

# Statistical Analysis of Downhole Physical Property Measurements: Classification and Predictive Analysis

Susanne MacMahon<sup>1</sup>, Senior Geoscientist, Earthfx Inc., Toronto, Ontario

Gary Hodgkinson<sup>2</sup>, Senior Geophysicist, Debeers Exploration Canada, Toronto, Ontario

Dirk Kassenaar<sup>3</sup>, Director of Application Development, VIEWLOG Systems, Toronto, Ontario

Bill Morris<sup>4</sup>, Professor, McMaster University, Hamilton, Ontario

*MacMahon S. E., Hodgkinson G., Kassenaar D.J., Morris W.A., Statistical Analysis of Downhole Physical Property Measurements: Classification and Predictive Analysis, in Proceedings of the 8<sup>th</sup> International KEGS/MGLS Symposium on Logging for Minerals and Geotechnical Applications, Toronto 21-23 August 2002*

## Abstract

One of the challenges of geophysical interpretation, whether purely descriptive (Exploratory Data Analyses) or mathematical (Classical Analysis), is the ability to accurately analyze and quantify the results. Most EDA techniques are graphical in nature with a few quantitative techniques. The reason for the heavy reliance on graphics is that by its nature the main role of EDA is to explore the patterns and clusters in the data. For classical analysis, the data collection is followed by the imposition of a model (normality, linearity, etc.) and the analysis, estimation, and testing that follows is focused on the parameters of that model. The Classical Analysis approach may include both deterministic and probabilistic models. Deterministic models include, for example, regression models and analysis of variance (MANOVA) models. The Exploratory Data Analysis approach does not impose deterministic models on the data. The two approaches differ substantially in focus. For classical analysis, the focus is on the model-- estimating parameters of the model, and generating predicted values from the model. Most log analysts will use a mix of graphical and classical quantitative techniques. The skill involved, is balancing the EDA approach, which is more subjective, against the more rigorous model fitting and testing algorithms. The questions that arise from the exploratory stage can then be further quantified and verified in the Classical Analysis stage.

The Guacho Kuč, 1999 Borehole Logging Project resulted in a vast amount of downhole geophysical data, 71 holes, each with eight physical properties, a data set that lends itself naturally to multivariate statistical analysis. The challenge was to accurately classify and quantify the physical properties to allow for lithology predictions. The creation of the relational database allowed for a quantitative, integrated interpretation of the data from multiple boreholes. This was a primary factor in establishing the physical property relationships between holes and establishing confidence in the classification results. Multiwell cross plotting (EDA) allowed for preliminary qualitative assessments in establishing the lithoclass boundaries. Multiple analysis of variance, MANOVA, is the

---

<sup>1,3</sup> 71 Cranbrooke Ave., Toronto, Ontario, M5M1M3, Canada. Email: susanne@earthfx.com, dirk@viewlog.com

<sup>2</sup> 1William Morgan Drive, Toronto, Ontario, M4H 1N6, Canada. Email: ghodgkinson@debeerscanada.com

<sup>4</sup> Applied Geophysics Research Group, Scholl of Geography and Geology, McMaster University, Hamilton, Ontario, Email: morriswa@mcmaster.ca

primary step to establish the underlying structure of the main and interaction effects on the categorical variables (lithoclass) and the multiple dependent interval variables (physical property parameters). Analysis showed that the fundamental modelling assumptions for multivariate analysis were met and that in all cases the results were significant at the 0.01 level. This provided a preliminary indication as to which physical property(s) would best discriminate certain lithoclasses. In general MANOVA can be viewed as a rigorous test for the subjective interpretations that result from EDA. Multiple discriminant analysis (MDA) shares many of the same principles as MANOVA, but is used primarily to classify the categorical dependant, investigate the difference between the lithoclasses and test the theory that the lithologies are classified as predicted. Analysis established that MDA was not the best method for classifying and predicting physical rock properties beyond a primary level of lithoclass distinction i.e. between distinct kimberlite phases and granite host rocks.

Analysis of the statistical parameters used in the predictive algorithms involved interpretation of the correlation matrix, matrix eigenvectors and principle component scores (PCA) to determine if the lithoclasses were uniquely defined. Ideally, the significant majority of the variation in the correlation matrix should be described by the first three eigenvectors i.e. PC1, PC2 and PC3. The most important factors are determined by the component matrix in PC Analysis. Analysis established that three parameters, Point Resistance, Sonic Velocity and Spontaneous Potential, best describe the uniqueness of each lithoclass. The accuracy of the classification and prediction depends primarily on the uniqueness of the downhole physical property responses of each lithoclass. If inconsistent or non-unique lithoclasses are used, classification results will be poor. The use of too few physical property parameters can inhibit classification, by not providing enough discrimination between units. Conversely, the use of too many parameters may smear the classes and result in poor discrimination. The log analyst must bring together both quantitative observations (EDA) and qualitative results (Classical Analysis), while simultaneously integrating his or her intuition of what is geologically viable.

## **Introduction**

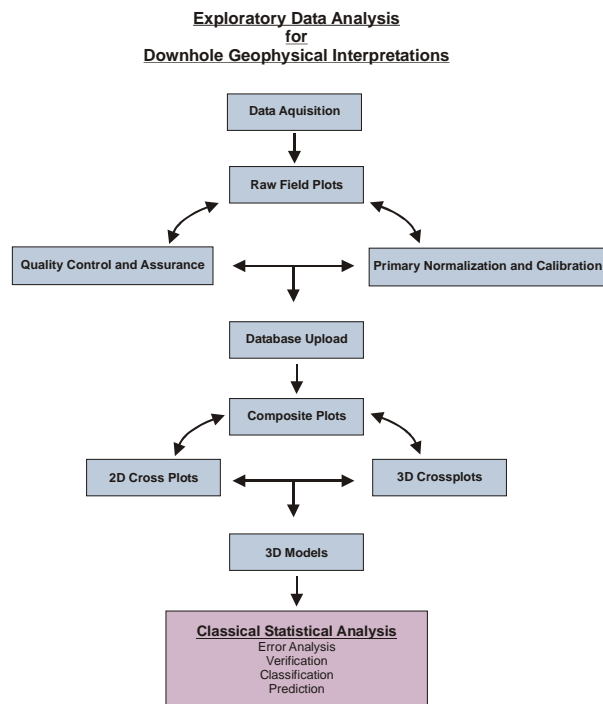
One of the challenges of geophysical interpretation, whether purely descriptive (EDA) or mathematically (Classical) is the ability to accurately analyze and quantify the results. Exploratory Data Analysis (EDA) is an approach/philosophy for data analysis, which employs a variety of techniques (mostly graphical) to;

- Maximize insight into a data set;
- Uncover underlying structure;
- Extract important variables;
- Detect outliers and anomalies;
- Test underlying assumptions;
- Determine optimal factor settings.

Most EDA techniques are graphical in nature with a few quantitative techniques. The reason for the heavy reliance on graphics is that by its nature the main role of EDA is to

explore. Thus for classical analysis, the data collection is followed by the imposition of a model (normality, linearity, etc.) and the analysis, estimation, and testing that follows is focused on the parameters of that model. For EDA, the data collection is not followed by a model imposition; rather it is followed immediately by analysis whose goal is to infer what model would be appropriate. The two approaches differ substantially in focus. For classical analysis, the focus is on the model-- estimating parameters of the model, and generating predicted values from the model. For exploratory data analysis, the focus is on the data-- its structure, outliers, and models suggested by the data.

A critical early step in any analysis is to identify (for the geophysical interpretation problem at hand) which of the above questions are relevant. That is, the researcher needs to identify which questions need to be answered and which questions have no bearing for the problem at hand. The next important step, which is invaluable for maintaining focus, is to prioritize those questions in decreasing order of importance. EDA techniques are tied in with each of the questions. Most data analysts will use a mix of graphical and classical quantitative techniques to address these problems. EDA and classical techniques are not mutually exclusive and can be used in a complimentary fashion.



**Figure I:** Statistical Analysis Model for Physical Rock Property Analysis

The MPV 1999 BHL Logging Project resulted in a vast amount of downhole geophysical data, 71 holes each with eight physical properties, a data set that lends it self naturally to multivariate statistical analysis. The initial challenge was to accurately analyze and quantify the physical properties, and from those properties develop a geophysical model for each kimberlite pipe. Developing and employing this stepwise strategy to achieve the end result, played an integral role in the early interpretation stages The questions that

arise from the exploratory stage can then be further quantified and verified in the Classical Analysis stage. Below is outlined the statistical model for physical Property Analysis.

### Geophysical Interpretation from Physical Properties

Multiparameter borehole geophysical measurements were collected at the Guacho Kuë property in the Lac de Gras area, Northwest Territories, to obtain *in situ* physical rock property data in kimberlites and their host rocks. Four pipes on the property were probed including; 5034, Hearne, Tuzo and Tesla. The tools used for this survey included; Natural Gamma, Neutron-Neutron, Gamma-Gamma, Inductive Conductivity, Spontaneous Potential, Point Resistance, Magnetic Susceptibility and 3 Arm Caliper. A total of 71 holes were surveys for the project including 28 NQ, diamond drilled holes and 43 reverse circulation holes ranging in diameter from HQ to 12.75” (33.15cm).

From the initial physical property logs three distinct kimberlite types were interpreted (K1, K2, K3), as were four country rock types (G1, G2, G3, G4). Subsets of these types were also identified for each kimberlite using statistical comparisons. A three-tier system of classification was used to facilitate the geological modeling of the bodies. A brief description of each kimberlite type see (MacMahon S. E., Wallace C., et al., 2002).

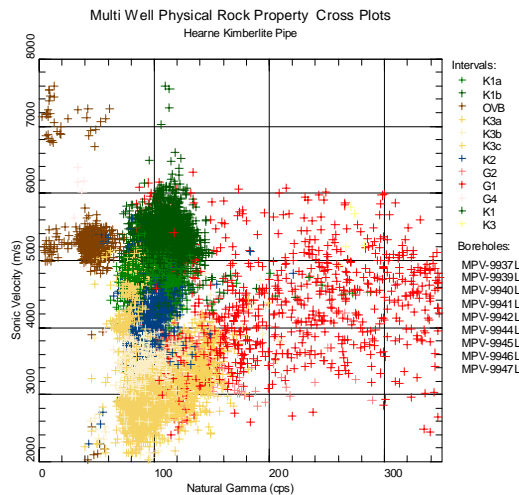


Figure II: 2D Multiwell Cross Plot of Hearne, Sonic Velocity (m/s) vs. Natural Gamma (cps)

On a cross plot, a lithoclass is represented by a cluster of points. The degree to which the points cluster in conjunction with the extent of separation between the individual clusters defines the “uniqueness” of each lithoclass. Multiwell cross plotting allows for the integrated solution, by plotting a series of holes together, testing the association, relationship and more importantly the correlatability of the lithoclasses across the body. The cross plots demonstrate the usefulness of exploratory data analysis as a methodology to be employed prior to any quantitative analysis. Examining Figure II, suggests that

Sonic Velocity and Natural Gamma are important factors in discriminating between the defined lithoclasses. Following the sequence for EDA, analysis of the data has inferred a model which can now be tested by classical statistical analysis. The subsequent step is to substantiate the probabilistic model via the rigorous mathematical methods associated with classical statistical analysis.

1. What are the most important factors?
2. Does any one factor have an effect?
3. What is the best function for relating a response variable to a set of factor variables?
4. Can accurate lithology and grade predictions from physical properties be made?

The next sections attempt to answer these questions using Classical Statistical Analysis, using MANOVA (Multiple Analysis of Variance), Discriminant Analysis and Factor Analysis.

### **Multivariate Analysis**

Multiple analysis of variance (MANOVA) is used to see the main and interaction effects of categorical variables on multiple dependent interval variables. MANOVA uses one or more categorical independents as predictors, and tests the differences in the centroid (vector) of means of the multiple interval dependents, for various categories of the independent(s).

There are multiple potential purposes for MANOVA.

1. To compare groups formed by categorical independent variables on group differences in a set of interval dependent variables.
2. To use lack of difference for a set of dependent variables as a criterion for reducing a set of independent variables to a smaller, more easily modeled number of variables.
3. To identify the independent variables which permits the most differentiation of a set of dependent variables.

Multivariate tests simultaneously examines each factor effect on the dependent groups. This is the most important table in the MANOVA output. Each factor (wellname and lithoclass in this example) has a main effect, as does the intercept. Interactions among the factors (here wellname\*lithoclass) are also assessed. Wilks' Lambda is commonly used if there are more than two groups, as there are in this example. The significance of the F tests show if that effect is significant. Eta-squared is the proportion of the total variability in the dependent variable accounted for by the variation in the independent variable. For the table below, the interpreted lithoclasses account for about ~30% of the variability in the parameters. Significance, of course, is the chance of making a Type I error (thinking you have something when you don't), whereas power (the last column, below) is the chance of making a Type II error thinking you don't have something when you do). Typically the power level should be high (eg., above .90).

Multivariate Tests<sup>d</sup>

Effect		Value	F	Hypothesis df	Error df	Sig.	Eta Squared	Noncent. Parameter	Observed Power <sup>a</sup>
Intercept	Pillai's Trace	.991	75016.837 <sup>b</sup>	9.000	6215.000	.000	.991	675151.536	1.000
	Wilks' Lambda	.009	75016.837 <sup>b</sup>	9.000	6215.000	.000	.991	675151.536	1.000
	Hotelling's Trace	108.633	75016.837 <sup>b</sup>	9.000	6215.000	.000	.991	675151.536	1.000
	Roy's Largest Root	108.633	75016.837 <sup>b</sup>	9.000	6215.000	.000	.991	675151.536	1.000
WELLNAME	Pillai's Trace	1.686	84.390	153.000	56007.000	.000	.187	12911.664	1.000
	Wilks' Lambda	.101	107.811	153.000	49784.455	.000	.225	14436.390	1.000
	Hotelling's Trace	3.438	139.608	153.000	55919.000	.000	.276	21359.990	1.000
	Roy's Largest Root	2.062	754.740 <sup>c</sup>	17.000	6223.000	.000	.673	12830.578	1.000
LITHONAM	Pillai's Trace	1.985	160.103	99.000	56007.000	.000	.221	15850.218	1.000
	Wilks' Lambda	.042	252.246	99.000	43839.737	.000	.298	18574.818	1.000
	Hotelling's Trace	6.054	379.944	99.000	55919.000	.000	.402	37614.491	1.000
	Roy's Largest Root	3.439	1945.524 <sup>c</sup>	11.000	6223.000	.000	.775	21400.763	1.000
WELLNAME * LITHONAM	Pillai's Trace	2.839	44.122	585.000	56007.000	.000	.315	25811.354	1.000
	Wilks' Lambda	.021	51.783	585.000	55463.442	.000	.351	29961.960	1.000
	Hotelling's Trace	5.742	60.983	585.000	55919.000	.000	.389	35675.069	1.000
	Roy's Largest Root	2.030	194.339 <sup>c</sup>	65.000	6223.000	.000	.670	12632.038	1.000

- a. Computed using alpha = .05
- b. Exact statistic
- c. The statistic is an upper bound on F that yields a lower bound on the significance level.
- d. Design: Intercept+WELLNAME+LITHONAM+WELLNAME \* LITHONAM

Figure III: Multivariate Tests

### Levene's Test of Equality of Error Variances

MANOVA assumes that each dependent variable will have similar variances for all groups and Levene's test examines this assumption. If the Levene statistic is significant at the .05 level or better, the homogeneity of variances assumption is met for all the parameters.

Levene's Test of Equality of Error Variances<sup>a</sup>

	F	df1	df2	Sig.
Sonic Velocity	21.398	93	6223	.000
3 Arm Caliper	40.351	93	6223	.000
Density	9.798	93	6223	.000
Spontaneous Potential	58.744	93	6223	.000
Point Resistance	28.344	93	6223	.000
Inductive Conductivity	69.590	93	6223	.000
Magnetic Susceptibility	46.379	93	6223	.000
Natural Gamma	67.137	93	6223	.000
Neutron	47.094	93	6223	.000

Tests the null hypothesis that the error variance of the dependent variable is equal across groups.

- a. Design: Intercept+WELLNAME+LITHONAM+WELLNAME \* LITHONAM

Figure IV: Levene's Test of Equality of Error Variances

### Homogeneous Subsets

Multiple range (homogenous subset) tests examine for homogenous subsets of groups based on their group means. The Tukey method also does range tests. Below is a table which lists all the groups (the categories on the independent variable) and their means for Sonic Velocity. Then for the .05 level of significance, additional columns will be printed, one for each subset where group means do not differ significantly. The subsets may overlap (a group may belong to more than one subset). Examination of the different

subset columns reveals for which groups (independent variable categories) the mean on the dependent do or do not differ.

		Sonic Velocity												
Lithology	N	Subset												
		1	2	3	4	5	6	7	8	9				
Tukey HSD <sup>a, c</sup>	K3c	151	3098.4728											
	K3b	582		3474.4594										
	G2	38			3986.2659									
	G1	677			4175.1588	4175.1588								
	K2	406			4313.5895	4313.5895								
	G3	139				4365.5873								
	K1a	509				4384.0774	4384.0774							
	K1	20					4720.7990	4720.7990						
	K1b	3402						4905.8537	4905.8537					
	G4	4							5131.0085	5131.0085				
	K1c	84								5343.0162	5343.0162			5343.0162
	OVB	305										5486.2406	5486.2406	
	Sig.		1.000	1.000	.104	.743	.081	.866	.641	.725	.977			

Means for groups in homogeneous subsets are displayed.

Based on Type III Sum of Squares

The error term is Mean Square(Error) = 194460.271.

a. Uses Harmonic Mean Sample Size = 33.037.

b. The group sizes are unequal. The harmonic mean of the group sizes is used. Type I error levels are not guaranteed.

c. Alpha = .05.

Figure V: Example of Homogeneous Subsets of Sonic Velocity

### MANOVA Results

Amalgamating the results for the previous example of MANOVA, from the example kimberlite pipe;

1. The physical properties differ in their covariance matrices.
2. All physical properties are normally distributed.
3. The interpreted lithoclasses account for 30% of the variability within the physical properties. This is a significant portion of the overall variability within the classes. Ideally, less than 10% would be better suited for the following analysis.
4. The homogeneity of variance assumption is met for all the physical properties, i.e. the physical properties do have equal variances.
5. The test of between subjects effects lends insight into which parameters account for the variability within the lithoclasses. Remember that Eta-squared is the portion of the total variability in the dependant variable accounted for by the variation in the independent variable.

LithoClass	Eta Squared
Sonic Velocity	42%
3 Arm Caliper	68%
Density	1%
Spontaneous Potential	31%
Point Resistance	47%
Inductive Conductivity	28%
Magnetic Susceptibility	58%
Natural Gamma	57%
Neutron	28%

Figure VI: Eta Squared for Defined Lithoclasses

The post hoc tests for multiple range or homogenous subset tests are also critical to the interpretation. In general no one parameter can accurately discriminate between each lithoclass. Sonic velocity is the best at discriminating between the lithoclasses with nine distinct subsets, followed by spontaneous potential, point resistance, 3 arm caliper, natural gamma and neutron each with six subsets, inductive conductivity and magnetic susceptibility with two subsets and density being the worst with only one subset. The next step in the statistical process is to establish, for each lithoclasses, which combination of parameters best describe how unique the class is and attach an error to that combination.

MANOVA gives additional insight and evidence of viability, into the subjective lithoclass classifications, which were derived from the Exploratory Data Analysis. As none of the underlying assumptions for MANOVA testing were violated, the log analyst can proceed with the prediction analysis.

### **Multiple Discriminant Analysis (MDA)**

Multiple discriminant analysis (MDA) is an extension of discriminant analysis and a cousin of multiple analysis of variance (MANOVA), sharing many of the same assumptions and tests. MDA is used to classify a categorical dependent which has more than two categories, using as predictors a number of interval or dummy independent variables. MDA is sometimes also called *discriminant factor analysis* or *canonical discriminant analysis*.

There are several purposes for MDA:

- To investigate differences among groups.
- To determine the most parsimonious way to distinguish among groups.
- To discard variables which are little related to group distinctions.
- To classify cases into groups.
- To test theory by observing whether cases are classified as predicted.

### ***Discriminant Functions for Multiple Groups***

When there are more than two groups, then we can estimate more than one discriminant function. For example, when there are three groups, we could estimate (1) a function for discriminating between group 1 and groups 2 and 3 combined, and (2) another function for discriminating between group 2 and group 3. In context, we could have one function that discriminates between Kimberlite and Granite, and a second function to discriminate between K1 and K2 versus G1 and G2, therefore then number of discriminant functions can be defined as (N-1).

### ***Standardized Canonical Discriminant Function Coefficients***

The standardized discriminant function coefficients in the table below serve the same purpose as beta weights in multiple regression: they indicate the relative importance of the independent variables in predicting the dependent.

Standardized Canonical Discriminant Function Coefficients									
	Function								
	1	2	3	4	5	6	7	8	9
Sonic Velocity	.520	-.299	.716	-.460	-.134	-.407	-.074	.126	-.120
3 Arm Caliper	.368	.608	-.258	-.063	.375	-.421	.368	.152	-.277
Density	-.010	-.051	.004	-.009	-.079	-.110	.048	.443	.902
Spontaneous Potential	-.132	.608	-.208	-.170	.507	-.573	-1.070	-.425	.266
Point Resistance	-.103	-.907	-.132	.397	.493	.551	.914	.268	-.163
Inductive Conductivity	.142	.111	.109	.007	.022	.515	-.566	.660	-.190
Magnetic Susceptibility	.345	.213	.185	.073	.142	.553	.300	-.600	.359
Natural Gamma	-.588	.303	.284	-.628	.255	.241	.195	.052	.046
Neutron	-.179	.371	.337	.919	.133	-.155	.085	.050	-.065

Figure VII: Standardized Canonical Discriminant Function Coefficients

### ***Structure Matrix***

The structure matrix table below shows the correlations of each variable with each discriminant function. The correlations serve like factor loadings in factor analysis, that is, by identifying the largest absolute correlations associated with each discriminant function the researcher gains insight into how to name each function.

### ***Structure coefficients vs. standardized discriminant function coefficients.***

The standardized discriminant function coefficients (above) indicate the partial contribution of each variable to the discriminant function(s), controlling for other independents entered in the equation. The structure coefficients (below) indicate the simple correlations between the variables and the discriminant function or functions. The structure coefficients should be used to assign meaningful labels to the discriminant functions. The standardized discriminant function coefficients should be used to assess each independent variable's unique contribution to the discriminant function. Physical Property combinations provide insight of the composition of the rock unit. The first function is defined by enhanced Magnetic Susceptibility and Caliper, but is inversely correlated to Natural Gamma. The second function has is described by Point Resistance and is inversely proportional to Caliper .

Structure Matrix

	Function								
	1	2	3	4	5	6	7	8	9
Natural Gamma	-.635*	.293	.426	-.438	.239	.134	.229	.099	-.020
3 Arm Caliper	.457	.500*	-.314	-.142	.306	-.205	.401	.323	-.146
Sonic Velocity	.362	-.306	.748*	-.174	.281	-.317	-.056	.029	-.044
Neutron	-.272	.251	.626	.650*	.097	-.174	.052	.072	-.036
Point Resistance	.012	-.533	-.006	.096	.836*	.026	.037	.073	.009
Spontaneous Potential	-.011	-.216	-.014	.009	.803*	-.178	-.462	-.197	.158
Inductive Conductivity	.310	.181	.070	.030	.190	.576*	-.424	.545	-.144
Magnetic Susceptibility	.517	.241	.220	.025	.185	.533*	.085	-.448	.315
Density	.067	.025	-.043	.014	.007	-.158	.128	.467	.856*

Pooled within-groups correlations between discriminating variables and standardized canonical discriminant functions  
Variables ordered by absolute size of correlation within function.

\*. Largest absolute correlation between each variable and any discriminant function

Figure VIII: Structure Matrix

## MDA Results

MDA is an extremely robust method used to discriminate between different groups and classifying cases into different groups. As a cousin to MANOVA the two methods share some similar assumptions with some exceptions. One of the main reasons why multiple discriminant analysis does not work well as a method for prediction is that the group sizes of the dependent variables (lithoclass) vary considerably. Secondly, MDA is highly sensitive to outliers, which in turn effects the mean, variance and the standard deviations of the predictors. If there are correlations between these factors then the validity of the significance tests can be compromised.

Closer examination of the classification results clearly indicate were some of the problems arise;

1. G1 (47.9%), G2 (60.5%) and G3 (51.9%) are not uniquely enough defined, to allow for discrimination between one another.
2. G4 (75%) and K3b (66.2%) fail to discriminate between one another.
3. K1 (95%) and K1c (88.1%) have excellent classification results, whereas K1a (66.6%) and K1b (55.7%), fail to discriminated well from one another.
4. K3c (94%) and OVB (83.3%) also have excellent classification results.

An interesting test of the underlying assumptions, would be to turn back to the exploratory data analysis stage and re classify the lithoclasses to only an primary level as opposed to a secondary level. In doing so the assumption of group sizes being not being grossly different is met. The underlying assumptions of normality and homogeneity are still significant (i.e. rejecting the null hypothesis), but more importantly the correlations between the mean, variance and standard deviation are no longer correlated. Below are the classification results from MDA on the primary lithoclasses with 85.5% being correctly classified.

Classification Results<sup>b,c</sup>

Original	CASENUM	Predicted Group Membership											Total		
		G1	G2	G3	G4	K1	K1a	K1b	K1c	K2	K3b	K3c		OVB	
Count	G1	325	109	149	0	18	9	3	9	6	13	36	0	677	
	G2	0	24	3	0	0	1	4	0	0	4	2	0	38	
	G3	20	11	72	1	0	2	12	5	0	0	16	0	139	
	G4	0	0	0	3	0	0	0	0	0	1	0	0	4	
	K1	0	0	1	0	19	0	0	0	0	0	0	0	20	
	K1a	2	36	0	1	18	341	63	16	18	5	7	2	509	
	K1b	0	3	111	1	330	681	1897	157	196	1	6	19	3402	
	K1c	0	0	0	0	0	4	0	74	6	0	0	0	84	
	K2	0	0	0	2	70	60	4	3	229	32	5	1	406	
	K3b	2	50	2	1	1	23	1	4	26	389	82	1	582	
	K3c	0	0	0	1	0	2	0	0	0	3	143	2	151	
	OVB	0	0	0	41	0	0	0	0	0	7	1	256	305	
	%	G1	48.0	16.1	22.0	.0	2.7	1.3	.4	1.3	.9	1.9	5.3	.0	100.0
		G2	.0	63.2	7.9	.0	.0	2.6	10.5	.0	.0	10.5	5.3	.0	100.0
		G3	14.4	7.9	51.8	.7	.0	1.4	8.6	3.6	.0	.0	11.5	.0	100.0
		G4	.0	.0	.0	75.0	.0	.0	.0	.0	.0	25.0	.0	.0	100.0
K1		.0	.0	5.0	.0	95.0	.0	.0	.0	.0	.0	.0	.0	100.0	
K1a		.4	7.1	.0	.2	3.5	67.0	12.4	3.1	3.5	1.0	1.4	.4	100.0	
K1b		.0	.1	3.3	.0	9.7	20.0	55.8	4.6	5.8	.0	.2	.6	100.0	
K1c		.0	.0	.0	.0	.0	4.8	.0	88.1	7.1	.0	.0	.0	100.0	
K2		.0	.0	.0	.5	17.2	14.8	1.0	.7	56.4	7.9	1.2	.2	100.0	
K3b		.3	8.6	.3	.2	.2	4.0	.2	.7	4.5	66.8	14.1	.2	100.0	
K3c		.0	.0	.0	.7	.0	1.3	.0	.0	.0	2.0	94.7	1.3	100.0	
OVB		.0	.0	.0	13.4	.0	.0	.0	.0	.0	2.3	.3	83.9	100.0	
Cross-validated <sup>a</sup> Count		G1	324	110	149	0	18	9	3	9	6	13	36	0	677
		G2	0	23	4	0	0	1	4	0	0	4	2	0	38
		G3	20	11	72	1	0	2	12	5	0	0	16	0	139
		G4	0	0	0	3	0	0	0	0	0	1	0	0	4
	K1	0	0	1	0	19	0	0	0	0	0	0	0	20	
	K1a	2	36	0	1	18	339	64	16	18	6	7	2	509	
	K1b	0	3	112	1	330	681	1896	157	196	1	6	19	3402	
	K1c	0	0	0	0	0	4	0	74	6	0	0	0	84	
	K2	0	0	0	2	70	61	4	3	228	32	5	1	406	
	K3b	2	52	2	1	1	23	1	4	27	385	82	2	582	
	K3c	0	0	0	1	0	2	0	0	0	4	142	2	151	
	OVB	0	0	0	41	0	0	0	0	0	9	1	254	305	
	%	G1	47.9	16.2	22.0	.0	2.7	1.3	.4	1.3	.9	1.9	5.3	.0	100.0
		G2	.0	60.5	10.5	.0	.0	2.6	10.5	.0	.0	10.5	5.3	.0	100.0
		G3	14.4	7.9	51.8	.7	.0	1.4	8.6	3.6	.0	.0	11.5	.0	100.0
		G4	.0	.0	.0	75.0	.0	.0	.0	.0	.0	25.0	.0	.0	100.0
K1		.0	.0	5.0	.0	95.0	.0	.0	.0	.0	.0	.0	.0	100.0	
K1a		.4	7.1	.0	.2	3.5	66.6	12.6	3.1	3.5	1.2	1.4	.4	100.0	
K1b		.0	.1	3.3	.0	9.7	20.0	55.7	4.6	5.8	.0	.2	.6	100.0	
K1c		.0	.0	.0	.0	.0	4.8	.0	88.1	7.1	.0	.0	.0	100.0	
K2		.0	.0	.0	.5	17.2	15.0	1.0	.7	56.2	7.9	1.2	.2	100.0	
K3b		.3	8.9	.3	.2	.2	4.0	.2	.7	4.6	66.2	14.1	.3	100.0	
K3c		.0	.0	.0	.7	.0	1.3	.0	.0	.0	2.6	94.0	1.3	100.0	
OVB		.0	.0	.0	13.4	.0	.0	.0	.0	.0	3.0	.3	83.3	100.0	

- a. Cross validation is done only for those cases in the analysis. In cross validation, each case is classified by the functions derived from all cases other than that case.
- b. 59.7% of original grouped cases correctly classified.
- c. 59.5% of cross-validated grouped cases correctly classified.

Figure IX: MDA Classification Results

## Principle Component Analysis

Principle Component analysis uses the correlation matrix to try to determine which sets of variables cluster together.

Correlation Matrix<sup>a</sup>

	Sonic Velocity	3 Arm Caliper	Density	Spontaneous Potential	Point Resistance	Inductive Conductivity	Magnetic Susceptibility	Natural Gamma	Neutron
Correlation	Sonic Velocity	1.000	-.070	.016	-.317	-.403	.143	-.306	-.180
	3 Arm Caliper	-.070	1.000	.195	-.200	-.274	.418	.442	-.258
	Density	.016	.195	1.000	.000	.008	.057	.080	-.086
	Spontaneous Potential	.317	-.200	.000	1.000	.761	-.008	-.021	-.081
	Point Resistance	.403	-.274	.008	.761	1.000	-.053	-.128	-.167
	Inductive Conductivity	.143	.418	.057	-.008	-.053	1.000	.495	-.230
	Magnetic Susceptibility	.306	.442	.080	-.021	-.128	.495	1.000	-.331
	Natural Gamma	-.180	-.258	-.086	-.081	-.167	-.230	-.331	1.000
	Neutron	-.134	-.211	-.036	-.070	-.165	-.077	-.085	.517
Sig. (1-tailed)	Sonic Velocity		.000	.107	.000	.000	.000	.000	.000
	3 Arm Caliper	.000		.000	.000	.000	.000	.000	.000
	Density	.107	.000		.490	.255	.000	.000	.002
	Spontaneous Potential	.000	.000	.490		.258	.045	.000	.000
	Point Resistance	.000	.000	.255	.000		.000	.000	.000
	Inductive Conductivity	.000	.000	.000	.258	.000		.000	.000
	Magnetic Susceptibility	.000	.000	.000	.045	.000	.000		.000
	Natural Gamma	.000	.000	.000	.000	.000	.000	.000	
	Neutron	.000	.000	.002	.000	.000	.000	.000	

a. Determinant = 7.200E-02

Figure X: Correlation Matrix for PCA

The "Total Variance Explained" table below shows the eigenvalues, which are the proportion of total variance in all the variables which is accounted for by that component. A component's eigenvalue may be computed as the sum of its squared component loadings for all the variables. A component's eigenvalue divided by the number of variables (which equals the sum of variances because the variance of a standardized variable equals 1) is the percent of variance in all the variables which it explains. The ratio of eigenvalues is the ratio of explanatory importance of the components with respect to the variables. If a component has a low eigenvalue, then it is contributing little to the explanation of variances in the variables and may be ignored as redundant with more important components. Figure XI, shows nine components, one for each variable. The "Initial Eigenvalues" and "Extraction Sums of Squared Loadings" columns are the same, except the latter only lists components, which have actually been extracted in the solution. The "Rotation Sums of Squared Loadings" give the eigenvalues after rotation improves the interpretability of the components. Varimax rotation was employed, which minimizes the number of variables which have high loadings on each given component. Note that the total percent of variance explained is the same (see the cumulative value for component 3 – 64.312%) but rotation changes the eigenvalues for each of the extracted components. That is, after rotation each extracted component counts for a different percentage of variance explained, even though the total variance explained is the same.

The Component Matrix, Figure XII, gives the component loadings. This is the central output for Principle Component Analysis. The component loadings, are the correlation coefficients between the variables (rows) and components (columns). Component loadings are the basis for imputing a label to the different components. Loadings above .6 are usually considered "high" and those below .4 are "low."

**Total Variance Explained**

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	2.322	25.796	25.796	2.322	25.796	25.796	2.119	23.546	23.546
2	2.153	23.917	49.713	2.153	23.917	49.713	2.110	23.448	46.994
3	1.314	14.600	64.312	1.314	14.600	64.312	1.559	17.318	64.312
4	.993	11.034	75.346						
5	.731	8.118	83.465						
6	.502	5.576	89.041						
7	.420	4.662	93.703						
8	.376	4.175	97.878						
9	.191	2.122	100.000						

Extraction Method: Principal Component Analysis.

Figure XI: Total Variance Explain for PCA

Figure XII, gives the unrotated solution and Figure XIII, the rotated solution. Normally the rotated solution will be significantly easier to interpret. Looking at the rotated matrix, the first component has high loadings from three variables: Spontaneous Potential, Point Resistance and Sonic Velocity. Because these three variables items sort on the same component, this is a justification for combining these items in a single group, i.e. the "most important physical property parameters". Magnetic Susceptibility, Inductive Conductivity and 3 Arm Caliper are associated strongly with the second component. Neutron and Natural Gamma are associated strongly with the third component.

Interesting that Density is not associated with any of the extracted components, indicating that Density may not play a significant role in classification or discrimination.

**Component Matrix<sup>a</sup>**

	Component		
	1	2	3
Sonic Velocity	.231	.591	.528
3 Arm Caliper	.718	-.337	-1.15E-02
Density	.239	-2.01E-02	-3.59E-02
Spontaneous Potential	-4.53E-02	.858	2.922E-02
Point Resistance	-6.76E-02	.921	-8.05E-02
Inductive Conductivity	.692	-4.04E-02	.319
Magnetic Susceptibility	.767	-1.88E-02	.371
Natural Gamma	-.658	-.272	.418
Neutron	-.434	-.171	.783

Extraction Method: Principal Component Analysis.  
a. 3 components extracted.

Figure XII: Component Matrix

**Rotated Component Matrix<sup>a</sup>**

	Component		
	1	2	3
Sonic Velocity	.685	.388	-.247
3 Arm Caliper	-.329	.662	-.286
Density	-2.59E-02	.197	-.139
Spontaneous Potential	.847	-.110	-.104
Point Resistance	.886	-.185	-.199
Inductive Conductivity	2.666E-02	.762	-3.95E-02
Magnetic Susceptibility	5.865E-02	.850	-3.26E-02
Natural Gamma	-.186	-.365	.717
Neutron	-1.31E-02	-1.05E-02	.911

Extraction Method: Principal Component Analysis.  
Rotation Method: Varimax with Kaiser Normalization.  
a. Rotation converged in 5 iterations.

Figure XIII: Rotated Component Matrix

### General Observations

PCA determines the factors which can account for the total (unique and common) variance in a set of variables. This is appropriate for creating a typology of variables or reducing attribute space. PCA is appropriate for most social science research purposes and is the most often used form of factor analysis.

There are conceptual similarities between multiple discriminant function and factors in principle component factoring, but they are mathematically different in what they are maximizing. MDA is maximizing the difference between values of the dependent. PCA is maximizing the variance in all the variables accounted for by the factor.

Results of PCA are not affected by standardization, which is built into the procedure. Note, however, that standardization (subtracting the mean, dividing by the standard deviation) scales data in a sample-specific way. If the log analyst's purpose is to compare factor structures between two or more samples, then one should use the covariance matrix rather than the correlation matrix as input to factor analysis. However, the covariance method has problems when the variables are measured on widely different scales (ex., sonic velocity ranging from 2000 to 8000 m/s as opposed to Density which ranges from 1.2 to 3.9 g/cc). To overcome this problem multisample standardization is recommended (subtracting the grand mean of all samples, and dividing by the standard deviation of all cases) prior to computing the sample covariance matrix .

### Statistical Application to Auto Interpretation/Prediction

#### Auto Prediction Analysis

The selection of the classification algorithm depends on the form and type of the available log interpretation knowledge. The statistical pattern recognition techniques (listed below) were selected for the following reasons;

Simultaneous Matrix Solution. Statistical pattern recognition algorithms simultaneously consider all the measurement features during classifications through the use of matrix techniques. This approach ensures that missing or incorrect measurements are minimized and that a solution will be determined.

The Construction of a Knowledge Base. The user first must define and locate a number of litho classes (or grade classes) in various representative boreholes. Following the test interpretation, the statistical parameters for each litho class (or grade class) are compiled into the knowledge base. As new classes are interpreted the knowledge base can incorporate the new information, thus growing through statistical “learning”.

Knowledge Base Evaluation. The use of techniques such as Principle Component Analysis, Factor Analysis and other multivariate statistical algorithms can then evaluate the system classification techniques.

### **Classification Methods and Lithology Prediction Results**

The Multiwell Analysis/Cross Plotting/Principle Component Analysis and Classification methods Predictive Analysis were completed using VIEWLOG DB/PRO/MW(stats). Below are outlined the three classification methodologies used in the prediction analysis. All three methods were tested, with the Minimum Intra-Class Distance (MICD) providing the best results Figure XVI and Figure XVII. All predictions were subsequently calculated using the MICD classifier. The primary reason for success of this classifier is that the probability distribution is not as dependent or sensitive to normally distributed data.

**Minimum Euclidean Distance (MED):** The MED technique bases classification on the distance between the feature vector  $\underline{x}$  and the prototype of the class  $\underline{z}_k$ . The Euclidian distance between the prototype of class k and the unknown vector is:

$$d_{MED}(\underline{x}, \underline{z}_k) = [ (\underline{x} - \underline{z}_k)^T (\underline{x} - \underline{z}_k) ]^{(1/2)}$$

The class exhibiting the minimum distance to the vector  $\underline{x}$  forms the interpretation. This technique is most suitable for classes with spherical and distinct distributions in n-space. No knowledge of the actual distribution of the class around the prototype  $\underline{z}_k$  is utilized in the metric, and thus calculation of the covariance matrix  $E_k$  is not required.

**Minimum Intra-Class Distance (MICD):** If the covariance matrix  $E_k$  is known, the shape of the distribution in n-space can be used to enhance the classifier. The classifier selects the class k exhibiting the minimum of the following distance metric:

$$d_{MICD}(\underline{x}, \underline{z}_k) = [ (\underline{x} - \underline{z}_k)^T E_k^{-1} (\underline{x} - \underline{z}_k) ]^{(1/2)}$$

The advantage of the MICD classifier over the MED classifier is that the shape of the distribution of data points within n-space, represented by  $E_k$ , is accounted for during classification. This allows better classification when classes exhibit elongate or non-spherical distributions. The technique, however, does not require any assumptions concerning the underlying distribution of the data. For example, the data within a class need not be normally distributed.

**Maximum A-Posteriori Classification (MAP):** When the class conditional probability density function is known, the probability that the set of measurements,  $\underline{x}$ , is a member of a class k can be determined. The classifier simply suggests the class which has the greater probability given the particular value of  $\underline{x}$ . Written in terms of a posteriori probability,

$$\underline{x}, \text{ Class}_k \text{ if } P(\text{Class}_k | \underline{x}) > P(\text{Class}_i | \underline{x}) \quad i = 1 \dots n$$

This classifier is referred to the Maximum A-Posteriori classifier since  $P(\text{Class}_k | \underline{x})$  is the a-posteriori class probability given the observed pattern  $\underline{x}$ .

Using Bayes Theorem, the MAP classifier may be written in terms of a probabilistic model:

$$\underline{x}, \text{ Class}_k \text{ if } p(\underline{x} | \text{Class}_k) P(\text{Class}_k) > p(\underline{x} | \text{Class}_i) P(\text{Class}_i)$$

Where  $P(\text{Class}_k)$  is the a priori class probability and the common denominator of  $p(\underline{x})$  has been dropped since  $p(\underline{x}) > 0$  in all regions of interest.

For multivariate Normal, or Gaussian, probability density functions:

$$p(\underline{x} | \text{Class}_k) = (2B)^{-n/2} |E_k|^{-1/2} \exp[-1/2 (\underline{x} - \underline{z}_k)^T E_k^{-1} (\underline{x} - \underline{z}_k)]$$

Where:

- k = class number
- n = number of features
- $\underline{z}$  = class mean vector
- $\underline{x}$  = unknown measurement vector
- $E_k$  = Covariance matrix for class k
- $|E_k|$  = determinant of the Covariance matrix

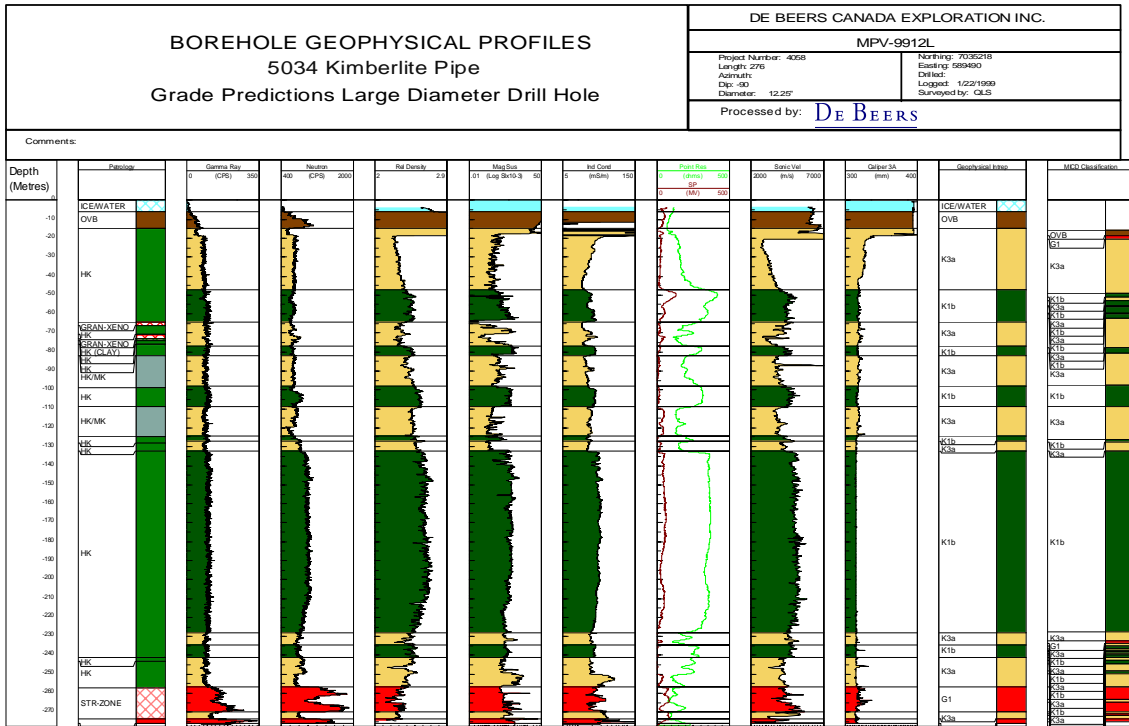


Figure XIV: Composite Plot of MPV-9912L, including observed Geology, Geophysical Interpretation and Auto Predictions

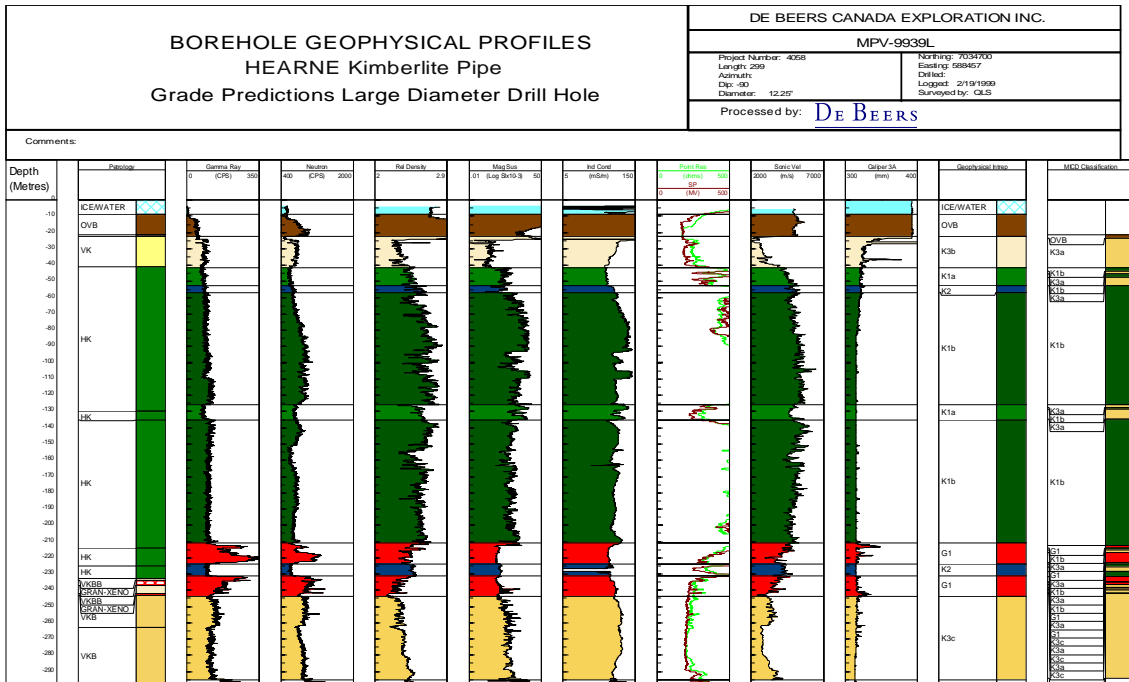


Figure XV: Composite Plot of MPV-9939L, including observed Geology, Geophysical Interpretation and Auto Predictions

## Conclusion

The critical criteria that underlies the entire classification and predictive analysis is determined whether or not there is any structure within the multivariate dataset. Specifically if there were no structure to the data, then there would be no valid statistical method that would provide a solution. Multiple analysis of variance, MANOVA, is the primary step to establish the underlying structure, the main and interaction effects of the categorical variables (lithoclass) on the multiple dependent interval variables (physical property parameters). MANOVA testing established that the underlying fundamental assumptions for multivariate analysis had been met and that in all cases the results were significant at the 0.01 level and gave preliminary as to which physical property(s) would best discriminate certain lithoclasses. For the less experienced log analyst the MANOVA process may be iterative. If the underlying assumptions, in particular the individual distributions, fail, examination of the original picks and cross plots from the exploratory data analysis stage would have to be revisited. In general MANOVA can be viewed as the rigorous test for the interpretations that result from EDA.

Multiple discriminant analysis shares many of the same principles with MANOVA, but is used primarily to classify the categorical dependant, investigate the difference between the groups and test the theory that the cases are classified as predicted. MDA was not the best method for classifying and predicting physical rock properties to a secondary (K1a, K1b etc.) lithoclass level, but MDA analysis works extremely well on a primary level (Granites, K1, K2, K3). That then evokes the question; at what level in the Exploratory Analysis Stage is it sufficient to optimally describe the visual observations, yet maintain geological accuracy? The solution is not easy. It becomes one of a balance between quantitative observations and qualitative results, while simultaneously integrating the log analyst intuition of what is geologically viable.

The most important factors are determined by the component matrix in Principle Component Analysis. The Figure XIV, shows that Point Resistance, Spontaneous Potential and Sonic Velocity account for the maximum variance in Component 1. Thus, those three parameters best describe the uniqueness of each lithoclass. The ability of each lithoclass to be unique, i.e. the more parameters that describe the uniqueness, the better the prediction results. Post Hoc testing in MANOVA, specifically multiple range tests (homogeneous subsets), will also provide a preliminary assessment of the most important factors.

All of the physical properties that define the lithoclasses play an interactive role in the analysis. This is best seen in the MANOVA test for between subject effects. Although some of the parameters are more significant in their role (see above), each parameter is significant to the prediction analysis. The accuracy of the interpretation depends primarily on the uniqueness of the downhole physical property responses of each lithoclass. If inconsistent or non-unique lithoclasses are used, classification results will be poor. The use of too few physical property parameters can inhibit classification, by not providing enough discrimination between units. Analysis of the statistical parameters used in the predictive algorithms (correlation matrix, matrix eigenvectors and principle

component scores) should be examined to determine if the lithoclasses are uniquely defined. In this study, the significant majority of the variation in the correlation matrix should be described by the first three eigenvectors i.e. PC1, PC2 and PC3. The eigenvalue for a given factor measures the variance in all the variables which is accounted for by that factor. That is, eigenvalues measure the amount of variation in the total sample accounted for by each factor. A factor's eigenvalue may be computed as the sum of its squared factor loadings for all the variables. The ratio of eigenvalues is the ratio of explanatory importance of the factors with respect to the variables. If a factor has a low eigenvalue, then it is contributing little to the explanation of variances in the variables and may be ignored as redundant with more important factors.

*MacMahon S. E., Wallace C., Kassenaar D.J., Morris W.A., Multiwell Analysis of Downhole Physical Rock Properties of Kimberlite: Guacho Kuč, Northwest Territories, in Proceedings of the 8<sup>th</sup> International KEGS/MGLS Symposium on Logging for Minerals and Geotechnical Applications, Toronto 21-23 August 2002*