Probability, Statistics & Computer Modelling in Geology

Alan Rogerson - Melbourne, Australia

1. Introduction

Modern geology was founded by James Hutton at the end of the 18th century as an observational, descriptive, qualitative, and subjective science. It was the pioneer work of Pearson (1901) and Spearman (1904) in statistics applied to agriculture, biology, medicine, and psychology, that led to the first major inroads of quantitative techniques in geology - and the models they used were above all statistical (Agterberg, 1974). The period of greatest growth of quantitative methods in geology occurred after 1950 when it was discovered that both statistical and multivariate techniques could be used on the computer and could therefore be applied for the first time to the enormous quantity of geological numerical data.

There are many ways of describing the modelling process itself, but as used in geology, it tends to follow the following stages (Rogerson, 1988 and 1989a,b):

(i) data acquisition/sampling;
(ii) data processing/analysis;
(iii) hypothesis testing;
(iv) prediction.

Not all of these stages are seen in the use of statistical models, but perhaps the most vital aspect of the quantitative approach is that it highly *conditions* the initial stage of data collection. We do not collect and organise data in a neutral or random way - the world is too full of information for this to be effective in the scientific sense and therefore we start from a pre-defined and narrow interpretation of those "facts" we are interested in collecting (Rogerson, 1989a). In other words, there is already an implicit *hypothesis* to be tested which the collected data helps us to resolve. Thus the above cycle could just as easily *start* with the hypothesis to be tested. Our concept of the hypothesis *defines* the relevant facts, which are then sought in the geological world.

2. Mathematical modelling in geology

The application of statistics and probability to geology in general serves one of two conceptually different purposes. Since real-world data is usually too numerous and/or too complex to be easily comprehended, we may use models to effectively *reduce* the information to a more representative or manageable form (a simple example is replacing a set of measurements by their mean and variance). On the other hand, models may be used for the opposite purpose: to apparently *increase* the data by using interpolative or extrapolative models on limited sets of data. For example, curve and surface fitting models are widely used in geology to replace a finite number of known values with an infinite (continuous) set, which may not even include any of the original data.

What both seemingly contrary uses have in common is that we have replaced original data with a *model* which is easier for us to understand, visualise, or manipulate, even though the model appears to carry less (or more) information. Both types of model occur frequently in statistics when we replace finite samples by their mean and variance (less information) or by a particular continuous distribution with the same mean and variance (more information). In a fundamental sense, the process of modelling becomes a delicate transference of information through numerous stages of data collection, analysis, and modelling, in which the analogy of a "modelling refinery" is probably more informative than, say, that of fitting together the pieces of a jigsaw.

In petrology, for example, a wide variety of statistical techniques is used to analyse the mineral or chemical composition of rock samples. An interesting ambiguity is mentioned in Le Maitre (1982, p.2) that sheds light on our concept of unbiassed sampling:

> "A 'sample' to a statistician is a number of objects taken from a population, whereas a 'sample' to a geologist is usually a single object or specimen ... To avoid bias the sample should be chosen at random, so that every object in the population has an equal chance of being collected, but ... The human tendency to collect oddities is revealed by the usual rock collection, which contains more strange than common rocks".

Once the rock samples have been collected and measurements made of their chemical composition, the process of statistical modelling begins. Calculating the mean and standard deviation for values of a particular chemical is useful in replacing the data with a central value and a measure of the deviation from that centre. These two statistics can then be useful for study and comparison with other samples from the same or different populations. From the data we can make observations about the variability in measurements, a factor which must be taken into account in all geological modelling. In particular, it is pointless to use highly sophisticated and delicate statistical models on data whose accuracy or reliability is low. This illustrates the important general principle of relating the data to its geological context at all stages in the modelling process. The wide variety of geological data leads to modelling using many kinds of probability distribution, including the binomial, poisson, and log-normal.

Another typical problem in geology is the visual or mathematical representation of a finite number of data points as a continuous curve or surface in space. When the

geologist collects information on the distribution of minerals, ores, or chemicals in the earth, this can be represented using three-dimensional coordinates: (x,y,z). We now try to find a function z = f(x,y) which is a "good" fit (or even, according to some criterion, a "best" fit) for the data points. Many geological situations produce coordinate data points for which it is useful to find interpolated or extrapolated values - the real problem is to decide which function f(x,y) should be used to fit the data points. Sophisticated statistical optimisation methods may therefore be needed (Davis, 1981). As there is an infinite range of possible functions, we are in effect hypothesis testing, and geologists usually resolve the best-fit problem by choosing some criterion of *minimal curvature*.

A typical example of this kind of modelling occurs in Merriam (1978, pp.135-155) who uses the computer to convert assay variation or drillhole data into a three-dimensional space-filling function. This may be imagined as a series of three-dimensional contours from which two-dimensional sections can be extracted by the computer. Firstly, the irregular data was located at three-dimensional coordinate points. A fixed mesh or grid of cubes was then generated by the computer to cover the volume concerned, and the data values transferred to this grid using a statistical relocation function. Following this a local smoothing function of 25 points was applied to interpolate unknown node values on the grid using a principle of minimum total curvature. Finally, another computer sub-routine was used to extract the required two-dimensional sections. It is clear that this extensive use of statistical techniques through the averaging, relocating, and smoothing functions is not possible without the aid of the computer to effect the calculations, and also to retain in its memory the final "space function model" of the original data. Nowadays many (if not most) statistical and probabilistic models used in geology are implemented using a computer.

3. Modelling using the computer

In geology, the computer is used when the data and/or the necessary calculations are numerous, when we need to control and manipulate a database, when we wish to apply complicated mathematical formulae, and when we wish to model using a simulation that must be repeated a large number of times (e.g. Monte Carlo methods). It is, however, one thing to realise that a geological problem can be solved using the computer, and another to find or write a program that will work. In practice, other criteria also enter the picture - the existence of ready-made programs, how the data can or should be coded, and whether it is financially feasible (which in turn brings in criteria of time and/or memory minimisation). These obvious practical difficulties and limitations remind us that the computer is far from being a miracle machine that always makes our life easier. On the contrary, the formulation of more sophisticated models and simulations is nowadays a highly complex and demanding field and computer applications within geology continue to increase every year. Some global idea of the range of computer-based techniques now used in geology is given in Merriam (1981, pp.24-42).

In probability modelling geologists are confronted not only with deterministic or time-independent models, but more and more frequently are using markov processes and even more uncertain processes, of which the general random walk is perhaps the most well-known case. Krumbein (see Merriam, 1976, pp46-47) reminds us that "some random elements do enter many geologic phenomena" and therefore the computer is

essential for modelling long-run simulations by varying numerical parameters. In paleontology, according to Merriam (1981, pp267-268), "the computer has taken its place with the microscope and hand lens as a basic research tool", leading to a search for general statistical laws, aided by the "computer catalyst" without which it would be impossible to treat the markov processes involved in a "massive" way.

A final example will perhaps be of special relevance to our meeting here in New Zealand, which has its own record of volcanic activity. I refer to an attempt to model the repose period patterns of volcanoes (Merrian, 1976), where the author distinguishes six renewal type markov models to simulate different stages of volcanic activity, using simple transition parameters for the probabilities of passing from one state to another. The models are computer-simulated and the parameters varied to produce an effect similar to that obtained in reality. A computer was essential for the long-run trials involved.

4. Conclusion

Geology provides a host of case studies for the use of statistical, probability, and computer models covering representational and parametric statistics, deterministic and markovian processes, random walk, and Monte Carlo methods. For most of these applications, the computer has become essential for the storage and manipulation of data as well as for the rapidly developing and innovative field of computer modelling using simulations. Not only do the Earth Sciences provide a rich resource of examples for statistics and probability courses, but they also provide an example of the power and mutual benefit of linking statistics/probability/computer modelling with a modern day experimental science.

References

Agterberg, F P (1974) *Geomathematics*. Elsevier.

Davis, J C (1981) Use of the computer in petroleum exploration. In: D F Merriam (ed) *Computer Applications in the Earth Sciences : An Update of the 1970s*. Plenum, 125-143.

Le Maitre, R W (1982) *Numerical Petrology*. Elsevier.

Merriam, D F (1976) *Random Processes in Geology*. Springer.

Merriam, D F (ed) (1978) *Recent Advances in Geomathematics*. Pergamon.

Merriam, D F (ed) (1981) *Computer Applications in the Earth Sciences : An Update of the 1970s*. Plenum.

Rogerson, A (1988) Problem solving, modelling and applications of mathematics. *Proceedings of the Sixth ICME Congress*. Budapest.

Rogerson, A (1989a) Mathematical modelling in the sciences. In: W Blum et al. (eds) *Applications and Modelling in Learning and Teaching Mathematics*. Horwood.

Rogerson, A (19889b) The human and social context for problem solving, modelling and applications. In: M Niss et al. (eds) *Modelling, Applications and Applied Problem Solving*. Horwood.