CDD: 574.50285
574.509811

A PROPOSAL FOR DATA BASE DESIGN IN AMAZONIAN ENVIRONMENTAL RESEARCH

Carl M. O'Brien[1]

ABSTRACT - *Environmental research within the Amazon requires data from both macro - and micro- observational studies. Collectively botany, ecology, the earth sciences and zoology amass vast amounts of data. For these data to be utilised fully by computer-based methods requires careful management and planning. The choice of data base and design will pre-determine the usefulness of many observational studies. A characterisation of observational study into one of five types: either institutional, local, regional, national or international is proposed. Potentially, the design concepts for information storage and retrieval are different for each type of study but their commonality is discussed.*

KEY WORDS: Observational study, Computer-based methods, Information, Design, Network.

RESUMO - A pesquisa ambiental na Amazônia necessita dados resultantes de estudos observacionais a niveis macro e micro. Disciplinas como a botânica, ecologia, ciências da terra e zoologia reúnem muitos desses dados. Não obstante, sua utilização ótima através de métodos computacionais requer cuidadosa manipulação e planejamento. A seleção do tipo de banco de dados virá predeterminar a utilidade dos dados observados. Os dados obtidos de estudos observacionais podem ser divididos em cinco categorias: institucional, local, regional, nacional ou internacional. Potencialmente, embora o tipo de "design" para armazenagem e "retrieval" da informação varie de acordo com cada tipo de estudo seus pontos em comuns são discutidos no presente.

PALAVRAS-CHAVE: Estudo observacionais, Métodos computacionais, Informática, Planejamento, Rede.

[1] Museu Paraense Emílio Goeldi. Departamento de Ecologia. Av. Magalhães Barata, 376. Caixa Postal 399. CEP 66040.170. Belém-Pará.

INTRODUCTION

Environmental research within the Amazon is important both *per se* and for an understanding of man's impact on that environment. Botany, ecology, the earth sciences and zoology are all important aspects of this research and in conjunction with the disciplines of mathematics, physics and statistics may one day lead to a greater understanding of the Amazon as well as of the earth's ecosystem. Any potential inter-dependencies that exist between the plant and animal kingdoms and their environment need to be investigated, documented and understood.

To this end, observational studies at both the macro- and micro- level are routinely undertaken in botany and ecology, for example. These yield vast quantities of data which are important both to their respective discipline and as an aid to environmental impact assessment (Tietenberg 1988).

The physical collection of large amounts of data brings with it new problems! Management of data, planning and information retrieval are by no means trivial and each requires careful thought and consideration. The method chosen to store and manage data will pre-determine the usefulness of an observational study. Traditional handwritten methods of information storage based on forms, records and cards have been replaced by modern computer-based techniques such as spread-sheets and data bases. These have often produced more problems than they were first envisaged to solve.

Both mainframe and micro-based computers are routinely employed for information storage and retrieval but can lead to inter-communication process errors and data transfer problems, for example. The need to consider the potential sources of such problems and errors before they occur in a study is imperative.

This paper identifies possible causes of errors and problems when relying upon computer-based technologies; together with proposing a broad characterisation of observational study based on the type of data to be collected and the uses identified. A number of different design concepts for information storage and retrieval are discussed and recommendations made for the present and future.

DATA ASPECTS OF ENVIRONMENTAL RESEARCH

At first sight, the observational studies undertaken in the disciplines of botany, ecology, the earth sciences and zoology would appear to be different. A transition from handwritten methods of information storage to methods based on computer technologies would seem to present enormous problems and difficulties. In actuality, however, this is not the case!

For the successful utilisation of computer techniques in the field of information storage and retrieval one must merely identify the commonality that exists between the data collected by all environmental researchers.

All data may be characterised as belonging to one of two types : either the *character type* or a *numeric type*. The character data type may consist merely of the name of a researcher undertaking a particular study; while a numeric data type may consist of the positional co-ordinates where a particular plant species was observed in an Amazonian region of study. Character data may easily be stored digitally within both mainframe and micro-based computers; as may numeric data. However, the storage of the latter may cause conceptual difficulties (see, for example, O'Brien(1986)) particularly for the environmental researcher. Expert guidance from a Statistician may be needed for the correct classification and definition of data type.

A numeric data type may be further sub-divided into one of two categories: either a *quantitative category* or a *qualitative category*. The quantitative category is by far the most common (Sokal & Rohlf 1981) and widely encountered in Amazonian research. It may consist of measurements, for example, either on the heights and diameters of trees (Hafley & Schreuder 1977) or on the number of trees flowering and fruiting per month taken over a number of years (O'Brien & Pires 1992). The qualitative category is little understood in environmental research but frequently encountered by researchers in areas of medicine (Colton 1974). It may be likened to the character data type in that although numbers are used to represent the findings of a particular observational study no significance should be attributed to the numbers other than to the fact that each distinct number identifies a different observational finding. For example, in a study of amphibious lake populations one may be interested in recording the presence and absence of a number of species within a range of habitats. While the character data type may be defined and used with the strings

PRESENT and *ABSENT* to denote the presence and absence of each species within each habitat, respectively, an alternative is to use a numeric data type in the qualitative category. This makes the data more amenable to mathematical and statistical modelling (see, for example, O'Brien(1989)) and relies on recoding the character string *PRESENT* with the number one and recoding the character string *ABSENT* with the number zero. The zero-one qualitative data may then be modelled, for example, using techniques appropriate to the analysis of binary data (see, for example, Cox & Snell(1989)).

The characterisation of all data into either the character type or a numeric type helps to identify a commonality that exists between *all* observational studies. However, when considering data collectively one needs to distinguish further those variables that will *always* be measured and recorded in an observational study from those that will only *sometimes* be available. The latter category may include comments and remarks made by the environmental researcher during the course of a study. This character data type will not always be available but when it is must be formally retained in a way that will permit future retrieval in an easy manner.

The ways in which such data may be stored digitally cause design problems with respect to the most appropriate computer language to use for representation of the data and the best type of computer architecture to adopt. These are discussed further in the next Section. However, before leaving this discussion of the data aspects of environmental research it is important to identify one last concept. The concept of *static* and *dynamic* arrangements of data.

Observational studies which produce vast amounts of data may be either *unique* in the sense that they are "*one-off, never to be repeated*" studies or may be part of a *continuous* environmental research programme. The data collected in the former case may be termed *static* since once collected a set of data will remain unchanged apart for minor corrections arising from transcription errors. Data collected in the latter case may be termed *dynamic* since once collected the set of data will be periodically augmented by the addition of new observational data taken at later points in time. The dynamic collections of data are by far the most frequently encountered in environmental research observational studies both within, and outside, the Amazon.

The distinction between the static and dynamic collections of data will be returned to later when the concept will become important to the utilisation of computer techniques in Amazonian research.

The concepts presented in this Section are important to all studies involving data collection. It is important to recognise their existence! Once identified, only then may thought be given to the choice of computer-based technology most appropriate for information storage and retrieval.

COMPUTER-BASED TECHNOLOGIES

For environmental research observational studies and environmental impact assessment, the need to transfer from handwritten methods of information storage to computer-based methods is not in doubt (see, for example, Green(1979)). What is in doubt, however, are the choices to be made. While there is a plethora of computer-based techniques, some of which were mentioned earlier in the *Introduction*, one must never loose sight of the fact that the computer merely stores, manipulates and displays information. The choice of hardware and software configuration *must* be dependent upon the particular characteristics of *each* observational study. The current fallacy that one computer and one piece of software will be appropriate for all environmental research must be dispelled.

Computer programming languages are many and varied; and like computer hardware they are best suited to different applications. The traditional 3GLs (Third Generation Languages) such as the languages of BASIC, COBOL and FORTRAN; together with the newer languages of ADA and C, are most people's introduction to the world of computing. These provide an interface to the computer but their use requires a high-degree of sophistication and expertise on the part of the environmental researcher. The 4GLs (Fourth Generation Languages), however, like those incorporated into statistical software packages such as BMDP®, SAS®, and SPSS® and most data base languages such as SQL, represent a higher-level, easier to use interface to the computer. To use these successfully the researcher is required to think less about computation and to think more about problem-solving. Operating systems such as Microsoft®, MS-DOS®, UNIX® and VAX/VMS are now common place and have much the same functionality; if slightly different command syntaxes. WIMPs (Window, Icon, Mouse and Pointer) are becoming ever more popular with both the micro-based computer user and the traditional mainframe user. Today, for example, a micro-based computer user can routinely expect to find the WIMPS implementation provided by Microsoft®, Windows™ to be available on a computer operating under MS-DOS®. Data base systems such as micro-ISIS, dBase® and spreadsheets are widely available, easy to use but tend to be restricted to applications on micro-based computers.

Commercial software such as DBMS Copy has been produced which can take data in the format of one piece of computer software and convert them into the format of another piece of computer software. In this way, for example, data bases in the format of Lotus 1-2-3® may be transformed into data bases in the format of Paradox® and vice versa. Communication software such as kermit and Rainbow allows the physical transfer of both data and files from a micro-based computer to a mainframe computer (so-called*uploading*) and transfer in the opposite direction (so-called *downloading*).

Mainframe and micro-based computers should no longer be thought of as isolated machines devoid of a commonality but instead, may be thought of as exchangeable and interchangeable pieces of hardware. The choice of whether to use a mainframe or a micro-based computer for information storage and retrieval should no longer be based on prejudice and ignorance. Both have their place in environmental research and both should be used!

The particular choices of hardware and software are becoming increasingly less important. Management of data, planning and information retrieval are all far more important but require informed judgements to be made.

INFORMATION STORAGE AND RETRIEVAL

Apart from a choice of computer-based technologies, it is important to identify the potential uses to which the data collected as part of an environmental research programme may be applied. A characterisation of observational study based on the type of data to be collected and the uses identified for the study is proposed.

Five kinds of observational study can be identified : the*institutional*, the *local*, the *regional*, the *national* and the *international*. Each requires its own unique arrangement and collection of data but the transfer of data between one type of observational study and another may be necessary. At times, it will also be highly desirable.

Studies at the institutional level are to be considered as special cases of those at the local level, local level studies as special cases of those at the regional level, regional level studies as special cases of those at the national level, and lastly, national level studies as special cases of those at the international level. The movement from institutional level through to international level must*never* necessitate either the loss or neglect of existing collèctions of data. It must be possible to progress in a hierarchical, linear way from one level to the next. By way of an example, consider the following scenario.

An observational study is to be undertaken within the Amazon. At the *institutional* level, Museu Paraense Emílio Goeldi (MPEG) may, for example, conduct a small scale pilot study. This may later be developed into a local study concerned with environmental impact assessment involving joint collaboration with a *local* institution such as Universidade Federal do Pará (UFPA). At the *regional* level, that study may then be extended to encompass other institutions within the state of Pará. It may then be decided that the study needs to be extended to include institutions from a number of other states in Brazil and develops into a study at the *national* level. Finally, other countries become interested in the study because they have similar problems and hope for global solutions to be found and agreed. Collaboration develops between nations and a study at the *international* level develops.

Each of the five types of observational study could, in principle, require different computer-based technologies for information storage and retrieval. However, similar questions will need to be answered for each:

- *Should a mainframe or a micro-based computer be used?*

In the case of an observational study at the institutional level the choice is really one of personal preference. The same is true of a study at the local level. At the regional, national and international levels the choice must realistically be for a mainframe. Potentially a large amount of information will need to be stored. Remote access by computers between institutions will be necessary and fast response times will be needed.

- ***What computer software should be used ?***

Once again, at the institutional and local levels, the choice is really one of personal preference. Spreadsheets, data bases, 3GLs and 4GLs are all appropriate. Software exists to allow translation between different formats of data but tends to be expensive and to have a limited lifespan. At the regional, national and international levels the choice must be between 3GLs and 4GLs. Historically, FORTRAN has established its place in numeric and scientific computing. However, the problems of using the 3GLs to store large amounts of data are well-known. Data base management software has been developed to alleviate these problems and to guarantee the integrity of the data stored in an observational study. Most have a host language interface for calling application programs written in 3GLs directly. The Structured Query Language, SQL mentioned earlier, is the implementation of the data base model that is the accepted*standard* in research and development. It is a powerful but easy to use language, supported on both mainframe and micro-based computers. Its use is to be encouraged!

- ***Should data be stored on the diskettes of micro-based computers or on the magnetic discs and tapes of mainframe computers?***

The choice depends on whether the observational study will be static or dynamic, in the sense discussed earlier. If static, then the choice is unimportant and should be based on personal preference and familiarity. If dynamic, however, the choice must realistically be for the use of the storage media of mainframe computers; if necessary, downloading files where appropriate.

The need to use both mainframe and micro-based computers*together* for information storage and retrieval will be essential to the future success of environmental research. In a perfect world this might be all that needs to be considered. However, problems and errors will still occur.

PROBLEMS AND ERRORS

Prototyping of a computer-based solution is essential prior to the collection of the first piece of observational data in an environmental research programme.

Hypothetical data must be generated for the study and investigated in order to identify possible problems with a chosen computer-based technology. Simulation studies, as one might term these, are common in mathematics and statistics research but under-used in environmental research at the present time.

Network communication hardware and software are becoming ever more sophisticated. It is now routine both to send electronic messages from Brazil to England by BITNET, for example, and to connect to remote computers hundreds of miles away. The Data General ECLIPSE MV/9500 at MPEG in Belém, for example, may be remotely connected to mainframe computers in Brasilia and Rio. Statistical software packages maintained on these remote machines may be accessed and discriminant function analysis (Srivastava & Carter 1983) such as presented in Claytor et al.(1992) routinely undertaken. All network connections and routes must be checked and tested prior to the investment of substantial amounts of time and effort in the development of computer-based techniques for information storage and retrieval. There is little point in developing a sophisticated collection of environmental data in a remote part of the world if the scientific community at large can not have access to the information.

Problems regarding the updating of dynamic collections of data must be considered. Schedules should be provided both for minor corrections arising from transcription errors and for the routine addition of new information obtained in an observational study. These can cause immense problems if *not* seriously thought about before an environmental research programme begins!

CONCLUDING REMARKS

Observational studies in environmental research can yield vast quantities of data. The physical collection of large amounts of data can bring with it problems which require care when computer-based solutions are proposed for information storage and retrieval.

In this paper, concepts important to the successful utilisation of computer techniques have been discussed. A characterisation of data from observational studies has been proposed based on two broad types. The need to further distinguish between qualitative and quantitative numeric data has been discussed. The concept of dynamic and static arrangements of data has been argued.

3GLs and 4GLs have been discussed and a characterisation of observational study into one of five types proposed. The design concepts for each has been discussed with respect to information storage and retrieval using computer-based technologies. All present and future environmental research within the Amazon and environmental impact assessment either requires, or will require, the collection of data from observational studies.

Simple questions should be answered *before* a computer-based technique for information storage and retrieval is suggested. The need to investigate all possible causes of problems and errors prior to the adoption of a computer-based technique is crucial to the successful utilisation of any data collected in an environmental research programme.

ACKNOWLEDGEMENTS

I would like to thank José Guilherme Soares Maia, the present Director of MPEG for making my visit to Brazil possible. The research presented in this paper was undertaken with the financial assistance of Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq). Their granting of Pesquisador Visitante Estrangeiro (grant 300889/91-7) is gratefully acknowledged.

REFERENCES

CLAYTOR, R.R, MacCRIMMON, H.R. &GOTS, B.L. 1992. Continental and ecological variance components of European and North American Atlantic salmon (Salmo salar) phenotypes. *Biol. J. Linn. Soc.* 44 : 203-229.

COLTON, T. 1974. Statistics in Medicine. Boston, *Little Brown and Company.*

COX, D.R. & SNELL, E.J. 1989. The Analysis of Binary Data . 2. ed. London, *Chapman and Hall.*

GREEN, R.H. 1979. Sampling Design and Statistical Methods for Environmental Biologists. New York, *Wiley-Interscience.*

HAFLEY, W.L. & SCHREUDER, H.T. 1977. Statistical distributions for fitting height and diameter in even-aged stands. *Can. J. For. Res.* 7 : 481-487.

O'BRIEN, C.M. 1986. One view of the GLIM/IKBS initiative. *NAG Newsl.* 86(1): 11-20.

O'BRIEN, C.M. 1989. Working with GLIMPSE.. Oxford, *Numerical Algorithms Group Limited.*

O'BRIEN, C.M. & PIRES, M.J.P. 1992. A case-study using GLIM 3.77 : Modelling the flowering and fruiting patterns of tropical rainforest trees. Prof. Statist. 11(1): 2-5.

SOKAL, R.R. & ROHLF, F.J. 1981. Biometry . 2. ed. San Francisco, *Freeman and Company.*

SRIVASTAVA, M.S. & CARTER, E.M.1983. An Introduction to Applied Multivariate Statistics. New York, *North-Holland Publishing Company.*

TIETENBERG, T. 1988. Environmental and Natural Resource Economics. 2. ed. Illinois, *Scott, Foresman and Company.*

Recebido em: 20.03.92
Aprovado em: 11.09.95