IDEAS: an interactive computer graphics tool to assist the exploration geochemist

Robert G. Garrett
Mineral Resources Division

Garrett, R.G., IDEAS: an interactive computer graphics tool to assist the exploration geochemist; in Current Research, Part F, Geological Survey of Canada, Paper 88-1F, p. 1-13, 1988.

Abstract

A colour interactive graphics computer package (IDEAS) is being developed to meet the data analysis needs of exploration geochemists. At this time enough of the package exists for it to be useful production tool, and it is being used to aid data interpretation and prepare graphics for publication and posters. The current version includes the majority of the plotting, database, and simple statistical functions to be implemented. Some multivariate statistical procedures are in place, and it is in this field that the majority of the remaining work will be undertaken. The package uses a combination of Digital Equipment Corporation (VAX 8700) computing and Tektronix graphics display equipment. The development history of the project is outlined together with a description of current functionality and future enhancements.

Résumé

On met actuellement au point un progiciel d'infographie interactive en couleurs, pour donner aux géochimistes d'exploration les données dont ils ont besoin pour leurs analyses. Actuellement, ce progiciel est suffisamment élaboré pour constituer un outil de production d'une grande utilité, et facilite l'interprétation des données et la préparation du travail graphique pour la mise au point de publications et d'affiches. La version actuelle comprend la majeure partie du traçage, de la base de données, et des simples fonctions statistiques appliquées. Quelques-uns des procédés statistiques multivariés sont déjà en place, et c'est dans ce domaine que la majorité du travail restant sera entrepris. Le progiciel utilise une combinaison du parc informatique de la Digital Equipment Corporation (VAX 8700) et de l'unité d'affichage graphique Tektronix. Dans cet article, on esquisse l'historique du projet, et l'on décrit son développement actuel et les perfectionnements envisagés.

INTRODUCTION

The concept of an interactive computer graphics package tailored to the needs of exploration geochemists crystallized during a review of the Exploration Geochemistry Subdivision's computing requirements in 1979. Prior to that time several interactive graphics programs had been written for use on the Departmental CDC Cyber computing facilities. These were useful for anomaly recognition and general data interpretation; however, they were not "friendly" and use of the CDC Cyber computers placed certain restrictions on their operation.

Between 1980 and 1982, a functional specification was developed under a joint Department of Energy, Mines and Resources — Department of Supply and Services Unsolicited Proposal. Under this Unsolicited Proposal made by Hickling-Smith Inc., a consulting group with computer science and statistics experience, their staff undertook extensive discussions with members of the Subdivision and perpared a functional specification for an interactive graphics package. There was a hiatus in development from 1932 to 1984 when a Digital Equipment Corporation (DEC) VAX computer and colour graphics terminal equipment finally became available to the Subdivision.

During this hiatus a number of key decisions were made. Firstly, the U.S. Geological Survey's GRASP database system was selected as the foundation upon which to develop the new system (Bowen and Botbol, 1975). The reasons for selecting GRASP were several, but the most important were its:

1. Friendliness and extensive on-line help facilities,

2. Ability to process qualified data, e.g., greater than or less than, and blanks in a way consistent with Subdivision usage in geochemical data files,

3. Simplicity of the database — post retrieval processing interface, and

4. Presence in the public domain, making it possible to modify GRASP as required for the interactive graphics project, and distribute the resulting software in whatever way the Geological Survey saw fit.

Secondly, it was decided at an early stage that the graphics software would be written using the GKS (Graphics Kernel System) standard largely developed in Europe (1976-1982) for adoption by the International Standards Organization (ISO) (Enderle et al., 1984). Thirdly, the decision was made to write all software to Fortran 77 standards using the minimal number of system dependent features, and where possible any non-standard Fortran would be isolated in subroutines.

The acquisition of a DEC VAX 11/780 and the VMS Operating System in 1984 by the Geological Survey of Canada and the Department's Computer Science Centre for interactive graphics use permitted development to proceed. The U.S. Geological Survey kindly made a VAX version of GRASP available to the project. Additionally, the selection of VAX technology met requirements for technology transfer. This family of 32 bit machines is extensively used in the mineral exploration industry, and by Geological Surveys, in North America, Europe, the Far East and elsewhere.

Tektronix hardware was selected, because it could be configured as a flexible graphics workstation, consisting of a high resolution 16 colour display, digitizing tablet and plotter(s), with its own internal graphics central processing unit (cpu). The availability of this local computing power reduces the amount of graphics information to be sent from the remote VAX host to the workstation and permits local true zoom and pan functions, thus improving the overall response time. Integral to the decision to use Tektronix equipment was the availability of the company's PLOT10GKS software which met the ISO's level 2.B standard, i.e. interactive input of graphical information and graphical output.

IDEAS DEVELOPMENT

Database

The acronym of IDEAS, Interactive Data Exploration and Analysis System, was finally selected in 1983, and implementation of the functional design began in summer 1984 with the establishment of GRASP on the VAX host as soon as it was available for use. Initially GRASP could undertake the following tasks:

1. Build a database from an existing file of data organized by rows and columns, with up to 99 columns (variables) and an unlimited number of rows (cases/samples),

2. Handle below detection level and other special data qualifiers,

3. Extend a database, temporarily by columns and permanently by rows,

4. Link databases via a common key,

5. List variable names and their data types, i.e., integer, floating point, qualified floating point, dictionary items, long and short character strings,

6. Maintain an alternate dictionary for expanded dictionary names, e.g., mnemonic expansions,

7. Display a help file,

8. Review current session activity and status,

9. Define conditional and logical criteria to be used for subset creation,

10. List, after sorting if required, all or selected fields of the database or a subset,

11. Output, after sorting if required, all or selected fields of the database or a subset to a disk file for subsequent processing or transmission or another facility,

12. Browse through data records using every n-th record, where the user specifies n,

13. Temporarily exit to the operating system, and

14. Compute means, correlation coefficients and a simple linear regression.

Major development work commenced in summer 1985. GRASP uses a sequential file structure to store a database and any subsets, resulting in multiple copies of parts of the database, a feature that can hinder database management.

In particular, these multiple copies make editing and the addition of new derived variables difficult to undertake in a fail-safe manner. The first task was therefore to convert the database structure from sequential to keyed access, specifically the keyed Indexed Sequential Access Method (keyed ISAM), so that each database entry could be read or written individually. The subsets are maintained in a separate keyed ISAM file as lists of pointers (accession numbers) to the members of that particular subset, together with information on the size, parent, time of creation of the subset and a text description. This has many beneficial features, the single copy of the database can be edited and added to irrespective of the subsets. Most importantly, this use of pointers establishes a favourable environment for interfacing the database to the user through the graphics screen by use of GKS numbered graphic segments. Changing the database structure to keyed access allowed the addition of:

1. Subset management for up to 75 subsets,

2. The editing of numeric and short character string data, and

3. The permanent storage of new computed variables with the necessary extensions to the database files,

to the original GRASP software.

Subsequently, other modifications have been made to the database and management components of IDEAS. The original GRASP "sorted-list" facility only permitted ascending order sorts. This has been modified to provide either ascending or descending order sorts, with the latter as the default as usually exploration geochemists are more interested in the high values than the low. The original upper and lower case alphabetic data qualifiers have been replaced by the characters >, <, #, and . The last 3 qualifiers, #, and may be employed in any way the user wishes to indicate some feature of the data. For example, possible analytical interference, or a change in the least significant digit by 1 so that differences of zero between duplicate analyses do not occur. This latter problem is of importance when certain log-log quality control plots are prepared with IDEAS. As in GRASP, selective retrievals can be made on the basis of data qualifiers. However, in the instance of the > and < qualifiers, numeric post-retrieval processing is carried out in a special manner. During numeric post-retrieval processing by GRASP the data qualifiers are ignored, i.e., #5 is treated as 5; however, in IDEAS > and < the values are now respectively doubled or multiplied by 5/8th, i.e., >1 000 is converted to 2 000 and <8 is treated as 5, and the data qualifier is temporarily removed. A module has been implemented to provide a summary of qualified data use. The number of occurrences of missing, unqualified and each qualified data type are displayed together with the maximum and minimum values associated with each. This has proven particularly useful with Instrumental Neutron Activation Analysis (INAA) data where significant variations in the detection level may occur. To improve the response time for help requests, the original sequential help file has been replaced by a keyed ISAM file. This has resulted in a marked improvement, with responses being equally fast for all resquests. This is particularly advantageous as the IDEAS help file is significantly larger than the original GRASP help file, containing in excess of 250 entries totalling some 2 000 lines, and will continue to grow as the system develops.

Graphics

IDEAS really took shape once graphics work commenced after the database system modifications were completed. At the current time the graphics modules are limited to displaying plots with up to 7 000 data points. To date, modules have been written which permit the plotting of:

1. Histograms, including means for up to 3 auxiliary variables,

2. Cumulative probability plots with provision for graphical trimming,

3. Simple x-y plots,

4. Up to 9 data subsets as uniquely coloured and shaped symbols on an x-y plot,

5. A third auxiliary variable as one of up to 9 uniquely coloured and shaped symbols on an x-y plot,

6. Ternary diagrams,

7. Up to 9 data subsets as uniquely coloured and shaped symbols on a ternary diagram,

8. A fourth auxiliary variable as one of up to 9 uniquely coloured and shaped symbols on a ternary diagram,

9. A non-traditional diagram, the box-and-whisker plot,

10. Simple posting maps, using a rectangular co-ordinate scheme, where the value of an integer, floating point or qualified data variable may be displayed beside a cross indicating geographic position.

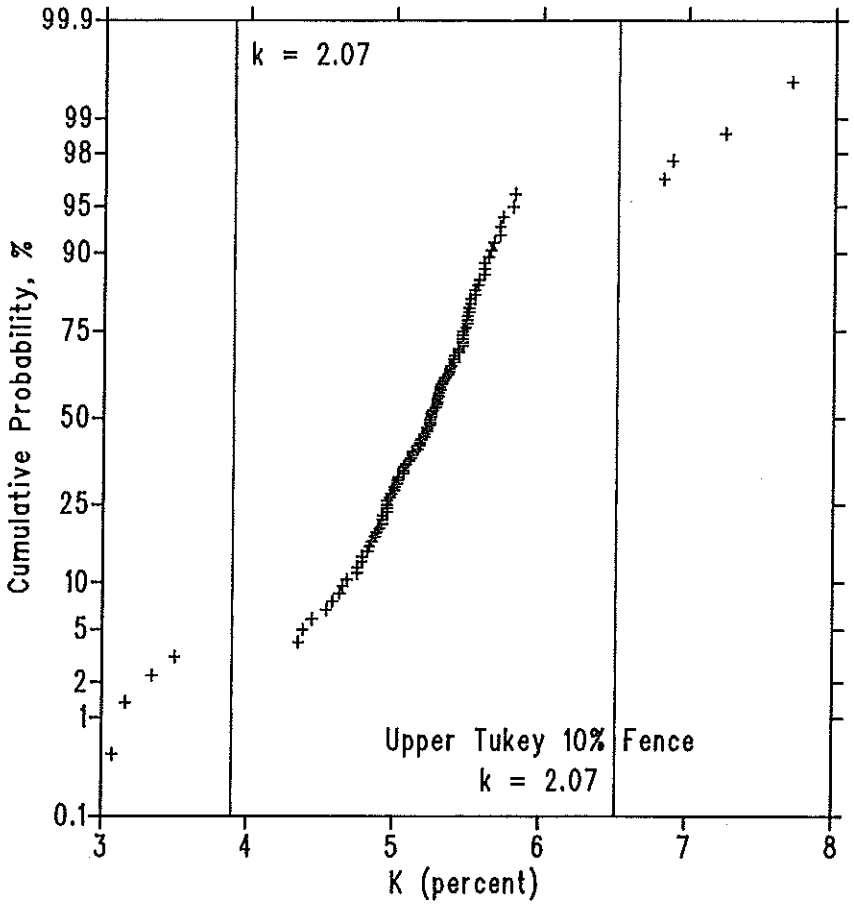
11. Up to 9 data subsets as uniquely coloured and shaped symbols on a simple map, and

12. A symbol map where the value of a numeric data item can be indicated by one of up to 9 uniquely coloured and shaped symbols.

In the auxiliary variable plots (5, 8 and 12) any data points outside of the symbolized range are indicated by a dot.

Some additional comments on the above procedures are presented below, where the numbers in parentheses refer to module numbers in the above list.

The graphical trimming of cumulative probability plots (2) permits outliers, at either the upper or lower extremes of the data, to be temporarily removed and identified by an up to 12 character "unique character identifiers". The percentiles and cumulative probability plots are then recomputed and displayed for inspection and possible futher trimming. To assist in the selection of trimming levels, "fences" at the approximate 10 % some-outside level may be added to the cumulative probability plots (Fig. 1) following Tukey's procedure (Hoaglin et al., 1986). The concept of a some-outside level (Hoaglin et al., 1986) in this application implies that up to 10 % of the data may plot beyond the "fences", 5 % at each extremity. If more than 5 % of the data fall beyond a "fence" outliers must be suspected. This trimming may be continued until an outlier free background population is isolated (Fig. 2), at which time a list of all the trimmed outliers may be displayed (Table 1) together with summary statistics for the remaining background data (Table 2). Up



Cut Off Value = 6.511

Points Trimmed = 4

Total Points Trimmed = 4

Total No. of Points = 110

Figure 1. Example of a cumulative probability plot. The fences correspond to a some-outside rate of 10%, i.e., 90% between the fences and 5% outside each fence. In this example the gaps between the central mass of data and the two sets of 4 outliers outside the fences are indicative of true outliers.

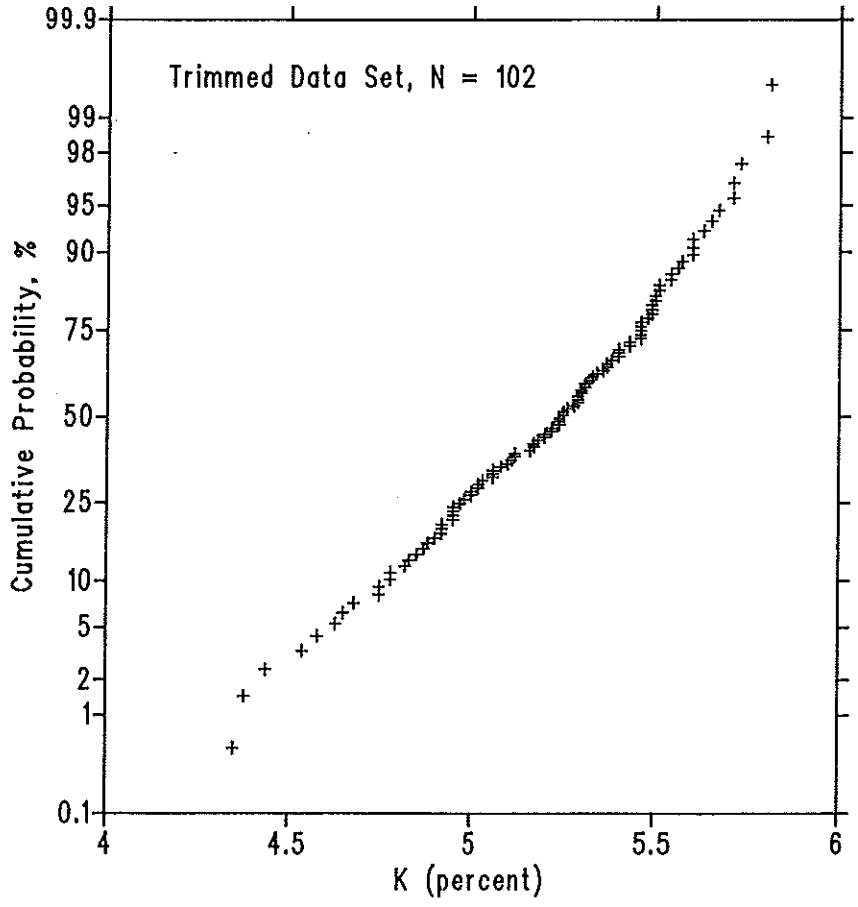


Figure 2. Example of a cumulative probability plot of trimmed data.

to 3 resulting groups, i.e., the lower-outlier, "core", and upper-outlier groups, may be selectively saved as subsets for future display or processing.

The histograms (1) may be annotated with 1 to 3 auxiliary variables. The means for these variables in the cases that fall within the histogram group of the primary variable are displayed to the side of the histogram (Fig. 3). This display has proven particularly useful in the study of the interrelationship of field observations to geochemical laboratory measurements. In the example (Fig. 3) the relative abundance of lakes less than 30 feet deep is clearly seen; but in addition information that indicates a decrease in Loss-on-Ignition and increase in Fe content with increasing depth is also simultaneously displayed. Although Mn generally behaves in a similar fashion to Fe it can also be seen that the sympathetic relationship with Loss-on-Ignition is less pronounced for Mn than if is for Fe in this particular instance.

Histograms, although a popular and widely-used display, lack objectivity by virtue of the arbitrary selection of the starting value, number of histogram groups and their width. In many instances box and whisker plots (9), which are based on the data minima, maxima, 5th, 25th, 50th, 75th and 95th percentiles, are more appropriate than histograms. In particular, they facilitate the graphical comparison of subsets for which data are available for the same variable (Fig. 4). Additionally, the box containing the range of the central 50 % of the data is notched to indicate the location of the median (50th percentile) and its 95 % confidence bounds. Where the notches for subsets do not overlap, the display provides a graphical distribution-free test indicating that the medians of the subsets may well be significantly different at the 95 % confidence level.

The auxiliary variable plots (5 and 8) permit the display of a trace element content in the context of major components, e.g., Zn versus the Fe and Mn in lake or stream sediments, or a trace element against Ca + Mg, Na + K and Fe + Ti in a ternary diagram for a differentiated series of igneous rocks. Both of these displays heve proyen useful. Figure 5 is an example where U (ppm) content has been plotted against the r1 (x) and r2 (y) differentiation indices (de la Roche et al., 1980). Each plotted point corresponds to a granitoid pluton, and it can be seen that the high U plutons fall into two clusters; one cluster corresponding to more felsic plutons and the other to syenitic bodies. It is worth noting that cation measures r1 and r2 were computed from the major element compositional data on a demand basis using the "define" command.

The spatial display modules (10-12) provide for rectangular co-ordinate systems, e.g., UTM or simple local grids; no provision has been made for geodetic co-ordinates or the projections required by them. In some instances of high data density on the posting display (10) the values may overplot on one another. To resolve this clutter problem a graphical editing procedure is provided so that the numeric values may be rearranged spatially so they do not overlap. As part of all the mapping modules (10-12) a facility has been implemented which allows a geological or topographic map to be placed on a digitizing tablet and be "slaved" to the map displayed on the screen. In this manner the position of any location of interest may be transferred to the screen as a highlighted open circle centred on the digitized position. Optionally, a marker can be left at this position for later annotation, or the target circle may be deleted.

In order to identify outliers the simple plot and the map posting procedures (3, 6 and 10) include a facility where up to eight points, selected from the screen display with the graphics cursor, may have their plotting symbol changed. The new unique symbol and "unique character identifier" of the selected point are displayed to the side of the plot. The example (Fig. 6) contains the same data and x-y axes as Figure 5, the outlier identification function has been used to identify the 1:250 000 NTS map sheet numbers and pluton field identifiers of the high U background syenitic plutons. If more than eight points are to be identified, the previous eight are set back to the default marker, and then additional sets of up to eight points may be identified.

To ease the task of preparing data subsets a graphical procedure has been implemented in the non-subset display routines, i.e., the x-y, ternary and map displays (3, 5, 6, 8, 10, 12). The user simply draws a polygon using the graphics cursor around the group of points to be included in a subset, this is followed by a prompt for the subset name and description. The subsets may be used for any future display or processing as the user desires.

Table 1. List of trimmed outliers. Note that the mean and standard deviation of the background (untrimmed) data are displayed. These are used in the computation of the SND (Standard Normal Deviate) values and their corresponding probabilities, which indicate the probability of an outlier being a member of the background population under the assumption of normality.

DATA SET NAME: ogdy.dat DATE: 2-JUL-87

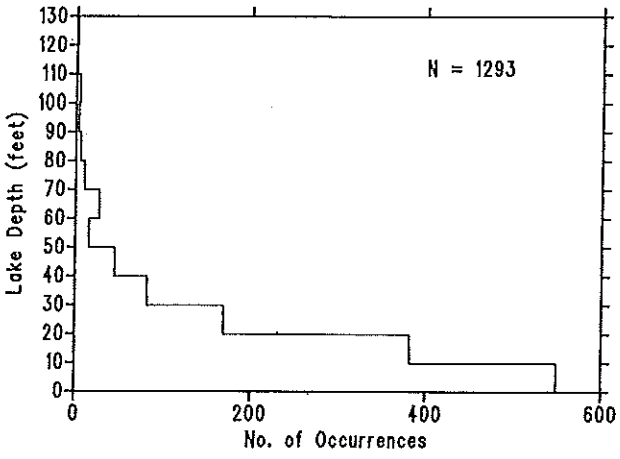
CORE DATA MEAN = 5.202 CORE DATA S.D. = 0.3266 SIZE = 102

TABLE OF OUTLIERS

NT	I	UNIQUE IDENTIFIER		VALUE	SND	PROB
1	69	105I	091	3.070	-6.530	0.00000
2	70	105I	092	3.160	-6.254	0.00000
3	68	105I	090	3.340	-5.703	0.00000
4	67	105I	089	3.500	-5.213	0.00000
5	81	105I	111	6.830	4.984	0.00000
6	82	105I	112	6.890	5.168	0.00000
7	79	105I	107	7.250	6.270	0.00000
8	80	105I	108	7.700	7.648	0.00000

Table 2. Summary Statistics for the graphically trimmed background population. Note that 2 estimates of the standard deviation are provided. One from the Interquartile Range, and the other via hte usual computation of mean and variance.

```
SUMMARY STATISTICS FOR: k

TABLE OF EMPIRICAL PERCENTILES          NUMBER OF OBSERVATIONS =  102

MAXIMUM VALUE                 5.8100
99.9th PERCENTILE             5.8090
  99th PERCENTILE             5.7993
  98th PERCENTILE             5.7296
  95th PERCENTILE             5.6690
  90th PERCENTILE             5.5970
  80th PERCENTILE             5.4900
  75th PERCENTILE             5.4600
  70th PERCENTILE             5.4000
  60th PERCENTILE             5.3100
  50th PERCENTILE             5.2450
  40th PERCENTILE             5.1640
  30th PERCENTILE             5.0230
  25th PERCENTILE             4.9725
  20th PERCENTILE             4.9260
  10th PERCENTILE             4.7800
   5th PERCENTILE             4.6310
   2nd PERCENTILE             4.4420
   1st PERCENTILE             4.3806
MINIMUM VALUE                 4.3500

MEDIAN =   5.2450     I-Q RANGE =  0.48750     S.D. =  0.36138
MEAN   =   5.2024     VARIANCE  =  0.10664     S.D. =  0.32655     CV% =   6.3
```



LOI (%)	Fe (%)	Mn (ppm)
11.00	3.500	3250.0
18.25	4.180	752.8
25.20	5.767	923.3
23.52	8.250	2195.0
23.56	3.895	713.1
18.77	4.800	2487.6
24.71	4.200	897.7
20.57	3.612	822.1
24.71	3.993	1232.1
25.79	3.549	694.7
29.05	2.752	457.3
39.59	1.907	341.8

Figure 3. Example of a histogram drawn by IDEAS and annotated with information for 3 auxilliary variables.

Univariate and bivariate statistics

A selection of statistical routines are available, including computations for.

1. Sumary statistics, e.g., range, mean, standard deviation, etc.,
2. Empirical percentiles,
3. Normality testing (3 tests),
4. Generalized power transformations (Box — Cox procedure),
5. One-way (two-level unbalanced) Analysis of Variance, both parametric and non-parametric Kruskal — Wallis procedures, for up to 9 groups, including homogeneity of variance tests (2),
6. Paired t-test,
7. Spearman rank correlation coefficients,
8. Pearson product-moment correlation coefficients, and
9. Simple linear regression.

The thrust of the univariate statistical and simple graphics modules is to facilitate graphical inspection and comparison of data in the spirit of what statisticians refer to as exploratory data analysis (EDA) (Turkey, 1977; Velleman and Hoaglin, 1981). Experience has shown that the trained or inquisitive eye studying graphical presentations is far more efficient at detecting patterns and outliers (anomalies) than statistical computations. Descriptive statistics will always have a place for the exploration geochemist and IDEAS generates those in common use. For confirmatory data analysis (CDA), or hypothesis testing, the normality tests and analysis of variance modules are provided.

In fact normality is less of a concern than many non-statisticians believe. Three normality tests are provided, the first two of which are commonly used in statistical studies;

they are the Shapiro — Wilk (Shapiro and Wilk, 1965), modified Anderson — Darling (Stephens, 1974) and Lin — Mudholkar tests (Nelson, 1983). However, it is not expected that the general user will or should make extensive use of these tests.

Analysis of variance (ANOVA) is useful dool for statistically determining if subset means are, or are not, significantly different (Anderson and McLean, 1974). In the context of this, the porperty of homogeneity of variance (homoscedasticity) is far more important than normality. Therefore two appropriate tests, the Bartlett (Anderson and McLean, 1974) and Levene (Levene, 1960) tests, are built right into the ANOVA module. The results of the ANOVA on the data subset means must be reviewed for credibility in tems of these homoscedasticity test results. In addition to the traditional ANOVA and computation of variance components (Table 3) a non-parametric (distribution-free) alternative, the Kruskal — Wallis test (Miller and Kahn, 1962), is undertaken. The ANOVA module also provides for the computation of subset means together with a multiple comparison following a procedure of Tukey's (Gill, 1978), Table 4. This is of assistance in identifying the actual subsets which have significantly different means from each other, as the ANOVA can only indicate that one or more means are significantly different from the remainder. Finally, in the case of a 2 group ANOVA with equal-sized samples a paired t-test may be undertaken if it is appropriate.

The paired t-test may be undertaken independently and is particularly appropriate if one wishes to determine if two procedures, perhaps concerning sample preparation and/or analysis, lead to similar or significantly different results when applied to a group of samples (Koch and Link, 1970). This procedure finds its most extensive use in IDEAS in studying the results of different analytical techniques applies to the same physical sample, e.g., gold determined by INAA versus gold by an acid leach and graphite furnace AAS procedure.

The inclusion of a module for determining optimal transformations to normality following the procedure of Box and Cox (Howarth and Earle, 1979) may not seem in the spirit of EDA. However, previous work has shown that use of this procedure on non-normal data after the user has removed any outliers or obvious members of other data populations can provide useful information. For instance, when an element is present dominantly in discrete mineral grains a Poisson distribution is likely; when a body of rock has undergone hydrothermal alteration a negative skew may develop; and where a process is time-dependent a reciprocal law may play an important role. The results of the optimal transform module can be effectively used to shed light on the term of the physical and chemical processes underlying the data distribution. In general the technique is used in this EDA mode rather than to determine the coefficients of the generalized power transform to fulfill the normality requirements of a CDA technique.

The provision of both Spearman and Pearson correlation coefficients permits the quantification of systematic monotonic non-linear pair-wise relationships as well as linear patterns. For this reason the Spearman coefficients are the default, as the presence of any systematic relationship is of interest, whether or not it is linear. Relationships can usually be linearized by an appropriate transformation if such is required. The Spearman coefficients are also more resistant to the effects of outliers, which can "lever" Pearson coefficients to unrealistically high values. In contrast, the Pearson coefficient is displayed in the x-y and bivariate regression plot modules, where various linearizing transforms would be inspected. The simple regression module is a natu-

Table 3. Example of an ANOVA table with the hypothesis test for equality of means and the computation of the variance components.

ONE-WAY ANOVA FOR VARIABLE: k

SOURCE OF VARIABILITY	SUM OF SQUARES	DEGREES OF FREEDOM	MEAN SQUARE	F-RATIO	SIGNIFICANCE
BETWEEN	2.299	2	1.149	3.12	0.9520 *
WITHIN	39.38	107	0.3680		
TOTAL	41.68	109	0.3823		

NS = NOT SIGNIFICANT, F < 0.95
* = F > 0.95 BUT < 0.99
** = F > 0.99 BUT < 0.999
*** = F > 0.999

SOURCE OF VARIABILITY	MEAN SQUARE	UNIT SIZE	VARIANCE COMPONENT	PERCENTAGE OF TOTAL	COEFFICIENTS	
BETWEEN	1.149	3	0.2798E-01	7.07	1.00	27.93
WITHIN	0.3680	110	0.3680	92.93	1.00	
TOTAL	0.3823		0.3960			

STD. ERROR OF THE MEAN = 0.1308

GRAND MEAN = 5.2035

APPROXIMATE (CONSERVATIVE) 95% CONFIDENCE LIMITS:	4.6485	5.7664
APPROXIMATE (RADICAL) 95% CONFIDENCE LIMITS:	4.9441	5.4628

Table 4. Example of the non-parametric Kruskal and Wallis statistic and the subset means display together with the matrix indicating significantly different means.

```
KRUSKAL & WALLIS NON-PARAMETRIC ANALYSIS OF VARIANCE

SUBSET    SIZE     SUM OF RANKS

alsk        6         223.5
grnt       36        2351.
grdr       68        3530.

KRUSKAL & WALLIS H STATISTIC =   6.1718
SIGNIFICANCE OF THE STATISTIC:   0.9543 *

DO YOU WISH TO DISPLAY THE SUBSET MEANS? y

SUBSET    SIZE       MEAN      STD. ERROR     95% CONFIDENCE BOUNDS

alsk        6       4.8783       0.2299        4.2873        5.4694
grnt       36       5.3942       0.1363        5.1177        5.6706
grdr       68       5.1312       0.5621E-01    5.0192        5.2432

TUKEY'S 95% CRITICAL DIFFERENCE =   0.44894

MATRIX OF SIGNIFICANT (95%) DIFFERENCES

   GROUP          1     2     3

3  grdr           -     -     #
2  grnt           *     #
1  alsk           #
```

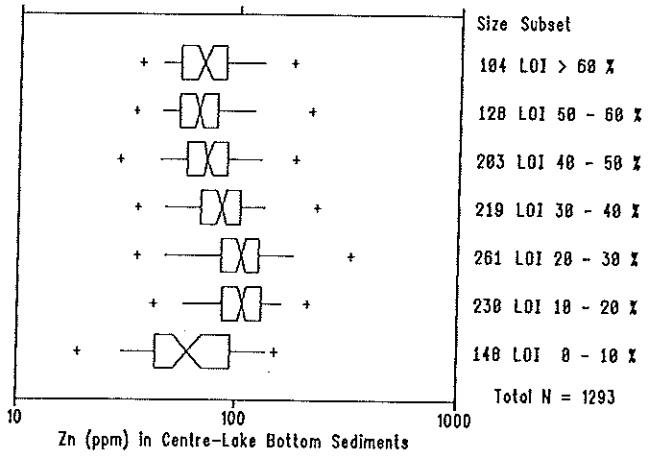


Figure 4. Example of a box and whisker plot drawn to illustrate the relationship between Loss-on-Ignition and Zn in centre-lake bottom sediments.

ral extension of the search for linearity. In addition to the regression coefficients a number of additional statistics are computed, hypothesis tests for the intercept being equal to zero and for the significance of the reduction in the sums of squares due to the regression are undertaken.

Multivariate statistics

Implementation of the multivariate statistical modules commenced in 1987. Currently there is an upper limit of 36 to the number of variables that can be processed in the multivariate procedures at any time. As with the limit on cases, i.e., 7 000, this was decided upon after discussions with the users; however, should more variables need to be processed simultaneously the required changes can be made quite easily.

To date one module is completed which undertakes a multivariate trimming (MVT) procedure employing the multivariate equivalent of a cumulative probability plot. The method is based on plotting the ranked Mahalanobis distances (c.f., univariate standard normal deviates) against Chi-square (Fig. 7), with degrees of freedom equal to the number of variables (Gnanadesikan, 1977). Initially after a summary statistics display (Table 5) a number of trial inspection plots, where arbitrary fractions of the most extreme data are temporarily trimmed and the Mahalanobis distances for all cases recomputed, are displayed (Fig. 8) in order to gain an insight on the data structure and an idea of how many gross outliers may exist (Garrett et al., 1982). Once a satisfactory starting plot is selected the extreme members, outliers, are graphically trimmed with the cursor and their "unique character identifiers" displayed. The univariate summary statistics and the correlation matrix for the trimmed data set are then presented with a new Chi-square plot. Trimming continues iteratively until the plot appears linear with a homogeneous distribution of points (Fig. 9). Finally, a list of all the trimmed outliers (Table 6) may be displayed, together with the summary statistics of the remaining "core" data (Table 7). As with the univariate probability plots the outliers and "core" data groups may be selectively saved as subsets for future display or processing. In the geochemical context, the "core" represents a multivariate background population, and the outliers are prime candidates for detailed interpretation in terms of mineral occurrences or other features of interest. The procedure is useful in determining the extreme members of a multivariate data set. By treating all variables simultaneously the "multiple jeopardy" problem is avoided, where accumulating the top few percent of each variable separately for detailed inspection usually leads to a disproportionately large list of candidates for further study. The MVT procedure is also a useful precursor to other multivariate techniques

where outliers could seriously distort the analysis, or non-multivariate-normality would contradict the assumptions of the techniques to be used. As such it is an effective exploratory data analysis tool in the multivariate domain.

PROVISIONS FOR HARDCOPY

A major feature of IDEAS is that through the use of GKS the graphical output may be edited, i.e., titles and labels changed, and additional annotative text added. A Tektronix multi-pen plotter allows camera-ready publication quality black line diagrams to be prepared directly during an IDEAS session, as has been done for this report. Currently the various colours on the video display are mapped to different pen widths. It was the need for publication quality black line diagrams that led to the decision to limit a number of the graphic displays to 9 data subsets or concentration levels as 10 unique marker types are available under GKS, and one is not suitable as a plot symbol. Alternately, the colour display on the screen may be dumped to an ink-jet plotter; this output is suitable for inspection and interpretation, or may be used in poster presentations.

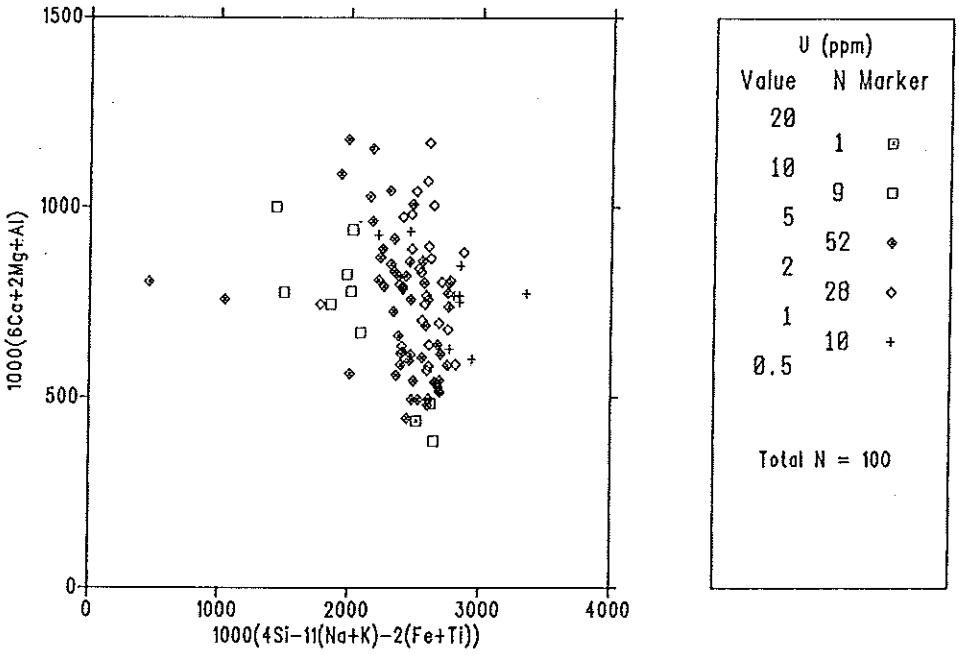


Figure 5. Example of an x-y plot coded by the value of an auxilliary variable.

Table 5. Initial Summary Statistics display of the multivariate data.

```
 5. ca
 6. na
 7. k
 8. ti
 9. mn
10. ba
11.

DATA SET NAME: ogdy.dat                                   DATE:  6-JUL-87

                MULTIVARIATE OUTLIER INVESTIGATION

        ITERATION: 0     POPULATION SIZE = 110     TRIM % = 0.0

  VARIABLE          MEAN          VARIANCE        STD.DEV          CV %

  1 si             30.74            3.627          1.904          6.20
  2 al             8.119           0.2550         0.5050          6.22
  3 fe             2.869           0.6144         0.7839         27.32
  4 mg             1.356           0.3174         0.5634         41.56
  5 ca             2.887           0.6086         0.7801         27.02
  6 na             1.687       0.3134E-01         0.1770         10.50
  7 k              5.203           0.3823         0.6183         11.88
  8 ti             3540.        0.8979E+06          947.6         26.76
  9 mn             645.0        0.2873E+05          169.5         26.28
 10 ba             984.2        0.8317E+05          288.4         29.30
```

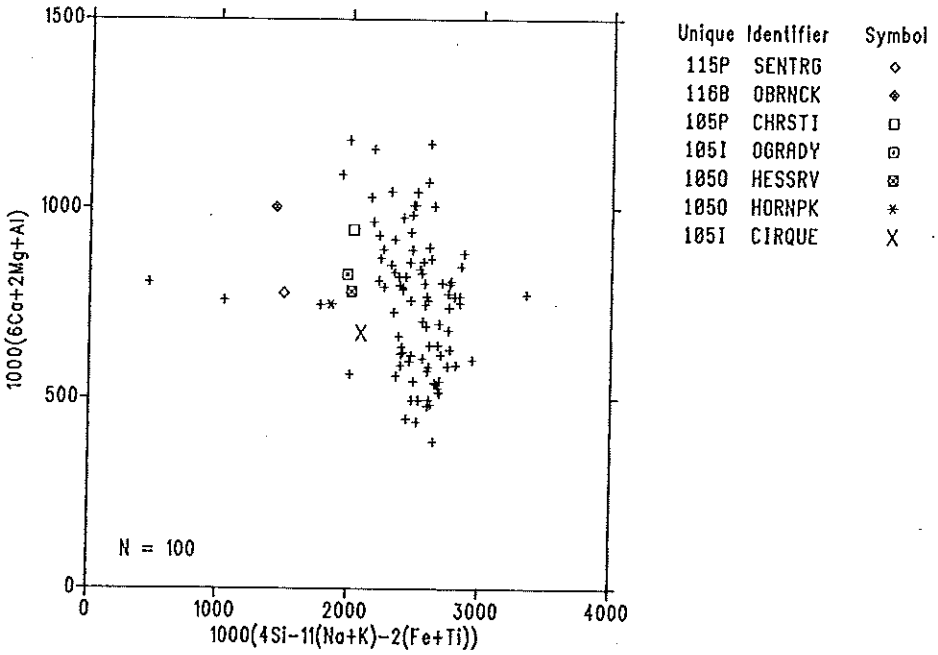


Figure 6. Example of an x-y plot labelled with the unique character identifiers via changes in plot symbology.

FUTURE DEVELOPMENTS

Future graphical developments call for the implementation of a "zoom-in" feature on the x-y plot and map displays (3-5, 10-12). The user will be prompted for the lower-left and upper-right corners of the area to be redisplayed. After determining graphically acceptable plot limits the area will be redisplayed with all appropriate title, axis labelling and legend information.

Two additional database related modules are planned. The first permits the logical combination of pre-existing subsets through the use of Boolean operators. The second permits variables to be conditionally recoded, this facility is particularly useful when working with multiple geochemical thresholds dependent on categorical field data, e.g., lithology.

In the realm of simple statistical procedures, a 2-dimensional contingency table module will be added to IDEAS. This tool it particularly useful for studying categorical and coded data, e.g., field data.

It is planned to prepare a number of additional multivariate graphical and statistical modules for inclusion in IDEAS. These include a canonical variable module, 2 cluster analysis procedures (one graphical and the other statistical), a principal component (factor) analysis module, a multilinear regression package, a logistic discriminant module, and finally, an empirical classification procedure which will include the ability to undertake multivariate analysis of variance.

In November, 1987 the Geological Survey of Canada upgraded its DEC VAX 11/780 to a VAX 8700. Both machines run the VMS Operating System and no conversion problems were encountered. Perhaps of greater interest are the possibilities for migrating the IDEAS software to smaller computers so as to provide a stand-alone or networked workstation. DEC offers its MicroVAX family of super-microcomputers, i.e., the II and 2000 machines. However, whereas this solution is appropriate in a regional or small office where several people or projects can share the cost of the computing resources, it tends to be too costly for a single user workstation. The obvious advantage of the above solution is the availability of VAX Fortran and the VMS operating system for the MicroVAXes. An alternative would be to use one of the new 32 bit super-microcomputers using the Intel 80386 or Motorola 68020 microprocessors. If this was done the

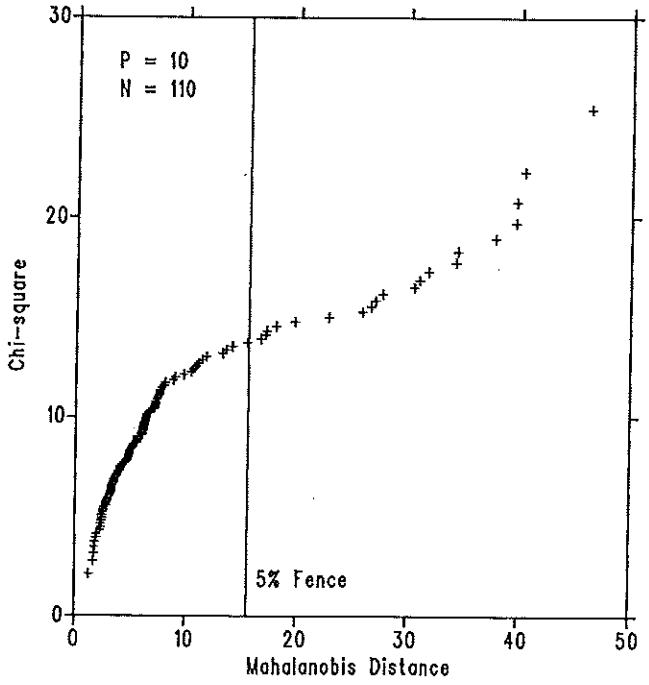
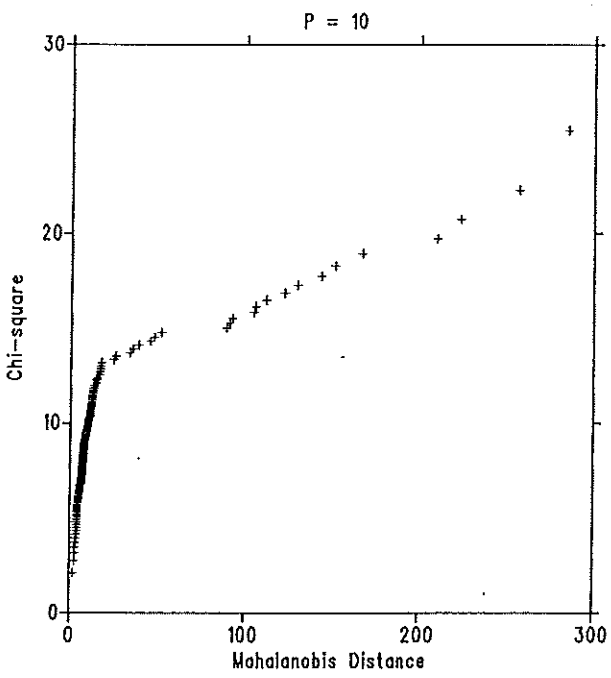


Figure 7. Example of a Chi-square (multivariate probability) plot. The fence corresponds to a some-outside rate of 5%. In this example 20 points lie above the fence, 14 in excess of the expected 6, which is indicative that some true outliers exist.



MVT % entered = 15.0

Cases trimmed = 17

Actual trim % = 15.5

Unique Identifier		D-sq
105I	108	105.8
105I	044	104.6
105I	015	92.4
105I	016	90.8
105I	112	89.0
105I	060	52.0
105I	114	48.0

Figure 8. Example of a multivariate trim robust start trial. Note the unique character identifiers of the most extreme members of the data set displayed by the plot.

Table 6. List of multivariately trimmed outliers. Note that the Mahalanobis distances (D-SQ) are with respect to the final kernel subset means and covariances (correlations). The H (I) column is a measure of leverage which can be computed from the Mahalanobis distance (Velleman and Welsch, 1981) and indicates the extent to which that outlier would perturb a linear model. The probability (DPROB) is the Chi-square based estimate of group membership in the kernel subset, in the example all the probabilities are less than 0.0005%.

DATA SET NAME: ogdy.dat DATE: 8-JUL-87

TABLE OF OUTLIERS

CRITICAL PROBABILITY FOR TRIMMING = 0.0087

NT	I	UNIQUE IDENTIFIER		D-SQ	H(I)	DPROB
1	80	105I	108	543.7	4.998	0.00000
2	68	105I	090	502.6	4.620	0.00000
3	70	105I	092	446.3	4.104	0.00000
4	79	105I	107	442.2	4.066	0.00000
5	21	105I	037	332.1	3.056	0.00000
6	22	105I	038	315.8	2.907	0.00000
7	69	105I	091	275.7	2.538	0.00000
8	67	105I	089	253.6	2.336	0.00000
9	81	105I	111	251.3	2.314	0.00000
10	74	105I	100	213.0	1.963	0.00000
11	1	105I	015	209.7	1.933	0.00000
12	2	105I	016	199.1	1.836	0.00000
13	40	105I	060	153.7	1.420	0.00000
14	24	105I	044	151.5	1.399	0.00000
15	39	105I	059	132.0	1.220	0.00000
16	82	105I	112	131.0	1.211	0.00000
17	8	105I	022	117.0	1.082	0.00000
18	42	105I	062	107.0	0.9910	0.00000
19	106	105I	138	80.37	0.7464	0.00000
20	73	105I	099	77.45	0.7196	0.00000

Table 7. Summary Statistics for the graphically trimmed multivariate kernel subset. Note the marked reduction, by in excess of 50% in some cases, in the coefficients of variation (cv%) due dominantly to decreases in the standard deviations with the removal of outliers.

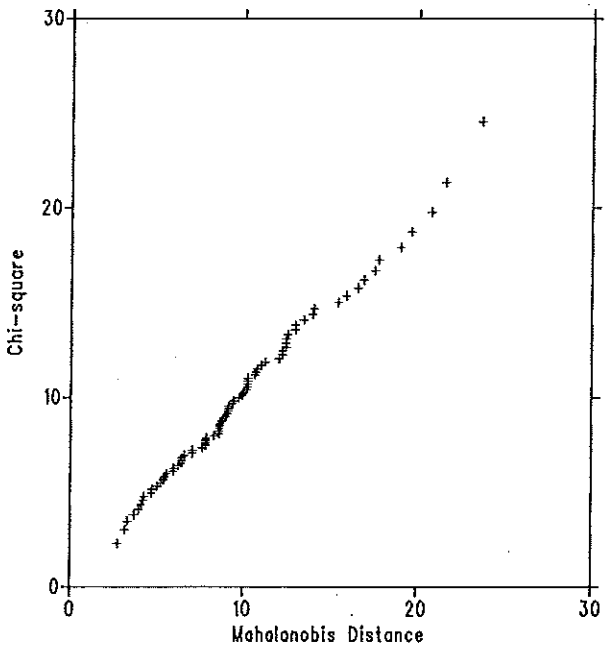
```
MOVE CURSOR TO NOTE STARTING POSITION AND PRESS A (PUCK) KEY
IS THE NOTE AND ITS POSITION CORRECT?

HAVE YOU FINISHED ADDING NOTES?


ENTER A <CR> TO CLEAR THE SCREEN AND CONTINUE
```

DATA SET NAME: ogdy.dat DATE: 8-JUL-87

MULTIVARIATE OUTLIER INVESTIGATION

ITERATION: 4 POPULATION SIZE = 80 TRIM % = 27.3

VARIABLE	MEAN	VARIANCE	STD.DEV	CV %
1 si	30.58	1.368	1.169	3.82
2 al	8.150	0.1617	0.4021	4.93
3 fe	2.976	0.1786	0.4227	14.20
4 mg	1.359	0.1446	0.3803	27.99
5 ca	2.979	0.1898	0.4356	14.62
6 na	1.654	0.6437E-02	0.8023E-01	4.85
7 k	5.253	0.6425E-01	0.2535	4.83
8 ti	3632.	0.2220E+06	471.2	12.97
9 mn	678.8	6267.	79.16	11.66
10 ba	1030.	0.2433E+05	156.0	15.15



Iteration No. = 4

No. of cases = 80

Cases trimmed = 30

Final trim % = 27.3

Figure 9. Example of a chi-square plot for a multivariately trimmed data set.

move to a UNIX, or UNIX-like, operating system might be made. Although such a move involves more effort it increases software portability and machine/vendor independence, which was one reason GKS was selected as the graphics software package.

Current plans call for the continued development of IDEAS until it contains all the features identified in the functional specification, and those others required by the Subdivision at this time. In parallel, a number of other software packages are being, or will be, evaluated or developed in the Subdivision. IDEAS will be integrated with these to provide a unified computing environment. Following a period, probably of one year, of extensive testing the final version of IDEAS will be released to the geoscience community. Several procedures for such a release are available, the two extremes being simple open-filing at a nominal cost, or release, possibly on a widely available microcomputer family, by a software house.

DISCUSSION

IDEAS is meeting the majority of the Geological Survey of Canada's exploration geochemist's needs, and those of other geoscientists with spatially distributed point data who are interested in detecting outliers (anomalies) and identifying and interpreting the underlying patterns of statistical and spatial variation in their data. IDEAS is considered to be an evolving system, the underlying structure facilitates the addition of further modules. As user experience grows, modifications and new features are being suggested and being integrated into the system as appropriate.

At this time the major use of IDEAS has been with regional geochemical and detailed study data derived from Federal — Provincial Mineral Development Agreement (MDA) funded surveys. Usage is now being extended to support other Geological Survey geochemical R & D projects and overseas CIDA surveys. To this end an additional Tektronix terminal, digitizing tablet and ink-jet plotter have been purchased so that "production" and "development" can continue without mutual interference.

To date user response has been fabourable, and the creation of IDEAS databases has been facilitated with the implementation of a software package that automatically generates an IDEAS data definition file from the Geochemical Information Services System (GISS) file description stored archivally with the data for all Subdivision surveys on the Departmental CDC Cyber facilities. Other users are compiling their databases on MS-DOS based microcomputers in their offices employing commercial software packages such as dBASE III and Rbase System V and then using the public domain data communications package KERMIT to transfer their data to the Geological Survey VAX for IDEAS use.

ACKNOWLEDGMENTS

A large group of people have been involved with the IDEAS project since its inception in 1979. During the Hickling — Smith Unsolicited Proposal (1980-81) T.I. Goss, G. Jackson and J. Nash all played important roles in preparing the functional specification. The U.S. Geological Survey made the VAX version of GRASP available and permitted R.W. Bowen to install it and provide valuable assistance in system familiarization in 1984. Between 1985 and 1987 a series of University of Waterloo Co-op program students worked on various aspects of IDEAS. C.E. Lee, M.S. Pearson, C.D. Lichtenfeld, F. Khalily Araghy and W.A. Quesnel all made significant contributions to the developments of IDEAS. Lastly, thanks are due to D.J. Ellwood of the GSC for his assistance and advice on many matters since the inception of the project. The role of management in the success of this long term project must be acknowledged. Progressing from a concept to a usable system has been a lengthy process, without their continued support over the years the project would have been stillborn.

REFERENCES

Anderson, V.L. and McLean, R.A.
1974: Design of Experiments — A realistic approach; Marcel Dekker, New York, 418 p.

Bowen, R.W. and Botbol, J.M.
1975: The Geological Retrieval and Synopsis Program (GRASP); U.S. Geological Survey, Professional Paper 966, 87 p.

de la Roche, H., Leterrier, J., Grandclaude, P., and Marchal, M.
1980: A classification of volcanic and plutonic rocks using r_1r_2-diagram and major element analyses; its relationships with current nomenclature; Chemical Geology, v. 29 (3-4), p. 183-210.

Enderle, G., Kansy, K., and Pfatt, G.
1984: Computer Graphics Programming: GKS — The Graphics Standard; Springer Verlag, New York, 542 p.

Garrett, R.G., Goss, T.I., and Poirier, P.R.
1982: Multivariate outlier detection — an application to robust regression in the earth sciences; Joint Statistical Meetings of the American Statistical Association, Cincinnati, Ohio, 1982. Abstracts, p. 101.

Gill, J.L.
1978: Design and Analysis of Experiments in the Animal and Medical Sciences; Iowa State University, Press, Ames, v. 1, 409 p.

Gnanadesikan, R.
1977: Methods for Statistical Data Analysis of Multivariate Observations; John Wiley, New York, 311 p.

Hoaglin, D.C., Inglewicz, B., and Tukey, J.W.
1986: Performance of some resistant rules for outlier labelling; American Statistical Association, Journal, v. 81, no. 396, p. 991-999.

Howarth, R.J. and Earle, S.A.M.
1979: Application of a generalized power transformation to geochemical data; Methematical Geology, v. 11 (1), p. 45-62.

Koch, G.S. and Link, R.F.
1970: Statistical Analysis of Geological Data; John Wiley, New York, v. 1, 375 p.

Levene, H.
1960: Robust tests for equality of variances; *in* Contributions to Probability and Statistics, ed. I. Olkin; Stanford University Press, Palo Alto, California, p. 278-292.

Miller, R.L. and Kahn, J.S.
1962: Statistical Analysis in the Geological Sciences; John Wiley, New York, 483 p.

Nelson, B.B.
1983: Testing for normality; Journal of Quality Technology, v. 15 (3), p. 141-143.

Shapiro, W.W. and Wilk, M.B.
1965: An analysis-of variance test for normality (complete samples); Biometrika, v. 52 (3-4), p. 591-611.

Stephens, M.A.
1974: EDF statistics for goodness of fit and some comparisons; American Statistical Association, Journal, v. 69, no. 347, p. 730-737.

Tukey, J.W.
1977: Exploratory Data Analysis; Addison — Wesley, Reading, Massachusetts, 688 p.

Velleman, P.F. and Hoaglin, D.C.
1981: Applications, Basics and Computing of Exploratory Data Analysis; Duxbury Press, Boston, Massachusetts, 354 p.

Velleman, P.F. and Welsch, R.E.
1981: Efficient computing of regression diagnostics; American Statistican, v. 35 (4), p. 234-242.