copy 3

COMMONWEALTH OF AUSTRALIA

DEPARTMENT OF NATIONAL DEVELOPMENT
BUREAU OF MINERAL RESOURCES
GEOLOGY AND GEOPHYSICS

RECORDS:

1967/1

THE USE OF COMPUTERS FOR THE STORAGE AND ANALYSIS OF GEOLOGIC DATA

by

T. Quinlan

The information contained in this report has been obtained by the Department of National Development, as part of the policy of the Commonwealth Government, to assist in the exploration and development of mineral resources. It may not be published in any form or used in a company prospectus without the permission in writing of the Director, Bureau of Mineral Resources, Geology and Geophysics.

15928/65

THE USE OF COMPUTERS FOR THE STORAGE AND ANALYSIS OF GEOLOGIC DATA

RECORD 1967/1

by

T. Quinlan

CONTENTS

THE USE OF COMPUTERS FOR THE STORAGE AND ANALYSIS OF GEOLOGIC DATA

RECORD 1967/1

by

T. Quinlan

INTRODUCTION

Approximately 3 months were spent in the United States of America and Canada to

(i) gain experience in the development and operation of computer applications in geology.

(ii) determine what categories of geological data are suitable for inclusion in a geological data processing system.

(iii) study the methods by which these data may best be analysed and displayed.

The following organizations and companies were visited:

5th July	Museum of Palaeontology, University of California, Berkely, California.
8th July	T.R.W. Systems, Redondo Beach, Los Angeles, California.
11th July	Shell Oil Coy., Houston, Texas.
13th July to 30th September	Branch of Geochemical Census, U.S. Geological Survey, Denver, Colorado.
24th August	Petroleum Information Corp., Denver, Colorado.
12th September	Marathon Research Centre, Denver, Colorado.
3rd October	Department of Geology, North Western University, Evanston, Illinois.
4th October to 6th October	Geological Survey of Canada, Ottawa.

During the period in Denver arrangements were made by geologists of the U.S. Geological Survey for me to participate in a programme to sample the Fountain Formation in the Rocky Mountains, in the statistical analysis of the results of analyses of these samples obtained from a Direct Reading Spectrograph, in the statistical analysis of pebble count data on the Fountain Formation, and the use of non polynomial terms for Trend Surface Analysis. The STPAC system of Computer programmes which has been developed by the U.S. Geological Survey was used for this work.

Geological Use of Computers

The use which is made of computers by geologists in North America has not changed significantly in the period since the review made by Garrison and Krumbein (1963). They concluded that 'computers are being utilized in classical earth sciences fields in only a few dozen institutions', and within these by only a small proportion of the staff. In addition they found that only 'rarely do users take advantage of the full spectrum of possible computer applications'.

This situation will not change until:

(a) data processing systems for the storage and retrieval of geological data have been implemented and

(b) suitable systems of computer programmes for the analysis and interpretation of geological data are available.

The development of each of these two facilities can only be undertaken by people with some training in geology and in the use of computers, and each represents a substantial investment of manpower and money by an organization.

Some of the features of the two will be discussed in turn, in the light of the experience gained by geological organizations in North America.

AUTOMATIC DATA PROCESSING SYSTEMS FOR GEOLOGICAL DATA

One of the results of the 'Information Explosion' is that rigid standards must be adopted for the storage of geological data, whether the system be one which uses technical files, index cards, edge punched cards, or 80 column punched cards. These may be irksome to maintain but they are required if the system is to function. An Automatic Data Processing System, in which it is assumed that a computer is used to perform the necessary house keeping arrangements, requires standards which are even more inflexible.

As a result individual organisations have concentrated on the development of limited ADP Systems to meet their specific needs. On the basis of their experience, some of these organizations now advocate the adoption of common standards and even systems, providing that theirs are the ones which will be used. Nevertheless most design their own, rather than adopting an existing one. This is largely because of their own particular requirements, the limitations of the hardware which is available, and a desire to improve the existing designs.

In spite of this trend some rationalization appears to be necessary, as the exchange of data between geologists within an organization can be hampered by semantic difficulties, differences of opinion on nomenclature. This becomes a distinct problem when the data is stored in an ADP System, because of its rigid and mandatory standards and it is one which is magnified when the exchange of data is between organizations. The Committee on the Storage, Retrieval and Automatic Processing of Geochemical Data of the International Union of Geological Sciences may provide useful guidelines for a solution to this problem.

Two recent developments are the use of free field input of data, with each field preceded by a variable identifier, and the development of the software package INFOL by Control Data Corporation. The latter is a general purpose information and data processing programme which will accept free field input, and which will handle any type of file. These should allow some reduction in the investment required to establish an ADP System, and some relaxation in the restrictions of punched card formats.

Automatic Data Processing Systems have been implemented in five areas of the science:

(a) Museum Indexing

(b) The maintenance of Bibliographies

(c) the storage of hydrological data

(d) the storage of geological and geochemical data

Numerous items of numeric data and information are currently being recorded from one locality for inclusion in a Mineral Resources Index, or relating to Petroleum Exploration Wells or stratigraphic sections, and more than one hundred 80-column punched cards are commonly required to record this information in an ADP System. A more suitable medium is a document or well ticket of page size, which is completed with an electric typewriter with a particular font. Magnetic tapes containing edited data are prepared from these documents by an optical page reader, which is linked to a small computer. Control Data Australia can provide this service at a cost of $40 per hour. An additional advantage is that data, which is in a computer acceptable but conventional form, can be cheaply stored and used by existing manual systems, without the obligation to incorporate it in an ADP System.

Museum Indexing

The Museum of Palaeontology of the University of California at Berkely and the Denver Research Center of the Marathon Oil Company use files of 80-column punched cards to index their collections.

The palaeontological collection of the Museum of Palaeontology is indexed with a format which covers two cards, which are cross referenced by a museum accession number. The first card contains the geographic and stratigraphic location together with the age which has been assigned to the fauna. The second card is used to record the major faunal elements, their generic and specific identification, the name of the collector and the date. Numeric codes are used to facilitate sorting on the basis of age, locality, and taxonomy.

Three 80-column punched cards with an intricate format are used by geologists of the Marathon Oil Company to index the samples, cores and cuttings which they have in store. In addition to recording the geographic position, age and character of each sample, provision has been made on the third card to record the type of laboratory work which has been undertaken. This index has proved useful to find material which is suitable for a particular research project.

The Maintenance of Bibliographies

The United States Geological Survey now make use of an ADP System to produce the Bibliography of North American Literature on a monthly basis. Each entry consists of a reference to the author, title, and publication, a limited number of key words which are used for indexing, and a maximum of twelve lines for an abstract. Entries are prepared by appointed specialists in the subject matter.

The Water Resources Center of Cornell University now publish annually a Permutated Title Index, sometimes called a KIWC (Key Word In Context) index, of publications in Hydrology. This index is available on a subscription basis.

The Storage of Hydrological Data

In 1964 a committee from the Water Resources Division of the United States Geological Survey designed formats for 15 80-column punched cards and an aperture card for the storage of hydrological data. The list of items recorded in the system is comprehensive, and includes hydrologic details of aquifers and springs, physical and hydrological properties of soils and rocks, the results of chemical and spectrographic analyses of surface and groundwater water levels, and water use. Individual cards are cross referenced by the latitude and longitude of bores, or by the identification number of a surface water station. Provision was made for the data to be coded directly onto the 80-column card on which it is later punched.

The Storage of Data From Petroleum Exploration Wells

The earliest attempts to use the facilities of Automatic Data Processing in geology were for the storage of this type of data. Some of these systems failed, but the experience gained was used to refine and develop the art, so that the records on magnetic tape are now the source of basic data for most exploration programs.

Six Well Data Systems are now in operation (Dillon, 1964), and together they cover much of the area of the United States. A number of them have been established by Petroleum Information Corporation as a business venture or as contractor to a cooperative group of Petroleum Exploration Companies. Use is made of 99 card formats for their system, of which 92 may be required for an individual well, together with the required number of continuation cards. A comprehensive editing programme is used to ensure that the factual data is correct. No attempt is made to adjudicate in those cases where different sources provide conflicting interpretations of formation tops and lithologies, all are included and identified so that the user has the choice of which to accept for his purpose.

The list of items for which information is included is extremely comprehensive, and it would serve as a good model for an Australian system. It is believed that the American Stratigraphic Company will adopt a similar format to place stratigraphic logs in magnetic tape files.

The Storage of Geological and Geochemical Data

The United States Geological Survey and the Geological Survey of Canada have created files of 80-column punched cards to serve as a storage and retrieval systems for geological and geochemical data. They contain the results of chemical and spectrographic analyses, and some of the physical properties of the samples, and the results of radioactive age determinations by the Canadian Geological Survey. Both organizations are preparing to use magnetic tape in place of punched cards as a storage medium.

The system used by the Bureau is basically similar to both the Canadian and the U.S.G.S. systems and it would appear to be adequate for our present needs.

THE ANALYSIS AND INTERPRETATION OF GEOLOGICAL DATA

A computer should be considered as merely another tool which can be used in geology, and as such it can not be used independently of geological reasoning. The user takes advantage of the machines ability to perform a large number of operations in a short time, to undertake more sophisticated methods of analysis than have been used in the past. As a consequence the geological user must be capable of a comparable standard of logical reasoning to be able to produce a rational geological interpretation of the results.

The basic operations of a computer are those of the addition and subtraction of numbers, and certain logical operations based on the values of numbers. These are used, in suitable combinations, to perform the more intricate operations (such as multiplication, division, and inter-register transmission) which are available in a machines repertoire, which in turn is used to execute statements written in a programming language, an example of this is the statement

X = LOGF (Y),

which should result in the value of the natural logarithm of the number represented by Y being given to X.

Thus it is necessary to express a geological problem in terms of numbers, that is to provide a numerical approach. In addition it is possible to represent characters and words as numbers by using an appropriate convention and in some cases to manipulate these once the basic rules are formulated. Initially computer methods should be used in those situations which can be described numerically, as these are much more straight forward to deal with, and the techniques which have been developed in other sciences (notably statistics) can be readily adapted and used to advantage.

It is assumed that one of the prime aims of geology is to determine:

(i) the occurrence of the constituents of rocks in the earth's crust,

(ii) the relationship which may exist between these various constituents, and

(iii) the areal distribution of these constituents, which may be conveniently displayed in diagrams and maps.

In this context the term constituent is used to include chemical elements, minerals, fossils, and contained fluids. A general term is required as the numerical techniques for analysis are common to many variables regardless of their type.

Estimates of Abundance

A suitable measure for the occurrence of the constituents of a rock is the expected value or concentration which can be expected from determinations made on large numbers of samples. Using the concepts of probability, it can be stated that if 100 samples are used, the concentration determined will be greater than the expected value for 50 samples, and will be less than or equal to it for the other 50. The Arithmetic mean is a convenient estimator for the expected value, in those cases where the frequency distribution curve is symmetrical. The efficiency of this estimate is low if the curve is asymmetric, as is commonly the case.

Moreover much of the statistical theory and many of the techniques which are available cannot be used directly. This is because they are based on the assumption that the frequency distributions are symmetric and can be approximated by the normal law. It is customary to apply one of a number of transformations to the basic data so that these assumptions are satisfied. These considerations are of particular economic importance for estimates of grade from assay values in the gold mines of South Africa (Krige, 1960; Sichel, 1952).

Computer programmes are available in the USGS STPAC System to:

(i) substitute appropriate values in place of those which could not be determined,

(ii) apply designated transformations to observed variables,

(iii) calculate the mean, standard deviation and the Fisher K statistics of variables,

(iv) produce histograms and contingency tables of selected variables,

(v) to calculate the standard errors of replicate determinations,

(vi) to estimate variance components within the limitations of the sampling design which was used.

Relation Between Rock Constituents

The Multivariate techniques which have been developed by statisticians can be used to develop hypothetical relations between rock constituents. Some tests of significance are available if the data has been obtained in a probabilistic context, but the final test should be based on geological reasoning. Suitable techniques are:

(1) Numerical Taxonomy, for the classification of 'individuals' which are characterized by certain attributes, into a number of groups in such a way that all members of a group are 'similar', and all groups are dissimilar. Lance and Williams (1966) have developed programmes which provide the facility for two sorting strategies (nearest neighbour and centroid) and for four similarity coefficients (the correlation coefficient, the squared Euclidean distance, a non-metric coefficient and an information statistic). The technique has been applied to few geological situations.

(2) Discriminant Analysis, a procedure for estimating the position of an individual in relation to a boundary that best separates previously defined groups of individuals. In addition the discriminant function obtained by the analysis can be applied to additional data which is collected within the same context as the original data. It has been used to distinguish marine from fresh water shales (Potter, Shimp, and Withers, 1963), to establish tectonic settings of sandstones (Middleton, 1962) and volcanics (Chayes, 1964), to classify depositional environments of carbonates (Krumbein and Greybill, 1964), and to distinguish ore-bearing from barren sediments (Griffiths, 1957).

(3) Factor Analysis, is the term for a variety of procedures developed for the purpose of analyzing the intercorrelations within a set of variables, or rock constituents. One of these, the procedure for principal-components analysis, is useful to determine the minimum number of independent dimensions needed to account for most of the variance in the original data. This procedure may be used to estimate the minimum number of properties of rock constituents which must be measured to provide an optimum amount of information on the composition of rock bodies. A second procedure, that of Vector Analysis (Imbrie (1963), can be used to determine the number of end members needed to account for the compositional variation in the data, and to calculate the proportion of each sample which can be attributed to each end member. The techniques of Factor Analysis have been used by Imbrie (1963) as a tool for facies analysis, by Harbaugh and Demirmen (1964) to examine the petrologic variation in the Americas Limestone, by Imbrie and Van Andel (1964) to map heavy mineral provinces.

(4) Analysis of Regression, the technique of extracting from an array of data the main features of the relationships between rock constituents which are hidden or implied in the array, with due regard to Kendall's (1957) warning that this seems to assume a much more indulgent behaviour on nature's part than we have any right to expect. Examples of the applications of this technique are given by Griffiths (1957, 1958, 1961) in studies on the composition, texture and bulk properties of sediments; Krumbein (1959) who related the firmness of beach sand to several attributes; and Vistelius and Hurst (1964) to study the relation of phosphorous to constituents of granite.

Areal Distribution of Rock Constituents

It is possible to use an x-y plotter to present the results of the analysis of geological data which has been done with a digital computer. This may be done in two ways, machine contouring of the data, or the presentation of trend surfaces fitted to the data.

(1) Machine contouring, computer programmes are available for the construction of contours on regularly spaced or gridded data. Most geological observations are made at points which are irregularly spaced, and programmes which will interpolate to provide values at grid points would have to be developed before machine contours could be drawn.

(2) Trend Surface Analysis, is the procedure by which observations of a rock constituent or property associated with a particular point in space is divided into two or more parts, those which are associated with systematic changes from one edge of the space to another, and those apparently associated with non-systematic changes within small areas of the space. This type of analysis and presentation of data has recently become fashionable and applied to many situations. Examples of these are the work of Allen and Krumbein (1962) on the particle-size index of zircon in sedimentary rocks; the composition of granites by Whitten (1961); the analysis of geochemical data by Connor and Miesch (1964); and structural development in sedimentary basins by Merriam and Lippert (1966). Recent unpublished work by Connor and Miesch has demonstrated the value of using stepwise regression techniques to derive the regression equation and of the inclusion of non-polynomial terms into the equation.

CONCLUSIONS

1. The system which is used by the Bureau for recording geological and geochemical data on 80 column punched cards is adequate for our present needs.

2. It is to be expected that the introduction of an Automatic Data Processing System would be warranted when the size of the punched card files exceeds 100,000 cards.

3. There is a present need for the development of a system of computer programmes, suited to the CDC 3600 machine of the Computing Research Section of C.S.I.R.O., which are capable of undertaking the analysis and interpretation of geological data.

4. The STPAC system of programmes which has been developed by the United States Geological Survey is suitable, and could be adapted to our needs.

5. An understanding of the methods used for the analysis of data is necessary if the geological interpretation is to be meaningful and not spurious.

REFERENCES

ALLEN, P., and KRUMBEIN, W.C., 1962 - Secondary trend components in the top Ashdown pebble bed, a case history, J. Geol., v. 70, p. 507-538.

CHAYES, F., 1964 - A petrographic distinction between Cenozoic volcanics in and around the open ocean: Jour. Geophysical Res., v. 69, p. 1573-1588.

CONNOR, J.J., and MIESCH, 1964 - Analysis of geochemical prospecting data from the Rocky Range, Beaver County, Utah, p. D79-D83, U.S. Geol. Surv. Prof. Pap. 475D.

DILLON, E.D., 1964 - Electronic storage, retrieval, and processing of well data, Am. Assoc. Petroleum Geologists Bull., v. 48, p. 1828.

GARRISON, W.L., and KRUMBEIN, W.C., 1963 - Computer utilization in the environmental and earth sciences : a reconnaissance of status and need, Tech. Rep. No. 3, ONR Task No. 389-135, Office of Naval Research, Geography Branch.

GRIFFITHS, J.C., 1957, - Petrographic investigation of the Salt Wash sediments, Final Reports, U.S. Atomic Energy Comm., RME-3151, 38 p.

GRIFFITHS, J.C., 1958, Petrography and porosity of the Cow Run Sand, St. Marys, West Virginia; J. Sediment. Petrol., v. 28, p. 15-30.

GRIFFITHS, J.C., 1961 - Measurements of the properties of sediments, J. Geol., v. 69, p. 487.

HARBAUGH, J.W., and DEMIRMEN, F., 1964 - Application of factor analysis to petrologic variations of Americus Limestone (lower Permian), Kansas and Oklahoma : Kansas Geol. Survey Special Dist. Publ. 15, p. 40.

IMBRIE, J., 1963 - Factor and vector analysis programmes for analyzing geologic data, Tech. Rept. No. 6, ONR Task No. 389-135, Office of Naval Research, Geography Branch, 83 p.

IMBRIE, J., and VAN ANDEL, Tj. H., 1964 - Vector analysis of heavy-mineral data : Bull. Geol. Soc. Amer., v. 75, p. 1131-1156.

KENDALL, M.G., 1957 - "A course in multivariate analysis", Charles Griffin & Company, Ltd., London.

KRIGE, D.G., 1960 - On the departure of ore value distributions from the lognormal model in South African Gold Mines. J. South African Inst. Min. and Met., v. 61, p. 231.

KRUMBEIN, W.C., 1959 - The "sorting out" of geological variables illustrated by regression analysis of factors controlling beach firmness, J. Sediment. Petrol., v. 29, p. 575-587.

KRUMBEIN, W.C., and GREYBILL, F.A., 1965 - An introduction to statistical models in geology : McGraw-Hill, 475 p.

LANCE, G.N., and WILLIAMS, W.T., 1966 - Computer programmes for hierarchical polythetic classification ("similarity analyses") Computer Journal, v. 9, p. 60-64.

MIDDLETON, G.V., 1962 - A multivariate statistical technique applied to the study of sandstone composition : Trans. Royal Soc. Canada, v. 56, ser 3, sec. 3, p. 119.

MERRIAM, D.F., and LIPPERT, R.H., 1966 - Geological model studies using trend-surface analysis, J. Geol., v. 74, p. 344-357.

POTTER, P.E., SHIMP, N.F., and WITTERS, J., 1963 - Trace elements in marine and fresh-water argillaceous sediments : Geochim. et. Cosmochim. Acta, v. 27, p 669-694.

SICHEL, H.S., 1952 - New methods in the statistical evaluation of mine sampling data, Trans. Instn. Min. Metall., v. 61, p 261.

VISTELIUS, A.B., and HURST, V.J., 1964 - Phosphorus in granitic rocks of North America, Bull. Geol. Soc. Am., v. 75, p. 1055-1092.

WHITTEN, E.H.T., 1961 - Quantitative areal modal analysis of granitic complexes, Bull. Geol. Soc. Amer., v. 72, p 1331-1360.