Classification of geochemical analyses using Artificial Neural Network Technology and Discriminant Analysis

A Mining Industry Application

Classifying multi-element geochemical data using SPSS Neural Connection™ and SPSS DISCRIMINANT

™ SPSS Neural Connection is a trademark of SPSS Inc.

Introduction

This paper compares the use of two statistical techniuqes- one "new" and one "old" - in classifying exploration stream sediment samples as being **mineralised or barren** on the basis of their multi-element analytical content. The "new" statistical technique is called Neural Network computing and the "old" statistical technique is called Discriminant Analysis.

Neural Networks

Neural network computing is an empirical modelling technique which is beginning to be used in a variety of applications (such as database marketing and process control) to complement and in some cases supplant, more traditional statistical methods (Furness,1992)[i]. As stated by Furness (1992)[ii] although the technique has been around since the 1940s it has only relatively recently begun to make an impact in commercial applications, fuelled by the falling costs of computing power and advances in the theory of neural networks.

Neural networks were one of the successes of the artificial intelligence work in the 1980s. Now , despite their limitations, businesses have begun to use them . They are built from webs of randomly connected electronic neurons and in design and function, closely resemble the human brain (Bournellis, 1996)[iii]. A brain is a wonderful thing to fake. It is made up of lots of switches (neurons) linked together by a lot of wires (synapses) - just like a machine.

A neural network is based on a modelling technique that observes the behaviour of biological neurons and uses the data to mimic the performance of a system. For example, if the data consists of the daily temperatures of say, Sydney over a two week period, the neural network will emerge with a simple curve that describes the way temperature rises in summer and falls in winter. This is achieved by varying the strength of connections (weights) between individual processors until the input yields the right output.

The features of neural network technology are that it improves its performance of a particular task by trail and error. It can also be a 'black box'. That is, the user doesn't need to know what mathematical equation describes its output (Bournellis, 1996)[iv].

A drawback of neural computing is the cryptic nature of the models. Neural networks have complex structures with large numbers of parameters
which do not lend themselves to intuitive interpretation. Indeed software such as *SPSS Neural Connection* can output information about the overall model structure, estimates of dependent variables and information about the fit of the model to the sample data, but not the individual model parameters. This 'black box' nature makes for difficulty in understanding neural network models (SPSS Inc. 1996)[v]

Discriminant Analysis

Little needs to be said about this well established traditional statistical technique (first introduced by Sir Ronald Fisher) which has been popular amongst geochemists for over two decades. Howarth (1971)[vi] states that with this method one wishes to formulate mathematical criteria which can be used to distinguish members of one class (e.g. sediments from streams draining a mineralised locality). from those of another (e.g. sediment from streams draining an unmineralised locality). These criteria could then be used to give an optimum classification of samples whose origin is unknown.

The procedure is:

1. Define the classes in which we are interested.
2. Select a statistically representative set of samples from each category to form the training set for designing the classifier (formulating the discriminant function).
3. Select those features (elements) which best distinguish between the chosen classes on the basis of the training set of samples. This selection may be carried out either empirically or by use of statistical methods to evaluate the effectiveness of a given feature set.

In order to evaluate the performance of the system correctly, a second testing set of data, consisting of samples whose class is known but which were not included in the training set, is presented to the classifier following the training phase. In time series analysis this same technique is called using a *hold out sample*.

As stated by Howarth (19971) [vii] although it is extremely tempting to put all the stream sediment samples for which one knows the answers (e.g. mineralised or unmineralised) into the training set, the results can be misleading.

A simple linear discriminant function (Davis, 1986) [viii] transforms an original set of measurements on a sample into a single discriminant score. That score, or transformed variable represents the sample's position along a line defined by the linear discriminant function. We can therefore think of the discriminant function as a way of collapsing a multivariate problem down into a problem which involves only one variable.

For the linear discriminant function to be "optimal", that is, to provide a classification rule that minimises the probability of misclassification, certain assumptions about the data must be met. Each group must be a sample from a multivariate normal population, and the population covariance matrices must all be equal.

Descriptive statistics and univariate tests of significance provide basic information about the distributions of the variables in the groups and help to identify differences among the groups. However, in discriminant analysis the emphasis is on analysing the variables *together*, not one at a time.

The Case Study

Example From Davis (1986) [ix]

The following problem is an example of the use of discriminant analysis. By using discriminant first we will have a base from which to make comparisons with neural network data analysis which will follow. Its solution may help gain a better understanding of the procedure.

A government survey group in northern Sweden is prospecting for **base metal deposits** in densley forested mountains. Airborne magnetometer surveys have proved to be of limited value, so a geochemical prospecting approach is being evaluated, based on stream-water analyses. Seven variables have been selected and two suites of measurements performed. Group 1 consists of measurements on streams draining areas with active mines or proven mineral deposits. Group 2 consists of similar measurments on streams draining areas that have been heavily prospected without results. A data listing is appended.

From these data, we will calculate the disriminant function between productive and non-productive regions. We will also determine if the difference between the two groups is significant, and investigate the relative importance of the variables used. For the purposes of this exercise , we will assume that the parent populations of the two groups are multivariate normal. The data listing also contains a set of measurments made on streams not known to have been prospected. On the basis of the discriminant function, **can any be selected as likely areas for prospecting ?**

Data Management

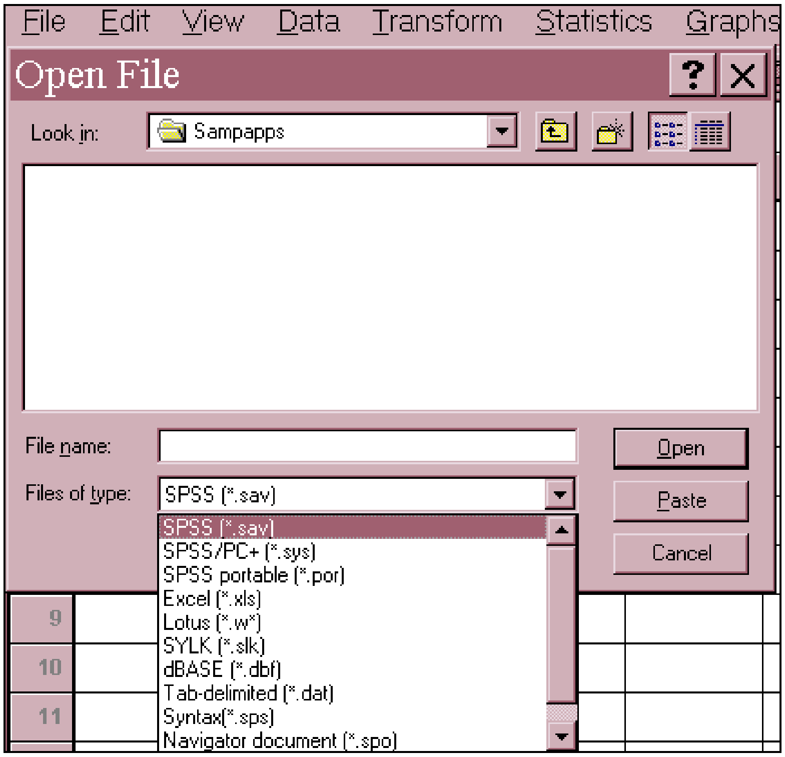
1. Opening the SPSS data file sweden.sav

We will assume that SPSS 7.0 for Windows has just been launched.
In order to analyse data, you must have data to run. You can use the main menu File command to open an existing SPSS file.

Click **File** in the main menu

Click **Open** in the File menu

SPSS displays the Open File dialog box which lets you specify the file name, directory (folder), and type of file you want to open. By default the dialog box displays SPSS files with the extension .SAV from the folder SPSS in the files list box entitled ‘Look In :’. You can change the specifications to read different types of files from different folders.
Our data file happens to be an Excel spreadsheet called *streams.xls* so we need to click on the down arrow button to display the drop-down list conatining the list of available file types that SPSS can import. We have also browsed our way to the directory (folder) which contains streams.xls.



Look in: Sampapps

File name:

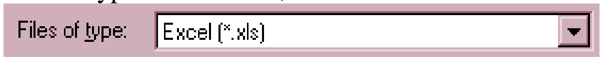
Files of type: SPSS (*.sav)

“Easily the slickest Windows-based statistics package... new users will love how smoothly the Windows dressing works.”

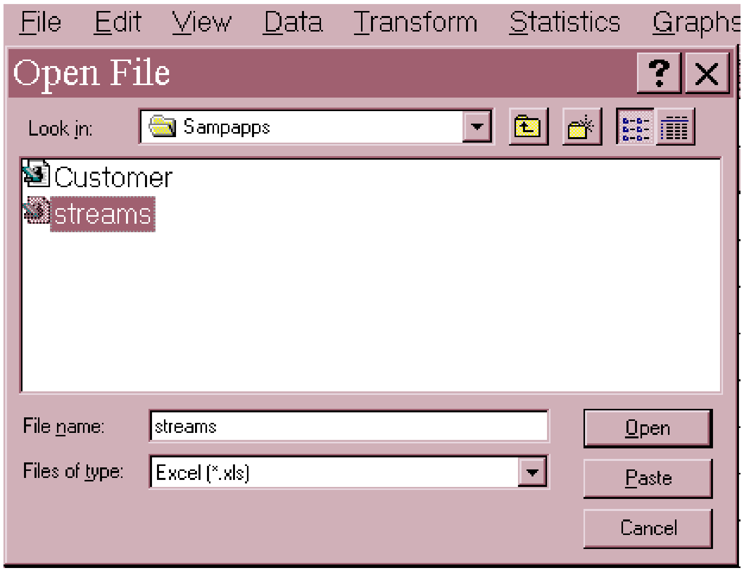
- InfoWorld

This spreadsheet uses the first row to store the field names , such as ‘Cu’, ‘Pb’ etc. We can extract this information from the spresdsheet to use as column headings in the SPSS spreadsheet , by proceeding as follows:

Select ‘streams’ from the list .The file extension is not shown because we know that the file type is a .xls file , as shown in the text box at the base of the dialog



Files of type: Excel (*.xls)



Look in: Sampapps

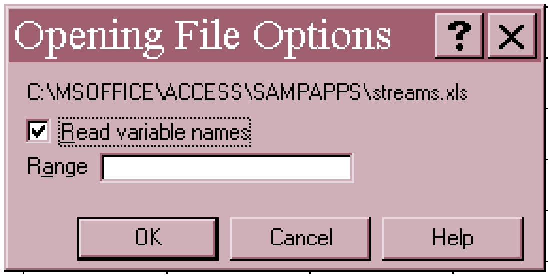
Customer
streams

File name: streams

Files of type: Excel (*.xls)

Double click *streams* in the ‘Look in:’ list box <u>or</u>
Click once on *streams* in the ‘Look in:’ list box and
Click the Open command button.

SPSS knows that you are opening an Excel spreadsheet so it asks you whether you want to use the information in row 1 of the Excel spreadsheet for field names when the data goes across to the SPSS spreadsheet.

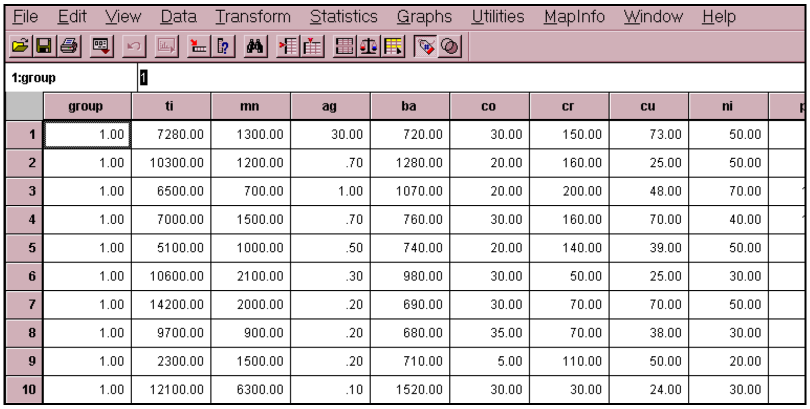


C:\MSOFFICE\ACCESS\SAMPAPPS\streams.xls

☑ Read variable names

Range

Click on OK and SPSS puts the contents of the file in the Data Editor window, as shown below.



1:group 1

	group	ti	mn	ag	ba	co	cr	cu	ni
1	1.00	7280.00	1300.00	30.00	720.00	30.00	150.00	73.00	50.00
2	1.00	10300.00	1200.00	.70	1280.00	20.00	160.00	25.00	50.00
3	1.00	6500.00	700.00	1.00	1070.00	20.00	200.00	48.00	70.00
4	1.00	7000.00	1500.00	.70	760.00	30.00	160.00	70.00	40.00
5	1.00	5100.00	1000.00	.50	740.00	20.00	140.00	39.00	50.00
6	1.00	10600.00	2100.00	.30	980.00	30.00	50.00	25.00	30.00
7	1.00	14200.00	2000.00	.20	690.00	30.00	70.00	70.00	50.00
8	1.00	9700.00	900.00	.20	680.00	35.00	70.00	38.00	30.00
9	1.00	2300.00	1500.00	.20	710.00	5.00	110.00	50.00	20.00
10	1.00	12100.00	6300.00	.10	1520.00	30.00	30.00	24.00	30.00

You can make changes to the individual cell entries in the file by highlighting a cell and changing a data value (just like Excel). You can also add new data to the file in the Data Editor window. You use the mouse or the arrow keys to move around the file to examine the different variables or the contents of particular cells.

We will save this file immediately as an SPSS .sav data file :

Click **File** in the main menu

Click **Save** in the File menu

The ‘Save Data as’ dialog will appear prompting you for a name for the .sav file. Type a name into the text box provided and click on OK.

2. Preparing the data for analysis

There are certain aspects of the data which we must address prior to using SPSS DISCRIMINANT. The first point is that in geochemistry **there is no such thing as an assay of “zero ppm”.** You can never have the total absence of any metal. It may be so low as to record ‘below the level of detection’ but never zero. We need to reset any “zero” value assays for any offending elements using the level of detection information which is summarised in the table below. All units of concentration are parts per million (ppm).

Table showing detection limits for elements used in this case study

Element	Ti	Mn	Ag	Ba	Co	Cr	Cu	Ni	Pb	Sr	V	Zn	Au
Detection limit	10	100	0.1	10	5	10	1	10	10	10	10	10	0.01

“SPSS’

As it turns out the elements gold (Au), cobalt (Co), Nickel (Ni) and lead (Pb) are the only ones with any zero assays. We will reset zero assays for

graphing capability is impressive.”

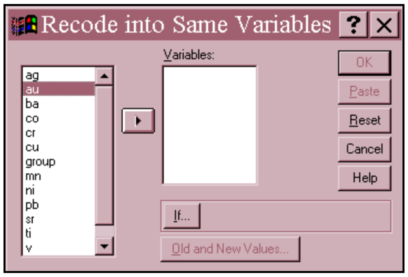
- InfoWorld

gold to a value which is half of the detection limit (i.e. 0.005 is half of 0.01) using SPSS’ recode facility.

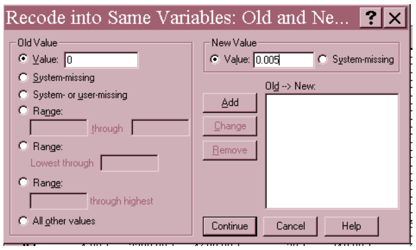
From the menus choose:

Transform
  Recode ▶
    Into Same Variables...

This opens the Recode into Same Variables dialog :



The source variable list contains the elements in our stream sediment data file. We will select gold (Au) for recoding by clicking on the variable name then clicking on the ▶ button to move it across to the Variables box. We then click on the Old and New Values... button .



Old Value: ● Value: 0
New Value: ● Value: 0.005

This opens the next dialog box where we define that we want to change all the gold assays from 0 to 0.005. We simply type 0 into the ‘Old value’ box and 0.005 into the ‘New value’ box , then click on Add Continue and OK to update the SPSS spreadsheet.

This procedure was performed on the other elements prior to running frequency tables on them to confirm that all the zeros had been changed.

3. Selecting Cases for the analysis

The first step in discriminant analysis is to select cases to be included in the computations. Any samples with missing assays for any element are removed from the data prior to analysis. With the case study data all elements have a full complement of assays so this won’t be required.

4. Analysing Group Differences

Group means

GROUP	TI	MN	AG	BA	CO	CR	CU	NI

1	8049.00000	1960.00000	2.25000	814.30000	30.50000	77.00000	36.50000	33.00000
2	3134.50000	700.00000	.17500	115.50000	15.75000	27.00000	69.85000	13.50000
3	4813.33333	1950.00000	.46667	243.33333	17.91667	40.00000	71.50000	25.00000
Tot	5490.21739	1410.86957	1.11522	436.00000	22.44565	50.43478	55.56522	23.47826

GROUP	PB	SR	V	ZN	AU
1	64.00000	146.00000	127.50000	147.00000	.01550
2	13.75000	812.00000	117.00000	79.50000	.01075
3	27.50000	595.00000	101.66667	100.00000	.01167
Total	37.39130	494.13043	119.56522	111.52174	.01293

Group standard deviations

GROUP	TI	MN	AG	BA	CO	CR	CU	NI
1	3142.41258	1243.25549	6.58503	330.53276	15.71958	54.97368	20.32499	15.92747
2	1283.00008	242.79079	.07864	49.14800	4.06364	16.88974	30.82937	7.08965
3	4171.99553	1724.81883	.52409	225.44770	11.44734	27.56810	37.95655	18.43909
Tot	3497.96226	1185.51035	4.40035	408.66069	13.31759	45.31159	31.96710	15.84191

GROUP	PB	SR	V	ZN	AU
1	27.22228	93.49360	90.13878	67.75342	.00826
2	8.56477	386.51752	51.81952	41.60908	.00591
3	31.89828	492.33119	70.82843	56.56854	.00683
total	32.14099	440.26041	72.07886	63.69955	.00735

The tables above contain the means for the 13 independent variables for ‘mineralised’, ‘barren’ , and ‘exploration’ samples (Groups 1, 2 and 3) along with the corresponding standard deviations. From these tables we can see that for all of the elements except copper, strontium and vanadium , mean concentrations are at least double in the ‘mineralised’ catchments compared to the ‘barren’ ones.

```
Variable   Wilks' Lambda        F         Significance
  --------   -------------   -------------   ------------
  TI            .55561        17.1962          .0000
  MN            .71726         8.4750          .0008
  AG            .94726         1.1972          .3119
  BA            .31614        46.5087          .0000
  CO            .70967         8.7958          .0006
  CR            .72128         8.3081          .0009
  CU            .72004         8.3596          .0009
  NI            .66189        10.9829          .0001
  PB            .44230        27.1094          .0000
  SR            .48342        22.9747          .0000
  V             .98583          .3090          .7358
  ZN            .74545         7.3415          .0018
  AU            .90254         2.3218          .1103
```

The table above shows significance tests for the equality of group means for each variable. If the significance level is small (less than 0.05) , the hypothesis that all the group means are equal is rejected. Large values of Wilks’ Lambda indicate that group means do not appear to be different, while small values indicate that group means do appear to be different. From the above table it is clear that **Ti, Ba, Pb, Sr, Ni,Co, Mn, Cr, and Cu are the variables whose means are most different for mineralised and barren stream water samples.**

Since interdependencies among the variables effect most mulitvariate analyses, it is worth examining the **correlation matrix** of the predictor variables. The matrix below is the pooled within-groups correlation matrix.

	TI	MN	AG	BA	CO	CR	CU	NI	PB	SR	V	ZN	AU

TI	1.00000												
MN	.54857	1.00000											
AG	-.02577	-.08977	1.00000										
BA	.52110	.48976	-.07026	1.00000									
CO	.42064	.10609	.06126	-.07486	1.00000								
CR	-.00181	-.33669	.28700	.16162	-.15184	1.00000							
CU	-.06052	-.30137	.19139	-.06344	-.09047	.25378	1.00000						
NI	.33034	-.10750	.21287	.39123	-.03189	.62458	.25842	1.00000					
PB	.17892	.15060	.03153	.18493	-.16819	.52450	.12930	.32446	1.00000				
SR	.04110	-.03078	-.05551	.05695	.00969	-.07984	.01957	-.01971	-.06930	1.00000			
V	.46935	.29118	-.04246	.00869	.64463	-.26813	-.10542	-.05879	-.15253	.07574	1.00000		
ZN	.38292	.25780	.10225	.28704	.08419	.16054	.13806	.14311	.33982	-.09559	-.05917	1.00000	
AU	-.06593	-.07801	.08974	-.17541	.06161	.01034	-.08953	-.24889	.23143	-.04358	.00015	-.14835	1.0
	TI	MN	AG	BA	CO	CR	CU	NI	PB	SR	V	ZN	AU

Vanadium:Cobalt and Nickel:Chromium have the largest correlation coefficients with 0.64 and 0.62 respectively.

A pooled within-groups correlation matrix is obtained by averaging the separate covariance matrices for all groups and then computing the correlation matrix. A total correlation matrix is obtained when all cases are treated as if they are from a single sample.

Descriptive statistics and univariate tests if significance provide basic information about the distributions of the variables in the groups and they help to identify some differences among the groups. **However in discriminant analysis (DA) and other multivariate statistical procedures, the emphasis is on analysing the variables together, not one at a time**. By considering the variables simaltaneously, we are able to incorporate information about their relationships.

In DA , a linear combination of the predictor variables is formed and serves as the basis for assigning our exploration stream water samples to groups. Thus , information contained in mulitple indpendent variables is summarised in a single index.

The linear discriminant equation:
$D=B_0 + B_1X_1 + B_2X_2 + ... + B_pX_p$
is similar to the multiple linear regression equation. The X's are the values of the independent variables (e.g. copper, lead , zinc etc.) and the B's are coeficients estimated from the data. **If a linear disciminínant function is to distinguish stream water samples draining ore from those samples draining barren ground, the two groups must differ in their D values**.

Therefore, the B values are chosen so that the values of the discriminant fucntion differ as much as possible between the groups.

The coefficients for the all the elements are listed overleaf.

```
Unstandardized canonical discriminant function coefficients

              Function 1


TI         3.50275673E-05     small and large values are
MN        -2.88499305E-04     sometimes displayed in scientific
AG              .0276627      notation. For example, the number
BA         3.01985809E-03     0.002252 might be displayed as
CO              .0469828      2.2252E-03.
CR             -.0103137
```

```
CU          -8.06797854E-03
NI                 .0240501
PB                 .0367362
SR          -2.25258445E-03
V            8.14492855E-05
ZN          -6.47220716E-04
AU                6.8552458
(Constant)       -2.3057124
```

Based on these coefficients it is possible to calculate the discriminant score for each stream water sample. Shown below are the element concentrations for the first five stream water samples. The discriminant score for the first water sample is obtained by muliplying the unstandardised coefficents by the values of the variables, summing these products, and adding the constant. For the first sample in our spreadsheet the score is:
D1= 0.000035027(7280) - 0.0002884(1300) + 0.0276(30) + 0.003019(720) + 0.0469(30) - 0.0103(150) - 0.008067(73) + 0.0240(50) + 0.0367 (70) - 0.002252(60) + 0.00008144(70) - 0.0006472(190) + 6.855(0.02) - 2.3057 =

This calculation was performed on the first five samples by using an SPSS COMPUTE statement to type in the above equation.

```
COMPUTE score = 3.50275673E-05 * TI - 2.88499305E-04 * MN +
.0276627 * AG + 3.01985809E-03 * BA + .0469828 * CO -.0103137 *
CR - 8.06797854E-03 * CU + .0240501 * NI + .0367362 * PB -
2.25258445E-03 * SR + 8.14492855E-05 * V - 6.47220716E-04 * ZN +
6.8552458 * AU - 2.3057124 .
FORMAT score(F6.3).
LIST cases = 5 / vars= TI to AU SCORE/
 format=numbered.
```

```
NO.   TI      MN      AG     BA      CO     CR     CU     NI     PB      SR      V      ZN      AU  SCORE

1  7280.00 1300.00  30.00  720.00  30.00 150.00  73.00  50.00  70.00  60.00  70.00 190.00   .02 3.511
2 10300.00 1200.00    .70 1280.00  20.00 160.00  25.00  50.00  70.00  90.00  50.00  50.00   .02 4.362
3  6500.00  700.00   1.00 1070.00  20.00 200.00  48.00  70.00 100.00 210.00  50.00 170.00   .01 4.315
4  7000.00 1500.00    .70  760.00  30.00 160.00  70.00  40.00 110.00 240.00  40.00 250.00   .01 3.388
5  5100.00 1000.00    .50  740.00  20.00 140.00  39.00  50.00  80.00  50.00  60.00 130.00   .02 3.101
```

Summary

Statistical techniques help you better use resources and give credibility to your ideas

Statistical software such as SPSS is the perfect complement to your spreadsheet. Spreadsheets are great for everyday tasks, such as tracking budget numbers and creating simple summary reports and graphs. However, there are times when you need more information from your data and you need to perform in-depth analysis. At these times, you need SPSS. Since SPSS was designed for in-depth analysis, **you get better information from your data.**

SPSS is the right choice to take your analysis to the next level. SPSS connects to your data regardless of where or how it is stored. SPSS can **uncover hidden patterns** and trends that rarely emerge using spreadsheet row-and-column maths. SPSS gives you great looking graphs and reports so you can easily communicate the results of your analysis. Together, SPSS and your spreadsheet can take your business data and translate in into meaningful information so you can make fully informed decisions - and make that **next oil discovery !**

Acknowledgements

The example data and case study discussed in this paper come from the first and second editions (1973,1986) of *'Statistics and Data Analysis In Geology'* written by Dr. John Davis of the Kansas Geological Survey in the USA.

Both John Davis and Garry Edser are members of the *International Association for Mathematical Geology*, which is dedicated to development and application of quantitative methods in the Earth Sciences. The IAMG is affiliated with the International Union of Geological Sciences (IUGS) and the International Statistical Institute (ISI). At the end of 1995 the IAMG had 555 members worldwide , of whom 30 come from Australia. Garry Edser has been a member since 1989. Garry has also been a member of the Australasian Institute of Mining and Metallurgy since 1978.

[i] **Furness,P.** (1992) 'Applying Neural Networks in Database Marketing: An Overview', *Journal of Targeting, Measurement and Analysis for Marketing*, August 1992, pages 152 - 168.

[ii] **Furness,P.** (1992) 'Applying Neural Networks in Database Marketing: An Overview', *Journal of Targeting, Measurement and Analysis for Marketing*, August 1992, pages 152 - 168.

[iii] **Bournellis,C.** (1996) 'Technology boost' in Feature on Artficial Intelligence, *Australian Personal Computer*, Volume 17, Number 1, page 116, published by Australian Consolidated Press

[iv] **Bournellis,C.** (1996) 'Technology boost' in Feature on Artficial Intelligence, *Australian Personal Computer*, Volume 17, Number 1, page 116, published by Australian Consolidated Press

[v] **SPSS Inc.** (1996), unpublished workshop manual entitled 'SPSS Neural Connection Workshop'

[vi] **Howarth, R.J.** (1971) 'Empirical discriminant classification of regional stream sediment geochemistry in Devon and east Cornwall', Trans. IMM, pages B142 to 149.

[vii] **Howarth, R.J.** (1971) 'Empirical discriminant classification of regional stream sediment geochemistry in Devon and east Cornwall', Trans. IMM, pages B142 to 149.

[viii] **Davis, J.C.** (1986) *'Statistics and Data Analysis in Geology'*, 2 nd Edition,pub. John Wiley and Sons, page 478.

[ix] **Davis, J.C.** (1986) *'Statistics and Data Analysis in Geology'*, 2 nd Edition,pub. John Wiley and Sons, page 490.