DATA REDUCTION IN PURE SCIENTIFIC RESEARCH

by

T. Pearcey

Commonwealth Scientific and Industrial Research Organisation,
Sydney, N.S.W.

INTRODUCTION

I would like here to talk around the question of the future computing requirements of scientific bodies who deal with large amounts of information derived from experimental records, and bodies as university research units, government scientific and industrial research organisations and the large industrial computing research laboratories. Many of these bodies are already faced with the problem of assessment of accumulating data. We may imagine that such an organisation will equip itself with a system for data analysis possessing sufficient capacity to deal with the rate of experimental recording.

I would distinguish data handling problems arising in such places from those arising in fields such as commerce, national census and statistics, and market research.

The latter fields have their special problems of data collecting, handling etc. and a reasonable time is allowed for the processes of reduction which are mostly of routine fairly uncomplicated kinds. In the former research fields data is likely to be provided at a fast rate as the result of a continuous experiment, reduction of results may be required within a short interval, for the next state of the study may depend upon their assessment. Further, the data provided for reduction and analysis may vary widely in significance from day to day since few laboratories of the type we are considering limit their activities to one very restricted field of study which would allow analysis to become a matter of pure routine. Thus, although new data may always be provided in the same medium, the processes of analysis will frequently differ from one day to the next and from project to project, and often more than one project will be providing data at the same time.

We expect the processes of data reduction in pure research projects to be more elaborate than those such as exist in commercial and national statistic problems. They are unlikely however to be as involved and complex as the sequence of processes required in many 'theoretical' scientific problems as arise in theoretical physics, engineering, and simulation studies.

Problems of data analysis are likely to arise in these fields which are highly instrumented; we can readily believe for instance that there will arise in these problems continuous micro-biological, chemical, fibre and spectroscopic studies, in x-ray crystallography of complex molecules, continuous recordings of cosmic rays, semi-conductor properties and so on. In particular I shall consider radio-astronomy especially as the typical case occurring in this country.

FEATURES AFFECTING DATA ANALYSIS

We shall be most interested in the way in which the special characteristics of highly instrumented data production affects the type of equipment required for its rapid and automatic reduction particularly in regard to its affect on the specification of the processing devices, on the recording of the raw data and final presentation of the results for assessment.

Some of the features have been mentioned above. The principle factors involved are,

1. The sequence of information and the rate of recording.
2. The type of experiment and the media of recording.
3. The processes of analysis, amount of computing, and variety etc.
4. The amount of editing and checking required of the raw data, and during later stages of the analysis.
5. The type of presentation and the media to be adopted.

I will discuss these features, although not in the above order.

We may view the problem of scientific data analysis as one of dealing with a steady daily supply of information from a variety of projects, that from each current project or experiment representing information in a sequence, arrangement and precision chosen by the observer to suit his needs. Each such set of data requires its specific treatment. The whole data handling system must handle the analysis of the daily inflow of data either completely, or to such a stage that it may later be collated with information from the same project previously produced; all this must be completed within a few hours, if not, a back log of data piles up. Clearly, in view of the variety, arrangement and precision of the data presented, the analysis system must be made very flexible; it must be possible to set it up to deal with any stage within only a few minutes.

Although any one laboratory may be specializing in a particular field, it will have a number of problems being studied with a variety of instruments. In the case of radio astronomy, studies are made at ever increasing resolution of the whole and parts of the sky at various frequencies, using interferometers, pencil beam transit type antennas, and later will include a high resolution giant steerable radio telescope. The receivers provide sequences of voltage outputs as a function of time representing the 'brightness' at different celestial positions, at different frequencies, or both, together with time signals and calibrating data at wider intervals. Commonly this data is traced by pen recorders on a moving chart. Each receiver produces different information which must be treated differently.

THE RECORDING RATE

It seems likely that experimental data will only infrequently be recorded to very high precision, although sometimes more precise values can be recorded as a set of small differences. Variables like time and frequency may often be known to a precision of 1 part in 10^6, but most data will not be better than 1 part in 10^3 or 10^4.

The rate of recording will depend upon the number of channels of information required simultaneously, and these may often amount to ten to thirty. In continuous recording by a set of pen recorders the sampling rates are effectively controlled by the time constants of the pen and rate of paper read, and are adjusted to show rates of fluctuation of the signal up to some chosen limiting frequency suitable to the experiment. In recording for high speed analysis on a single track of information data must be sampled during correspondingly shorter periods and the channels switched into use in some specific sequence chosen by the observer. We can readily see that sampling periods may in some cases be reduced to a few msec and the operation of holding or transferring the datum for record may require to be as small as a few μsec. This certainly suggests adoption of electronic processes for recording, and consequently strongly suggests adoption of digital representation.

Although the recording rates experienced by W.R.E. are at present considered high, there will inevitably be an increase in recording rates in most fields of investigation as techniques improve. In the Radio Physics Division of C.S.I.R.O. the data output from radio astronomical equipment, currently equivalent to about 100 binary digits per second is likely to rise to the order of 10^3 binary digits per second with the introduction of equipment now being considered.

AMOUNT OF DATA TO BE HANDLED

The data handling system must be able to keep up with the faily flow of information. The daily data may, as in some radio-astronomical project, consist of a full 24 hours of a continuous recording at a rate of some hundreds of binary digits per second thus presenting data at a rate of about half a million bits per day. The actual rate of presentation of data may differ widely depending upon the current projects and equipment in operation.

Data may be accumulated over a period, and final analysis and collation may be carried out only on the whole of the data and the end of the observing period. This may amount to a very high volume of data. For instance we may expect that a general survey of the sky covering the 20 cm hydrogen-line frequency (1420 Mcs), over a bandwidth of 1 Mc at intervals of 50 Kc, and with a high resolution steerable antenne giving a beam width of 12 min of arc, would require the recording of the equivalent of 10^8 to 10^9 decimal digits alone! However we may expect the average project to occupy only an order 10^6 to 10^7 decimal digits.

THE PROCESSES OF REDUCTION AND ANALYSIS

The processes of analysis of scientific data are likely to be complex and of considerable variety even within the restricted field of activity of specialized laboratories. Analysis goes beyond just the simple reduction of the raw data and includes such operations such as the fitting of corrected results to certain theories by testing for expected correlations between parts of the recorded items and fitting of data to chosen expressions for the evaluation of parameters, a search for special features, the transformation of data to different co-ordinate systems and sorting for the desired form of presentations soon. Analysis is always a form of condensation of information.

The first stages of analysis we may call 'reduction'. These consist of editing the raw data, supplying corrections for calibration factors which may vary non-linearly with the variable and may differ from one part of the record to another and from day to day, and for the correction of errors such as zero settings and other characteristics of the experimental equipment.

In regard to the editing stage, as is well known, experimental equipment is not always 100% serviceable, and, particularly radio equipment, is liable to temporary failure due to external interference. Further, observers are not infallible even if the data handling side itself never fails. Thus the raw data must be scanned for faulty sections. Blocks of faulty data may be removed and simple errors may be corrected by simple interpolation processes. Reduction thus provides a 'clean' copy of the data suitable for the later process of 'analysis' which may involve elaborate transformations.

Data provided at different times by the same project, may overlap the range of variables covered by a previous set of data. Upon collation of the two groups after reductions comparison checks must be made to ensure that, results are truly reproducible under the same circumstances. In this way 'systematic' errors due to unsuspected faults or changes in equipment may be detected and possibly removed.

The final stage of the analysis must always prepare the data for the media of presentation. In some cases only a short table of figures will be required. It is usually a poor method of analysis which provides a vast quantity of page printed results at high speed for visual assessment. Frequently it is desired to present a considerable amount of data in a manner which can be readily appreciated, and in which the degree of precision is not immediately important, in order that a decision can be made in regard to further analysis of sections to be selected. A graphical presentation suits the need, the contour diagram is particularly suitable in the manner in which it presents much information in a readily recognisable form.

The computation, sorting and re-arrangement of the data during analysis will be very varied but it is frequently a virtue for data reduction problems that a full analysis may be broken into a set of distinctly separate stages, e.g. a numerical transformation, followed by a sort or collation followed again by a further numerical change. This sequence may become extensive, a number of such sequences suitable for radioastronomy amount to 50 or more stages. It is clear that a considerable amount of computation will be required. This must be to a precision somewhat better than that of the raw data in order to avoid the accumulation of error in extended sequences of operations. It is expected that a precision of at least 6 decimal digits would be needed. A digital method of analysis is therefore indicated. The advantage of the digital method is its great flexibility. Processesses may be changed readily and the precision of computation may be programmed to suit the data when necessary.

In the handling of some 10^6 digits any human link in the chain of analysis, other than that of general supervision must be avoided. This calls for a fully integrated system in which the data recording, analysis and presentation equipment, are considered as a single flexible system.

RECORDING THE RAW DATA

The raw data must be recorded on a medium which may be accepted directly by the analysis part of the system, the processing unit, and it is reasonable to ask that the observer shall always record his data on the same medium, although the precision and sequence of what he records may be of his choice.

The raw data will normally be recorded separately from its treatment; the experimental equipment are frequently located away, even in the field, and treatment cannot occur until sufficient data is collected, and the processing unit will be expected to treat data from a number of different experiments during any one day.

The medium must handle the expected recording rate and must itself be readily handled and transported, and recording may be required to continue unattended over long periods. The choice of medium naturally falls to magnetic tape, as having a suitable packing density, cost economy, reliability in the field, ready transportability and rapid reading rate during analysis. It fits conveniently into a scheme of sequential sampling of time varying quantities. Normally a days' recording could be stored on one 1 000 ft of quarter inch width magnetic tape.

The details of how many channels and tracks of information shall be provided on the tape, whether it be quarter inch wide or greater depends upon the detailed requirements of the laboratory and the reduction scheme adopted. Clearly however special effort must be given by each laboratory to the detailed design of its instruments so that they may be tied into a digital tape recording scheme. Each requirement will differ from others; special purpose analogue-digital converters will be required although the incidental digital stores, pulse timing, sequencing, and tape transport gear may well be standardised and provided commercially.

Provision must be made for recording a number of channels such as time or other characterisation symbols, one or more variables in turn, together with occasional values for calibrations. The arrangement of these on the tape could be decided by the observer who may be provided with a plug-board by means of which he specifies his recording requirements. The recording gear will thus be a special purpose unit with plugable sequence for channel sampling.

That data will be digitally recorded automatically should not prevent the simultaneous use of the more conventional pen recorder for the same variables since these recordings are readily scanned by eye for interesting features later to be picked out for detailed analysis, and for detecting faulty sections of data due to periods of interference or failures of experimental equipment and adjustment.

PROCESSING THE DATA

From the foregoing it is clear that a central digital type of analysis is required the questions which arise are those of its capacity and speed.

In view of the variability of the detailed sequence of analysis it is clear that a programmable computer is called for, but since usually the analysis may be broken down into distinct stages, the programmes will not be very extensive, less so than those occurring in scientific computation but more than those in commercial data reduction.

Usually the data currently held in the computer will be a very small amount of the data recorded, so that the internal store of the computer may be relatively small, of the order of a 1 000 words. The process of reduction and analysis will consist of a series of passages of the data from one or more magnetic tapes to another each transfer comprising one stage of the process. At the end of each passage tapes must be adjusted for the following passage and at the same time the programme for the next stage may be fed into the computer.

If a sufficient number of tape transport units are available, the adjustment may often be made during the current tape passage.

An average daily amount of data fed into such a computer continuously could be scanned in about 10 min, transformation and recording of the output occurring at the same time. At this speed, an analysis requiring an average of 50 well organised tape passages could occupy only ten to twelve hours of working time.

In order to maintain this high rate of tape transport and to avoid difficulties in timing and with tape wastage and consequent loss of effective speed, the computer must operate at such a high speed that it will never, or only very rarely, require the tape to stop before calling for the next datum. If we accept a nominal figure of 200 commands to be performed before the next datum is called without stopping the tape unit, and a normal packing density of characters on the tape to be 80/in. then the average time allowed for an operation is about 10 μsec for single channel tape travelling at about 100 in./sec. The allowed operation times decrease as the number of parallel channels on the tape increases for the same rate of tape transport, whilst the total length of data tape decreases in the same manner. An expensive alternative to providing rapid computation would be to provide a sufficiently large buffer store between the tape input and output units and the computer. With a sufficiently fast computer this buffer store need accumulate only the datum currently under the heads or hold for immediate reading.

Clearly, computers of this order of speed, assuming fairly elementary command codes, will be of the parallel type. In view of the normally restricted precision required the word length may be only 20 bits with one address code and one command per word. Engineering techniques in parallel operating magnetic core internal stores are now only approaching the speeds called for. The present situation seems to be that operating speeds are limited by storage access time, not by the time required to perform arithmetic. With vacuum tubes driving the store the desired speeds could not be realised, with transistor drives there would be some difficulty. There is however hope that store of one μsec access time will become available in the future when operation times of three to five μsec will become realisable, and the advantages of wide use of high frequency transistors will be available.

There is no reason to suppose that the cost of the analysis unit will increase as its speed increases with the introduction of new techniques. If these units are to be commercially available their cost is likely to depend upon the amount of equipment involved e.g. the size of the store, and a simple command code usually goes with simplicity of functional design and equipment and to some extent also simplicity in programming.

Ancilliary equipment to the analysis unit will be magnetic tape transport units. At least three of these will be required for the purpose of merging and sorting and an additional unit would be convenient for holding programmes for rapid insertion into the machine.

FINAL PRESENTATION

We have already mentioned the importance of graphical presentation compared with the printed page.

Although on occasion a line printer is indispensible large lists of figures are difficult to access and the actual printing, if done from a magnetic tape via the computer, is very inefficient in the use of machine time. Large volumes of printed output should be provided by magnetic tape-printer auxiliary equipment which would occasion some considerable additional expense.

Presentation of final results in quantities, larger than can conveniently be page printed, must be carried out at a rate comparable with the speed of passage of magnetic tape, say over a period of about 15 min. It is suggested that results from magnetic tape be transcribed into graphical form via a facsimile type of recorder so successfully put into use at W.R.E. In view of the cost however this should be done via the computer and not by means of additional converter units.

For keeping track of the progress and correctness of an analysis a small quantity of selected information would be printed directly from the machine so that this information will be immediately available to the analysis staff.

ANALYSIS FOR RADIO ASTRONOMY

To illustrate the magnitude of a reduction problem facing some scientific laboratories I will mention one which we anticipate will arise in the division of radio physics of C.S.I.R.O. in the field of galactic radio astronomy.

For any one project data will be recorded sequentially and digitally, the sequence of data depending upon the project, and will consist of values of integrated field strength to 0.1% precision for various beam positions and frequencies, antenna direction co-ordinates to one minute of arc, and sidereal or solar time to 1 sec together with calibration factors for all signal channels. From this is required a list of radio sources, their magnitude or brightness temperatures, and contour diagrams of brightness of the whole or selected portions of the sky at scales specified by the observer.

We may assume that data will be recorded simultaneously on magnetic tape and on paper charts by pen recorder. A glance at the latter will indicate the main features of the record and may affect the process of analysis. In any case periods of poor data may be detected and noted at the head of the corresponding data tape prior to its first passage of reduction.

The first passage of the data tape will correct isolated freak errors and remove those portions considered faulty and allow for calibration changes and other known errors of the experiment.

The next stage involves removing those fluctuations in the signal which are known to be beyond the limits of the frequency bandwidth of the antenna system, and isolating those due to time scintillation. The signals will be converted to brightness and interpolated to a regular mesh of points in celestial co-ordinates. Redundant and duplicated data will be checked for consistency and then incorporated into a final set of values for the co-ordinate mesh. At this point a search for radio sources in two dimensions i.e. brightness maxima, will be made. Interpolation for the celestial locations and brightness, and possibly angular width, will provide a list which may require sorting in order of magnitude before printing.

Further, the brightness data, ordered in an array of celestial mesh, may require to be transformed to a similar array in galactic co-ordinates. This requires a considerable amount of interpolation and sorting, and will be followed by further operations ordering the brightness values for final production of the contour diagrams via the output facsimile recorders.

CONCLUSION

Finally, I would summarise as follows. Scientific data processing in the future will involve extended sequences of operations of analysis of considerable variety. At the rates of production expected, the whole process for recording the raw data to the presentation of results for assessment must be automatic in so far as human effort is directed only to general supervision and minor handling of the data involved.

Special effort must be applied by any laboratory to the correct design of equipment to record its data, especially to any analogue-digital conversion. The most suitable medium of automatic recording in the field is magnetic tape.

Processing will be carried out using a centrally situated digital processing unit comprising a medium storage capacity computer. At each stage of the analysis this unit will be suitably programmed and the process of reduction will consist of the sequence of tape passages reading and recording simultaneously. The operation rate of the unit will be so great that the processing will be limited mainly by the rate at which magnetic tape is passed. Operation times of 3 to 5 μ sec are desired.

Output of results will be largely graphical. Fast graph plotters, taking periods of the order of time of a tape passage are required. It is suggested that the facsimile type of recorder be used here.

DISCUSSION

Dr. I. Bassett, University of Melbourne.

For two dimensional smoothing is there any means by which many movements of the tape transport can be avoided to give the information for any point?

Dr. T. Pearcey (In Reply)

As far as I can see there would have to be extra tapes made of the same record so that the surrounding information could be obtained by simultaneous reading. This would avoid having to use the backing store for storing large amounts of data and would overcome the serial storage on the magnetic tapes.

Dr. M.V. Wilkes, University of Cambridge.

I think perhaps that the editing of the information provided from the records you mentioned may be a little bit more difficult than you suggest. Could you elaborate on this please?

Dr. T. Pearcey (In Reply)

This is purely a matter of degree and depends largely on the experiment and the data being recorded. Certainly in the case I have mentioned it would be possible to remove errors caused by say, interference or noise, by mechanical means provided they did not occur more than two or three times in a second. In general for large sections of data it is necessary to use human interpreters whilst for small sections it can certainly be performed mechanically.

Dr. G. Hill, University of Melbourne.

Would it be possible to use the radio data from the previous day to check that from the following day and remove errors this way?

Dr. T. Pearcey (In Reply)

Yes, this is a possible check but one would have to include the daily variation one expects together with effects which may have occurred on the previous day. This may prove to be difficult in some cases but it certainly is a method.