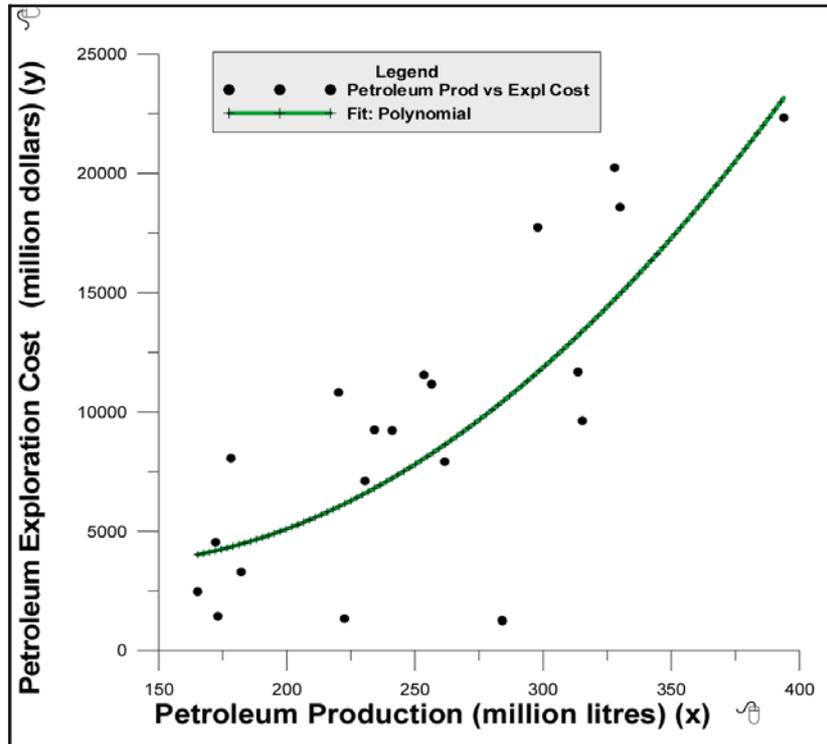
Irregular fluctuations of the actual data around the computed trend may be due to the irregular or random variations with various external factors. They may be political or drop in Australian dollar value or any other natural calamities during these periods. Petroleum actual exploration cost and petroleum production have been plotted to explore for any correlation (see Figure 24). Original exploration costs data are not user friendly. A polynomial equation constructed between these two dependent (if they are linearly proportional) variables, fairly match with the actual data as illustrated in Figure 24. One has to be careful in using this computed trend, though coefficient of determination is fairly good. The actual data appears to have irregular trends. These irregular variations in the actual exploration costs data could be due to unforeseen and external situations. But, the computed trend appears to follow the actual data trend and so one can make use of this model to predict the future petroleum production, having known the petroleum exploration costs.

Figure 24 Construction of polynomial equation between petroleum exploration cost and petroleum production data (see online version for colours)



A trend analysis has been made between petroleum production and petroleum export attributes (see Figure 25). The constructed computed trend matches with the actual data though there are irregular fluctuations. The constructed curve can provide future predictions of petroleum production and exports.

Figure 25 Construction of polynomial equation between petroleum production and petroleum exports data (see online version for colours)



Regression analysis done between two dependent variables, such as the number of surveys conducted and the number of hydrocarbon-producing wells, provides a poor correlation coefficient ($r = 0.47$, see Figure 26). Regression analysis done between number of surveys and number of structures provided a fair ($r = 0.655$, see Figure 27) correlation. The number of surveys vs. number of wells drilled has been used for regression analysis and a fair ($r = 0.68$, see Figure 30(a)) correlation has been established. These equations (X on Y and Y on X) may be used to predict the two variables. Figure 30(b) shows relationship and construction of a trend between *number of surveys* and *number of wells drilled* attributes. Similar analysis done between wells drilled and number of hydrocarbon-producing wells provides a good correlation coefficient ($r = 0.878$, Figure 28). Regression analysis carried out between number of structures interpreted and number of hydrocarbon-producing wells has provided a strong correlation coefficient ($r = 0.932$, Figure 29). Regression equations may be used for future prediction of the attributes.

Figure 26 Regression analysis between *number of surveys* and *number of hydrocarbon-producing wells* data (see online version for colours)

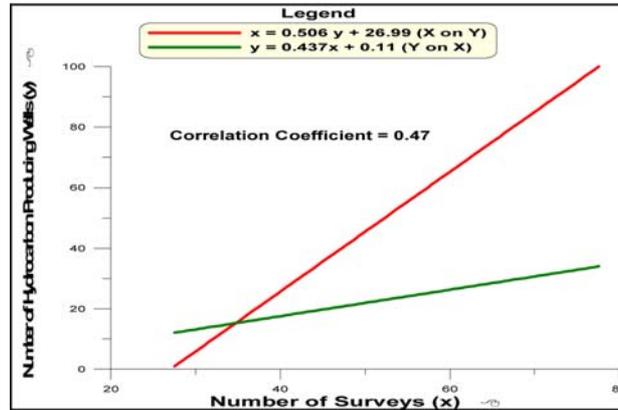


Figure 27 Regression analysis between *number of surveys* and *number of structures* interpreted data (see online version for colours)

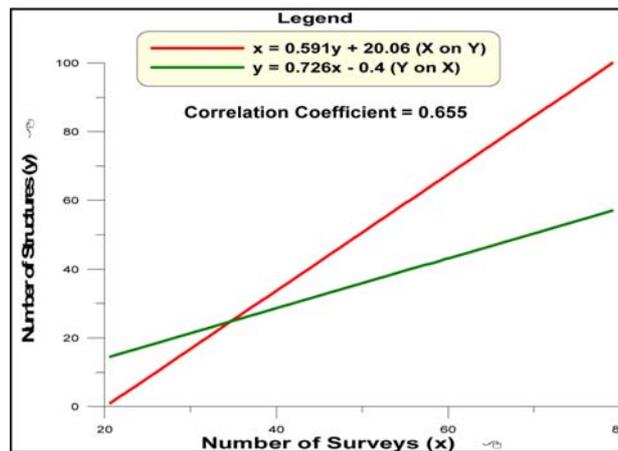


Figure 28 Regression analysis between *number of wells drilled* and *number of hydrocarbon-producing wells* data (see online version for colours)

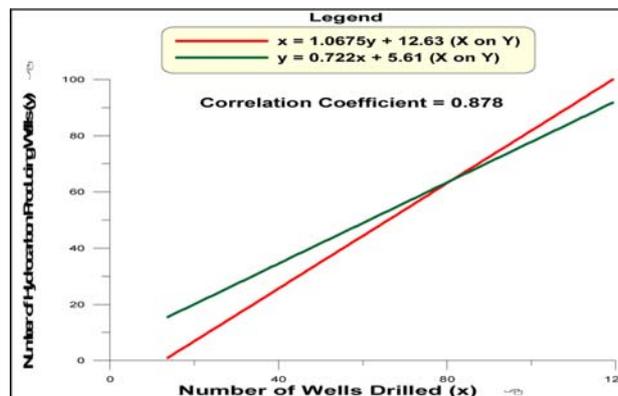


Figure 29 Regression analysis between *number of structures* and *number of hydrocarbon-producing wells* data (see online version for colours)

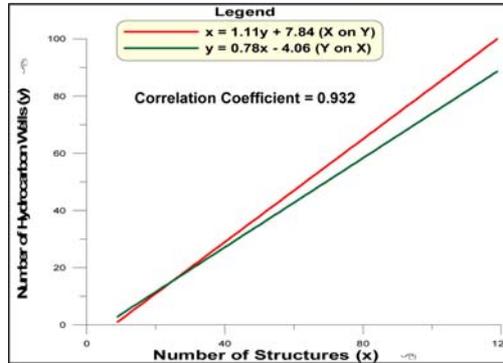
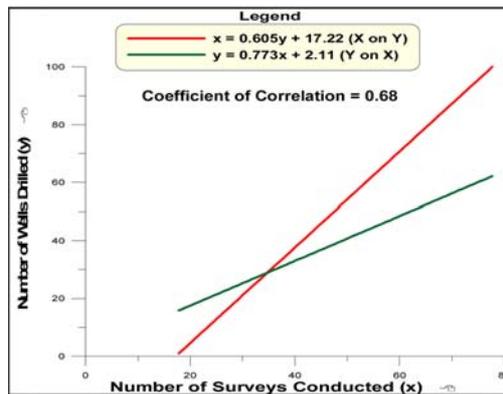
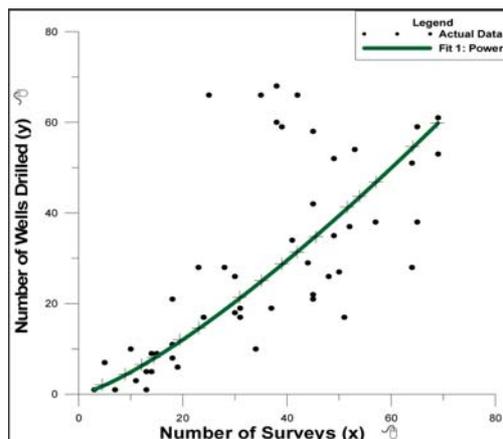


Figure 30 (a) Regression analysis between *number of surveys* and *number of wells drilled*; (b) polynomial fit between *Number of Surveys* and *Number of Wells Drilled* attributes and (c) polynomial fit between *Number of Surveys* and *Number of structures* attributes (see online version for colours)

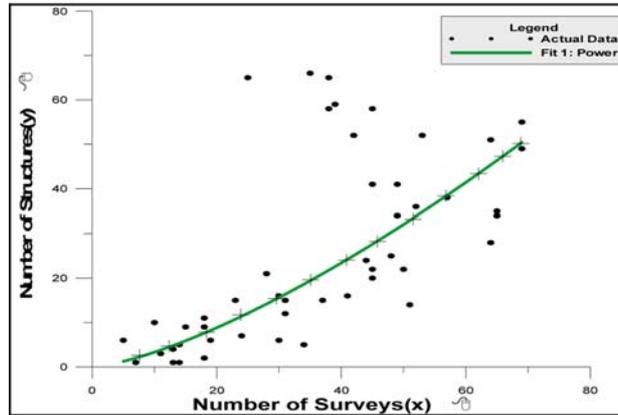


(a)



(b)

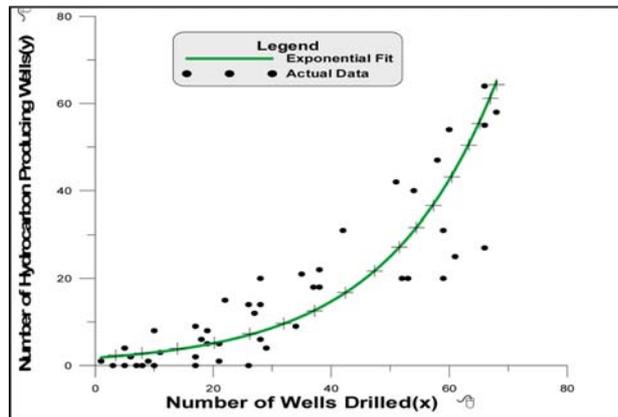
Figure 30 (a) Regression analysis between *number of surveys* and *number of wells drilled*; (b) polynomial fit between *Number of Surveys* and *Number of Wells Drilled* attributes and (c) polynomial fit between *Number of Surveys* and *Number of structures* attributes (see online version for colours) (continued)



(c)

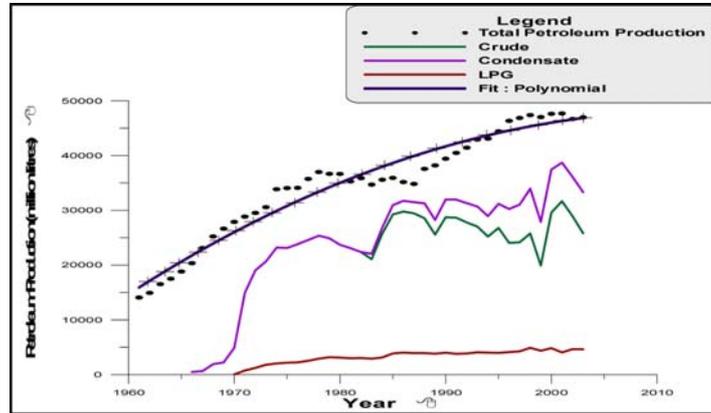
A correlation has been searched (as shown in Figure 31) between the number of wells drilled and the number of hydrocarbon-producing wells (attribute instances). One variable depends on the other. The computer trend is very gentle initially and much steeper with increase in the number of wells drilled. This suggests that with increasing number of wells drilled, there is much more chance of getting more number of hydrocarbon wells (with more petroleum production).

Figure 31 Construction of exponential equation between *number of wells drilled* and *number of hydrocarbon-producing wells* data (see online version for colours)



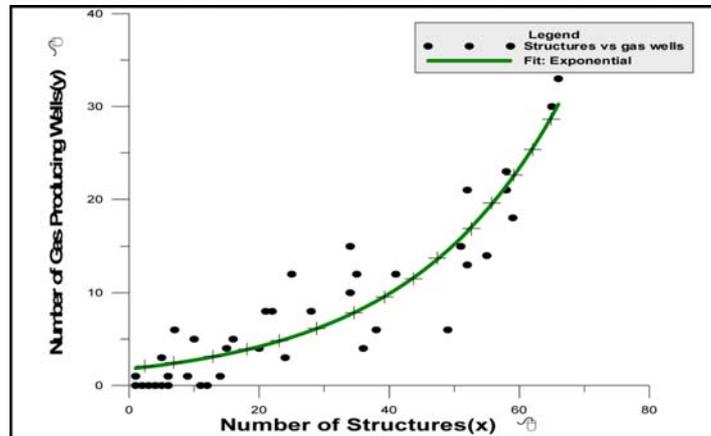
Similar trends in the data of petroleum products such as crude, LPG, condensate and a general periodic increase in petroleum products with regular fluctuations in crude and condensate after 1985 as demonstrated in Figure 32 are due to seasonal. Coefficients of determinations are strong as shown in Figure 32. This panicked shaped data patterns (Figure 7) indicate more demand for crude and strongly suggests global oil pricing.

Figure 32 Correlation analysis of Australian petroleum products data (see online version for colours)



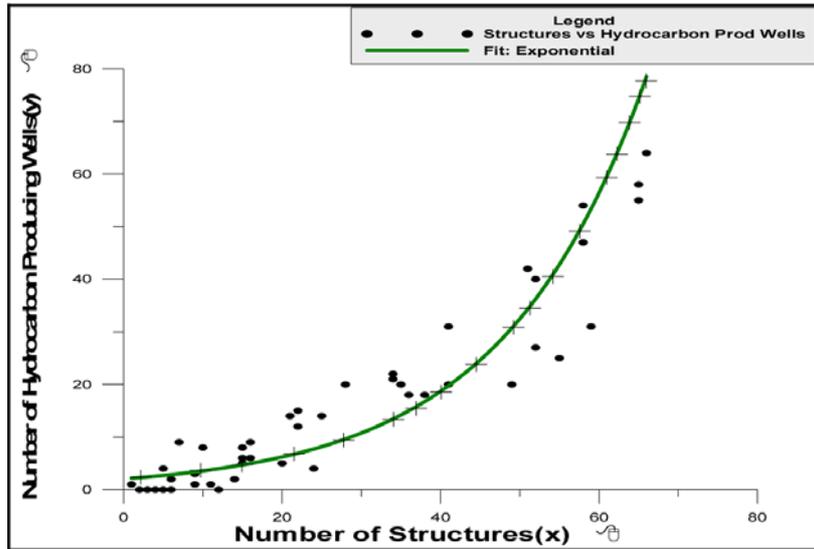
The number of structures interpreted vs. the number of gas-producing wells (attribute instances) has been correlated (Figure 33). The computer trend interprets a very gentle rise initially and the trend rises sharply at higher number of structures. This implies that with more number of petroleum structures interpreted, there is an increase in the number of gas-producing wells (see Figure 33 for details) in Australia.

Figure 33 Construction of exponential equation between *number of structures* and *number of gas-producing wells* data (see online version for colours)

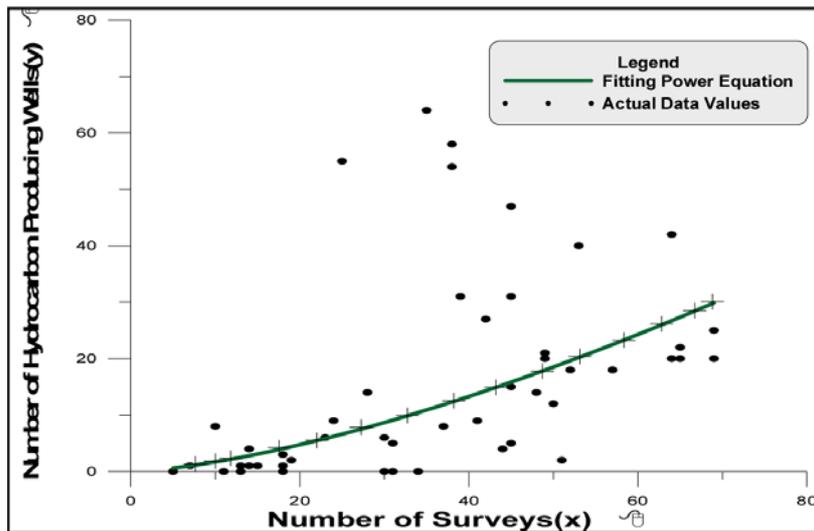


Attribute instances of *the number of structures* (and surveys) interpreted vs. *the number of hydrocarbon-producing wells* have been correlated (Figure 34). The computer trend interprets a very gentle rise initially slowly and the trend rises steeply at higher number of structures. This implies that with more number of petroleum structures interpreted, there is an increase in the number of hydrocarbon-producing wells. In case of the number of surveys vs. number of producing wells, there is noise present in the data. However, with increase in the number of surveys, there is corresponding increase in the number of hydrocarbon-producing wells.

Figure 34 (a) Construction of exponential equation between *number of structures* and *number of hydrocarbon-producing wells* data and (b) construction of power equation between *number of surveys* and *number of hydrocarbon-producing wells* attributes (see online version for colours)

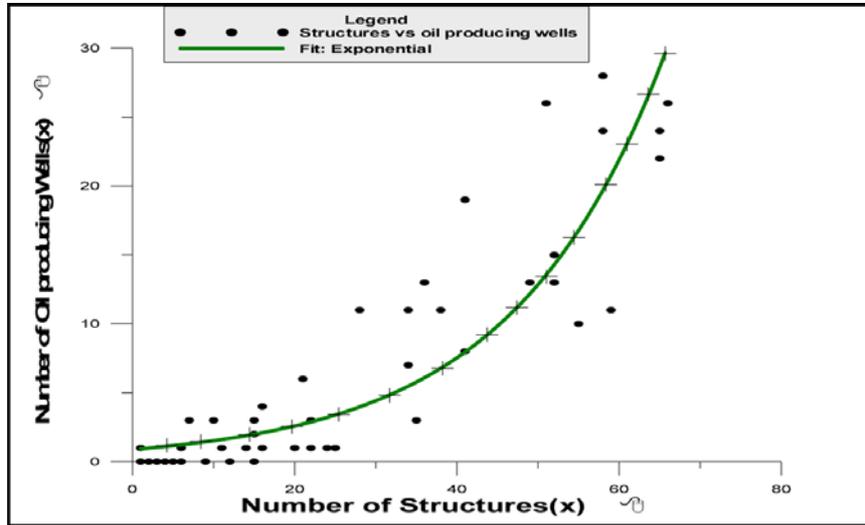


(a)



Attribute instances such as the number of petroleum structures are again correlated with the number of oil-producing wells as shown in Figure 35. In this case, the computer trend is much sharper even for smaller number of structures, which indicates that in spite of less number of petroleum structures interpreted, with increasing number of oil-producing wells.

Figure 35 Construction of exponential equation between number of structures and number of oil-producing wells data (see online version for colours)



The number of wells and the number of condensate wells plotted has provided a good correlation in the years 1971–1972 and 1993–2000, but fair correlation in between years 1985–1990 (Figure 36). There is quite good correlation drawn between the number of wells drilled and number of gas-producing wells in the years 1963–1970, 1971–1990 and 1991–2000 (Figure 37).

Figure 36 Correlation analysis between number of wells drilled and condensate-producing wells data (see online version for colours)

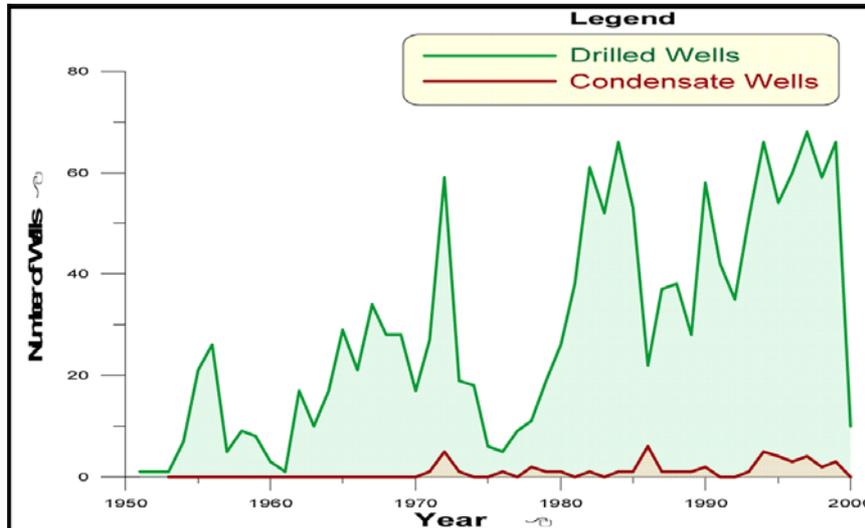


Figure 37 Correlation analysis between number of wells drilled and number of gas-producing wells (see online version for colours)

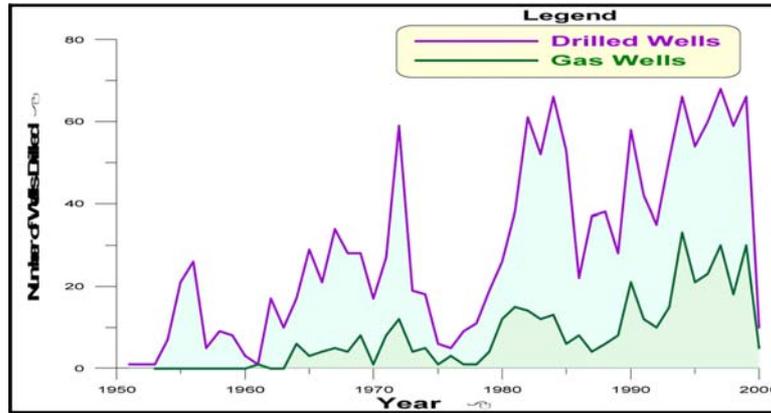
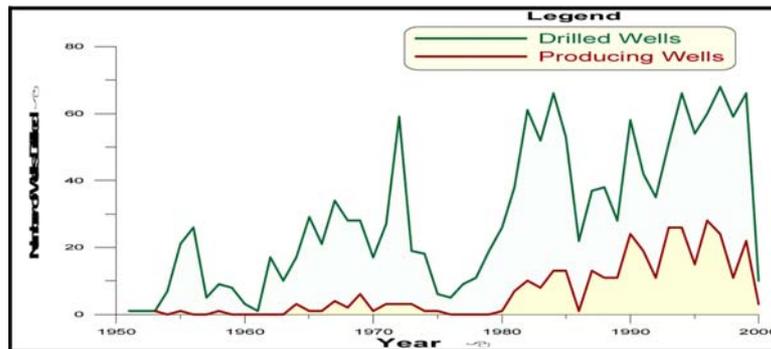
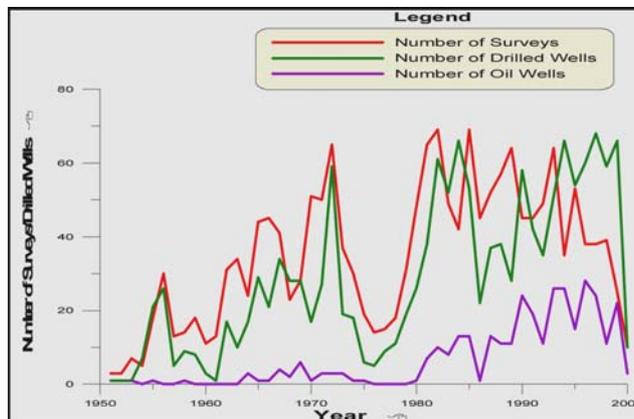


Figure 38 Correlation analysis between wells drilled and oil-producing wells (see online version for colours)



The number of drilled wells vs. the number of oil-producing wells has been correlated. These two variable attributes have dependence on each other and has quite good correlation as illustrated in Figure 38. In general, the number of surveys conducted and wells drilled in Australia do match, except in the years 1972 and 1997. Similar good correlation is observed between the number of surveys, number of wells drilled and number of oil-producing wells particularly in the years 1955, 1970 and 1980–2000 as narrated in Figure 39. As narrated in Figure 39, the number of wells drilled matches with the number of surveys conducted. In between years 1985 and 1990, less number of wells drilled in spite of increase in the number of surveys, in which period, number of oil-producing wells has also gone up. On the contrary, as demonstrated in Figure 39, in between 1995 and 2000, more number of wells drilled is drilled with similar increase in oil-producing wells. In the same period, surprisingly, oil business was in bad shape in the sense; the value of barrel went drastically down during 1997. Though economic slowdown is perceived during this period, oil and gas industry has been still active.

Figure 39 Trend analyses among number of surveys, number of wells drilled and producing wells for different period attributes (see online version for colours)



Data warehousing and mining methods facilitate so far the interpretation of predictions done during resources business operations and their exploration costs controls. There are potential opportunities and scope for further studies, especially in the context of development of data-mining procedures.

9 Conclusions and recommendations

Moving average, weighted moving average and exponential smoothing methods provide good resolution between responses of resources data attributes, but these computed responses do not match with the actual periodic resources data. Time series patterns of the resources data fairly match with the patterns computed by the dynamic programming approach. Polynomial, parabolic and exponential equations have been constructed between actual exploration costs and mineral discoveries made in different states of Australia. These responses have been computed for quarterly and annually. The matching between these data is fair to good. Similar equations have been constructed between onshore and offshore petroleum exploration costs and production, consumption and exports of petroleum products in Australia. A fair good match is observed between these variable attributes of resources databases indicating a strong correlation between data items. There is linear relation observed between petroleum production and consumption. Power equations have been constructed among various performance indicators of data items of the resources industry deriving almost linear relationship among these variables suggesting strong correlation and dependence between different attributes.

Regression analysis has been carried out between dependent data items of the resources data again signifying with good correlation coefficient values. Most of these data items are friendly in nature. Exponential equations have been constructed among performance indicators of the resources data revealing a fair match between actual data and computed data. In general, with the increase in exploration activity in the oil-bearing basins of the Australia, there is increase in petroleum production. There are periodic random and cyclic fluctuations in the actual data at periods, due to the inflation, change in the global petroleum supply and demand and fluctuations in the dollar value.

Rounding top reversals of data responses infer that with increase in period, the exploration scenario drastically changes, for example, decreasing is exploration costs and resources production even with increased exploration costs. In another case, with increase in petroleum production, there is decrease in petroleum products exports initially, but, subsequently, there is steep rise in exports later. It is interesting to observe that with the increase in seismic line kilometres (surface data coverage), there is steady increase in the sub-surface metres drilled (sub-surface data coverage). It appears that there is a fall in the metres drilled even with increase in surface line kilometres. Statistical models computed in this paper are useful and good guidance for managers involved in the resources exploration and production. Some of the actual data presented contain noise, however a good statistical trend estimated helps in understanding the attribute variation. All data-mining approaches may be tried in understanding the correlations, trends and patterns in the resources data. Analysis of the errors in correlation coefficients and the actual correlation coefficients computed for each pair of attribute values provide interesting trends and a strong positive correlation among resources data attributes.

Acknowledgements

The authors wish to thank Jeff Haworth, Manager, Data Management Group, Western Australian Department of Industry and Resources, for providing necessary data used in the present studies. The authors also acknowledge Amit Rudra, School of Information Systems, for useful discussions, and Head of the School of Information Systems, Curtin Business School, Curtin University, for necessary facilities for carrying out this research work at Curtin University computer laboratories.

References

- Aczel, D.A. (1993) *Complete Business Statistics*, 3rd ed., McGraw-Hill, New York, pp.1–869.
- Berenson, M.L. and Levine, D.M. (1992) *Basic Business Statistics, Concepts and Applications*, 6th ed., Prentice Hall, New Jersey, USA, pp.1–953.
- Bhatt, M., Flahive, A., Wouters, C., Rahayu, J.W., Taniar, D. and Dillon, T.S. (2004) ‘A distributed approach to sub-ontology extraction’, *AINA*, Vol. 1, pp.636–641.
- Dunham, H.M. (2003) *Data Mining, Introductory and Advanced Topics*, Prentice Hall Publications, New Jersey, USA, pp.10–200.
- Fayyad, U.M., Shapiro, G.P., Smyth, P. and Urthrusamy, R. (1996) *Advances in Knowledge Discovery and Data Mining*, MIT Press, Cambridge, MA, pp.229–247.
- Flahive, A., Rahayu, J.W., Taniar, D. and Apduhan, B.O. (2004) *A Distributed Ontology Framework for the Grid*, PDCAT: 3320, ISBN: 3-540-24013-6, pp.68–71.
- Gornic, D. (2000) *Data Modelling for Data Warehouses*, Rational Software White Paper, www.rational.com/worldwide
- Graham, P. and Desmond, J.K. (1992) ‘The use and misuse of statistical methods in information systems research’, *Information Systems Research*, No. 3, pp.208–229.
- Gregersen, H. and Jensen, C.S. (2002) *Conceptual Modeling of Time-Varying Information*, <http://powerdb.net/database>
- Gupta, S.P. (1990) *Practical Statistics*, M/S Chan & Co Publishers, New Delhi, pp.1–563.
- Hair, F.J., Anderson, R.E. and Tatham, R.L. (1984) *Multivariate Data Analysis*, 2nd ed., Maxwell Macmillan Publishers, New York, pp.1–449.

- Hoffer, J.A., Presscot, M.B. and McFadden, F.R. (2005) *Modern Database Management*, 7th ed., Prentice-Hall, Upper Saddle River, New Jersey, USA, ISBN: 0-13-145320-3.
- Mattison, R. (1996) *Data Warehousing Strategies, Technologies and Techniques*, Mc-Graw Hill Publishers, New York, USA, pp.100–450.
- Nimmagadda, S.L. and Dreher, H. (2007a) 'Design of petroleum company's metadata and an effective knowledge mapping methodology', *Proceedings of IASTED Conference – Intelligent Systems and Control*, 19–21 November, Cambridge, Massachusetts, USA, http://www.actapress.com/Content_of_Proceeding.aspx?proceedingID=466
- Nimmagadda, S.L. and Dreher, H. (2007b) 'Ontology based data warehouse modeling and mining of earthquake data: prediction analysis along Eurasian-Australian continental plates', *Proceedings of INDIN 2007*, 23 July, Vienna, Austria, pp.597–602.
- Nimmagadda, S.L. and Dreher, H. (2008a) 'Ontology-based data warehousing and mining approaches in petroleum industries', Chapter XI in Negro, H.O., Cisaró, S.G. and Xodo, D. (Eds.): *Data Mining with Ontologies: Implementation, Findings, and Frameworks*, Information Science Reference, IGI Global, Hershey, PA, USA, <http://www.igi-pub.com/reference/details.asp?ID=6844>
- Nimmagadda, S.L. and Dreher, H. (2008b) 'Ontology based data warehouse modeling – a methodology for managing petroleum field ecosystems', *Proceedings of 2nd IEEE International Conference on Digital Ecosystems and Technologies (IEEE DEST 2008)*, 26 February, Phitsanulok, Thailand, pp.221–228.
- Nimmagadda, S.L. and Rudra, A. (2004) 'Applicability of data warehousing and data mining technologies in the Australian resources industry', *Proceedings of 7th International Conference on Information Technology*, 20–23 December, Hyderabad, India, pp.1–10.
- Nimmagadda, S.L., Dreher, H. and Rudra, A. (2005) 'Ontology of Western Australian petroleum exploration data for effective data warehouse design and data mining', *Proceedings of INDIN 2005: 3rd International Conference on Industrial Informatics, Frontier Technologies for the Future of Industry and Business*, 10 September, Perth, Western Australia, pp.611–616.
- Nimmagadda, S.L., Dreher, H., Chang, E. and Rajab, M.R. (2006) 'New technologies in mature gulf basins – multidimensional modeling of ontologically derived historical petroleum exploration data properties for effective basin knowledge mapping', A poster paper presented and published in the *AAPG International Conference and Exhibition*, 5–8 November '06, Perth, Australia.
- Pujari, A.K. (2002) *Data Mining Techniques*, University Press (India) Pty Limited, Hyderabad, pp.7–67.
- Roiger, J.R. and Geatz, M.W. (2003) *Data Mining a Tutorial Based Premier*, Addison Wesley, pp.200–350.
- Rudra, A. and Nimmagadda, S.L. (2005) 'Roles of multidimensionality and granularity in data mining of warehoused Australian resources data', *Proceedings of 38th Hawaii International Conference on Information Sciences*, Hawaii, USA, ISBN: 0-7695-2268-8, pp.216–223.
- Rusu, L.I., Rahayu, J.W. and Taniar, D. (2005) 'A methodology for building XML data warehouses', *International Journal of Data Warehousing and Mining IJDWM*, Vol. 1, No. 2, pp.23–48.
- Shanks, G., Tansley, E. and Weber, R. (2003) 'Using ontology to validate conceptual models', *Communications of the ACM*, Vol. 46, No. 10, pp.85–89.
- Taniar, D., Rahayu, J.W., Lee, V. and Daly, O. (2008) 'Exception rules in association rule mining', *Applied Mathematics and Computation*, Vol. 205, No. 2, Elsevier Publishers, pp.735–750.
- Tjioe, H.C. and Taniar, D. (2005) 'Mining association rules in data warehouses', *International Journal of Data Warehousing and Mining, IJDWM*, Vol. 1, No. 3, pp.28–62.