Quaternary Proceedings No. 5, 1997 1- —

Towards an Expert System in Palaeoentomology

Philip I. Buckland, Yuan Zhou Don & Paul C. Buckland

Philip I. Buckland, Yuan Zhou Don & Paul C. Buckland, 1996. Towards an expert system in Palaeoentomology. In — — — — — — — . Quaternary Proceedings No. 5, John Wiley & Sons Ltd., Chichester, pp. 1- —.

Abstract

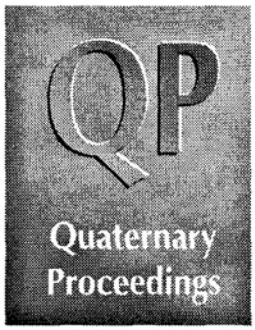
Studies of Quaternary insect fossils, principally of Coleoptera, the beetles, are now sufficiently frequent to warrant the construction of a database to maintain easy access to the record. BUGS, however, seeks to go beyond this and provide ecological and distributional data on the modern fauna to enable more precise reconstructions of past environments. This paper summarizes the program and its salient features and discusses the application of intra- and intersite statistics, which the database allows.

KEYWORDS: Coleoptera, Ecology, Quaternary, Databases, Statistics.

P.I. Buckland, Dept. of Archaeology, University of Umeå, Sweden, S-90653.

Yuan Zhuo Don, Dept. of Computer Science, University of Sheffield, S10 2TN, UK.

P.C. Buckland, Dept. of Archaeology & Prehistory, University of Sheffield, S10 2TN, UK.



WILEY
Publishers Since 1807

Introduction

The Insects form the most diverse group of animals on the planet, with species occupying habitats from the edge of the open ocean to the permanent snow patches of the highest mountains. One Order, the Coleoptera (beetles), not only occurs across such a range of habitats, but also largely consists of individuals which are sufficiently durable to provide some of the most frequent identifiable fossils in Quaternary sediments (Buckland & Coope 1991). Although often exceeded in numbers of preserved individuals by the Diptera (Skidmore 1996), particularly in lacustrine sediments (Sadler, in press), knowledge of beetle taxonomy, habitat and distribution is such that they provide an ideal group for the reconstruction of past environments, both on the regional and more intimate archaeological scale. In addition, their sensitivity to climatic and environmental perturbations make them a useful barometer of landscape change, either induced by human interference or natural change. With such a large and diverse group — the British fauna alone contains nearly 4,000 species — ecological and distributional data are widespread. Those involved in entomology know only too well how difficult the processes of procurring habitat information can be. Sources range from major international works (*e.g.* Koch 1989; 1992) to obscure regional, state, county or province journals stocked only by specific local or national libraries. In the past, collation of this information lay with individual card indices, leading to the frequent reinvention of the wheel.

The original version of the BUGS computer program (Sadler *et al.* 1992), a DOS-based system written in DBaseIII, has been available for some 5 years, and this began the creation of an electronic data retrieval system for insect ecology and distribution, speeding up the process of analysing insect data for both conservation and palaeoecological purposes. With data entry for much of the Holarctic beetle fauna, including the Palaearctic Lateglacial and Holocene record, BUGS has reached a stage where a jump can realistically be contemplated from what is simply the electronic storage of relational information, to an integrated system for the complex analysis of faunal assemblages, that is to say, a step towards an expert system for entomology and palaeoentomology.

BUGS — An Ecology Based Relational Database System for Coleoptera

The BUGS project, the result of international collaboration through NABO (North Atlantic Biocultural Organisation) and funded largely through NSF (Office of Polar Programs), is an ongoing research effort towards enabling easier access to ecological and palaeoecological data and methods. A parallel project, SLUGS, is currently entering data on terrestrial and freshwater mollusca (Keen, unpubl.). The current version of the program runs in the Microsoft Access v2 © environment, under Microsoft Windows© (and Windows 95©), and was written by Yuan Zhuo Don and Philip I Buckland, with the assistance of Paul Buckland and Jon Sadler. It can run with or without MS Access in its basic form, but requires MS Excel for full operation as a palaeoecology and ecology package.

Basic Data Access — Looking up a Species

On entering BUGS, users are confronted with a startup screen from which various accreditations can be viewed, including information about NABO. Following this the main data retrieval screen is displayed (Fig. 1).

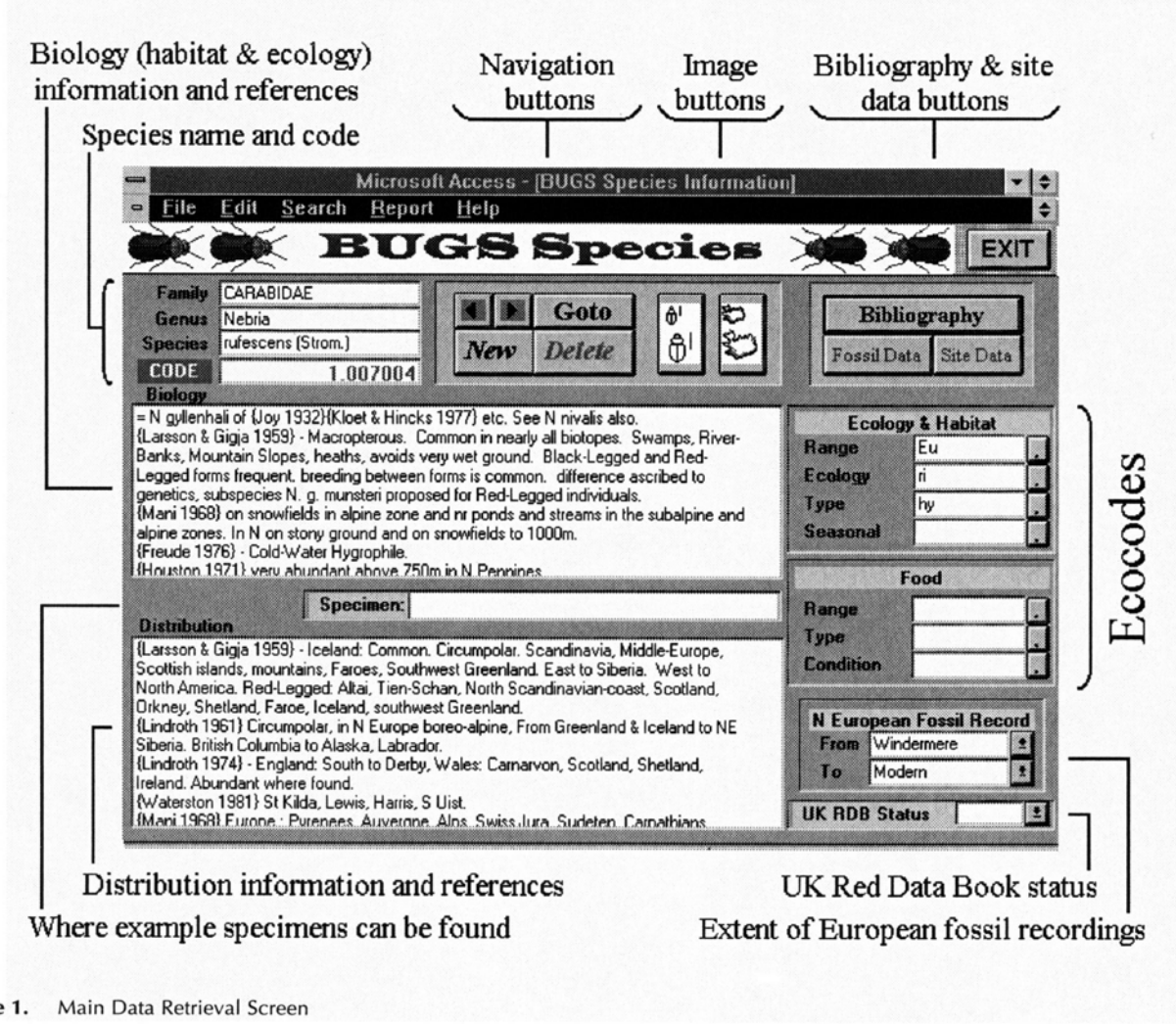


Figure 1. Main Data Retrieval Screen

From here one can access the full range of information concerning the ~5000 coleopteran species stored within the database, including the complete British, Faroese, Icelandic and Greenlandic faunas. The screen is made up of numerous boxes and buttons. In the top left are boxes which provide the Family, Genus and Species, and also its checklist code, although the latter is only displayed when the password is entered. The key field in the program is a coded checklist. This uses two digits for the Family, three for Genus and three for Species, and is based upon the Central European list by Lucht (1987). This is used as the unique identifier by BUGS to access all species, and sort lists into taxonomic order; a spare trailing digit allows for additions of species not present in Central Europe and for the use of higher taxonomic divisions, such as Subfamily, species (sp.) and species plural (spp.). To the right of these are a group of navigation buttons, allowing the user to move through the list sequentially, or search for a particular genus. The [GOTO] button introduces the standard ACCESS search box, initially constrained so as to search the Genus field. New species can be added to the database through this form, and any alterations made, although these are password protected features. A large text box below these buttons is the most commonly examined part of BUGS. It contains the information relating to the habitat requirements, or biology of each species, as extracted from various texts, the selection of which inevitably reflect the interests of the compilers. By clicking on the [Bibliography] button, one can view the full details of each reference, and in fact browse the entire bibliography, currently in excess of 1600 papers (Fig. 2). Similarly, the references used to compile the distribution data — the lower large text box — can be examined. Information on distribution is currently limited to the North Atlantic and European regions, but will be updated with time, initially to include the Arctic to Subarctic Holarctic beetle fauna.

The authors of BUGS are primarily concerned with the interpretation of fossil insect data, although such would be of little value without modern ecological information. They have been involved in numerous collection exercises throughout the North Atlantic region and have tried to ensure that BUGS is suitably entomologist-friendly, and is not merely a palaeoentomological research tool. Modern distribution and habitat data makes up the major part of the dataset, and facilities for viewing both modern and palaeo-distribution maps are present. Currently, the lack of graphics handling routines in Access v2 limits use to purely static scanned maps, although the links to an effective GIS are being contemplated. A text list of sites where the species has been found in a fossil context can be seen by clicking the [Fossil Sites] button (Fig.1, top right), and the relevant bibliographic data can be accessed via the included [Bibliography] button. A summary of fossil information is also given in the lower right part of the main screen. Below these is another display which shows modern ecological data in the form of the insect's Red Data Book status. This is a measure of the relative rarity of an insect. Initially these data have been derived from British

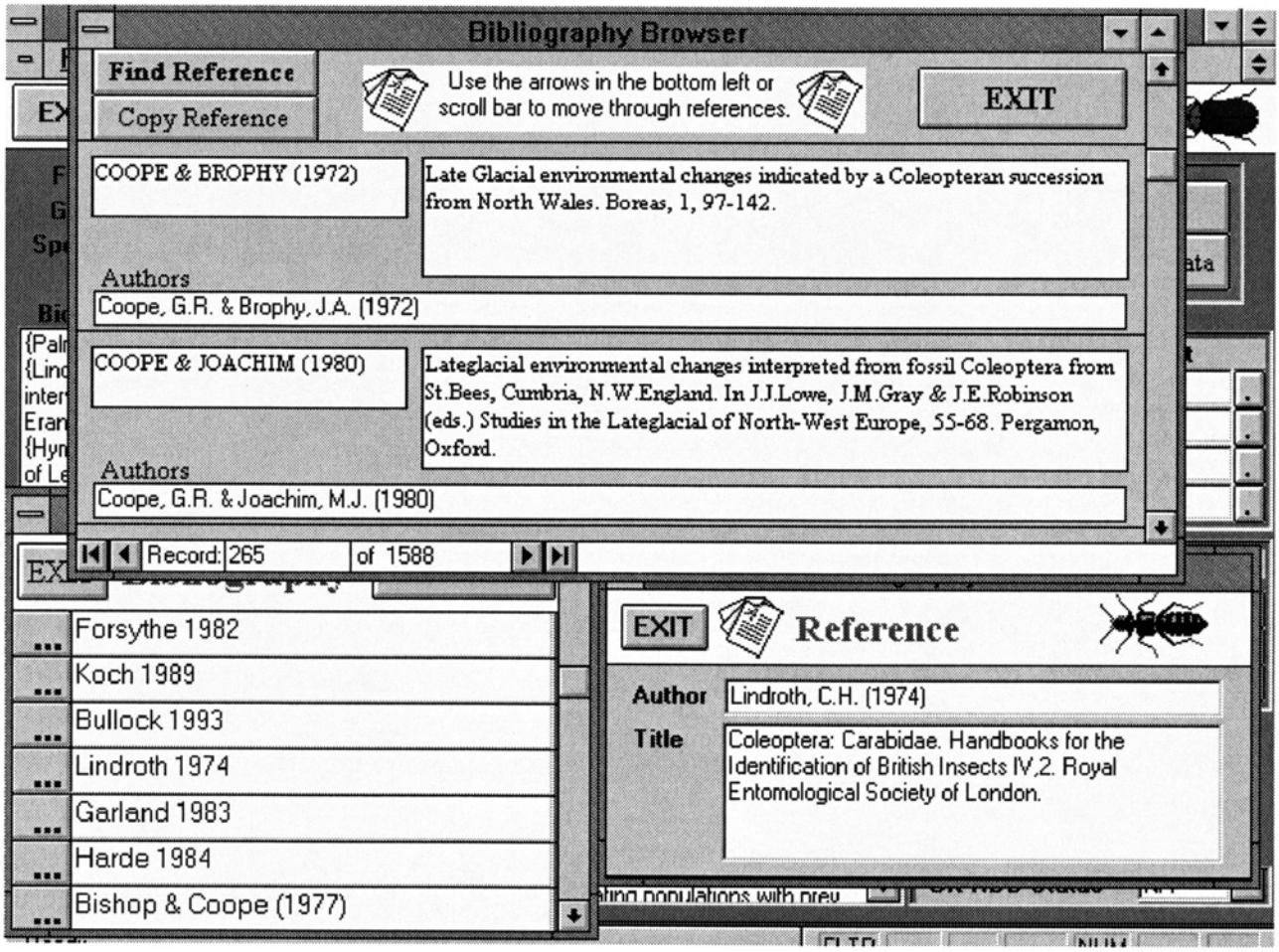


Figure 2. Browsing the BUGS Bibliography

sources (Hyman 1992; 1994), although the facility will be available to enter information from other countries by way of a drop-down menu.

Hyperlinks and Sample Data

Most of the above information about a species could be found, albeit not so easily, using the original DOS version of BUGS, but the realm of Dynamic Data Exchange, Object Linking and Embedding, and Hyperlinks that is Microsoft Windows© allows for more sophisticated data management. Apart from being an internationally standardised user interfacing system, Windows enables a number of complex programming problems to be solved relatively simply. The programmers are therefore afforded more time for problem solving, rather than (at least in theory) debugging. BUGS can store three images of each species — a black and white outline drawing; colour photograph (or digital microscope image); and any further diagrams for identification purposes. The latter two images also have pop-up memo fields for any additional notes.

Hyperlinking is a method by which different components of the Windows environment can be connected. This is particularly useful where it is necessary to link text in a non-linear manner (as is common on Internet Web Pages), or use the functions of an application which is external to the one currently being run. In BUGS the user benefits from being able to use Microsoft Excel © to enter spreadsheet data from within BUGS. Both modern and fossil lists of species consist of lists of MNI (minimum numbers of individuals) for the samples from a site, be it an abandoned Norse Farm or a Greenlandic pitfall trapping project. BUGS stores these data as "count sheets", that is the species list and assemblage data from a site, in MS Excel © spreadsheet files. Lists are automatically sorted into taxonomic order within Access ©, and sample codes and insect frequencies are typed in through Excel ©, which is called up by double clicking on the count sheet in BUGS (Fig. 3).

Site Data and Descriptions

The site data screen is found by either clicking on the [Site Data] button on the main screen (Fig. 1), or the [Site Data] button which appears when [Fossil Sites] has been clicked (the latter calling up the list of known fossil locations for the insect). Sites are sorted alphabetically by country, and then site name, the first site being a general help card which explains the screen layout. The site data screen has two sections (Fig.3). In the upper part of the screen is a collection of boxes showing various details about the site — its name, location, including latitude and longitude, dating evidence, and some interpretational text along with its associated references. Below this is a snapshot (*ie.* a picture representing the data) of the associated count sheet, which itself is an MS Excel © file. Double clicking on this

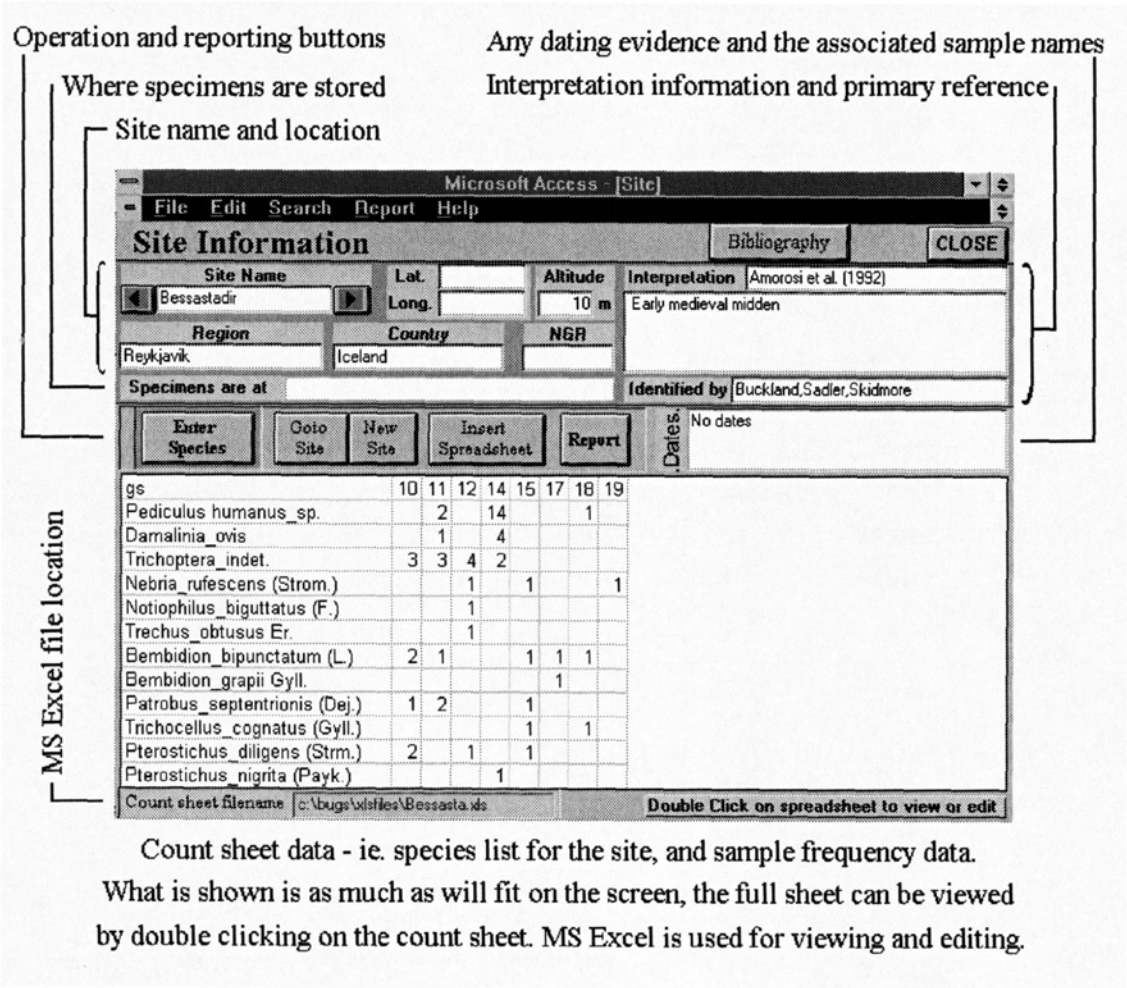


Figure 3. Information about sites and their faunas

starts Excel ©, and allows the user to view and edit data as a spreadsheet (the first column, the species list, is locked to prevent changes in species names causing errors within BUGS). In theory, the spread sheets could hold any information, since they are not limited by the database form. Excel © uses a system of workbooks, in this case the first sheet of the book is the BUGS count sheet. Subsequent sheets can be used for summary data and statistical functions.

Species and their Habitats

Before moving to more empirical methods, it is necessary to describe one of the more simple analysis tools of BUGS — the query form titled "Find bugs with specific habits..." (Fig 4). Keywords are simply typed into the relevant box, and logical operators (such as AND, OR, and NOT) are selected from drop down lists between the keywords. The present scope allows the user to search for up to three words each in both the biology and distribution fields, and to combine this with a search for particular values of the Red Data Book index. For example, one could look up all the species which have a relationship with (MOSS OR SALIX) AND TUNDRA (in Biology), AND are found in GREENLAND OR ICELAND OR CANADA (Fig 4). The query forms one of the most basic question-and-answer routines that databases provide, and as such is an invaluable research tool. The resulting list of species can either be viewed through the standard BUGS interface, output as a list of species names, or output with full details to file or printer.

An important element in the program, particularly when it is being used either in teaching or rapid site assessment, is provided by the [Report] button on the Site Information screen. The habitat, distribution and ecology information for any species list, obtained by identifications from sweep-netted, pit fall or fossil material, can be obtained by pressing this button. The relevant information from the database is joined with the list, which can then be output as a simple report, in .RTF format, to be picked up with WORD6, or other word processing packages.

Towards an Expert System

A step towards the creation of an expert system has been the incorporation of coded ecology descriptors — Ecocodes, as we like to call them. These two letter codes, mostly modified from Koch (1989; 1992), at present cover over 120 terms, which are divided into seven sets (Table 1). Those which describe habitat — range, ecology, type, and seasonality — are in the first group. Food range, type, and condition make up the second group. Codes can be combined, so, for example, the large Carabid beetle *Carabus granulatus* L. has the habitat range description "Eusihy" — eurytopic (Eu) silvicolous (si) (in woodland) and hygrophilous (hy).

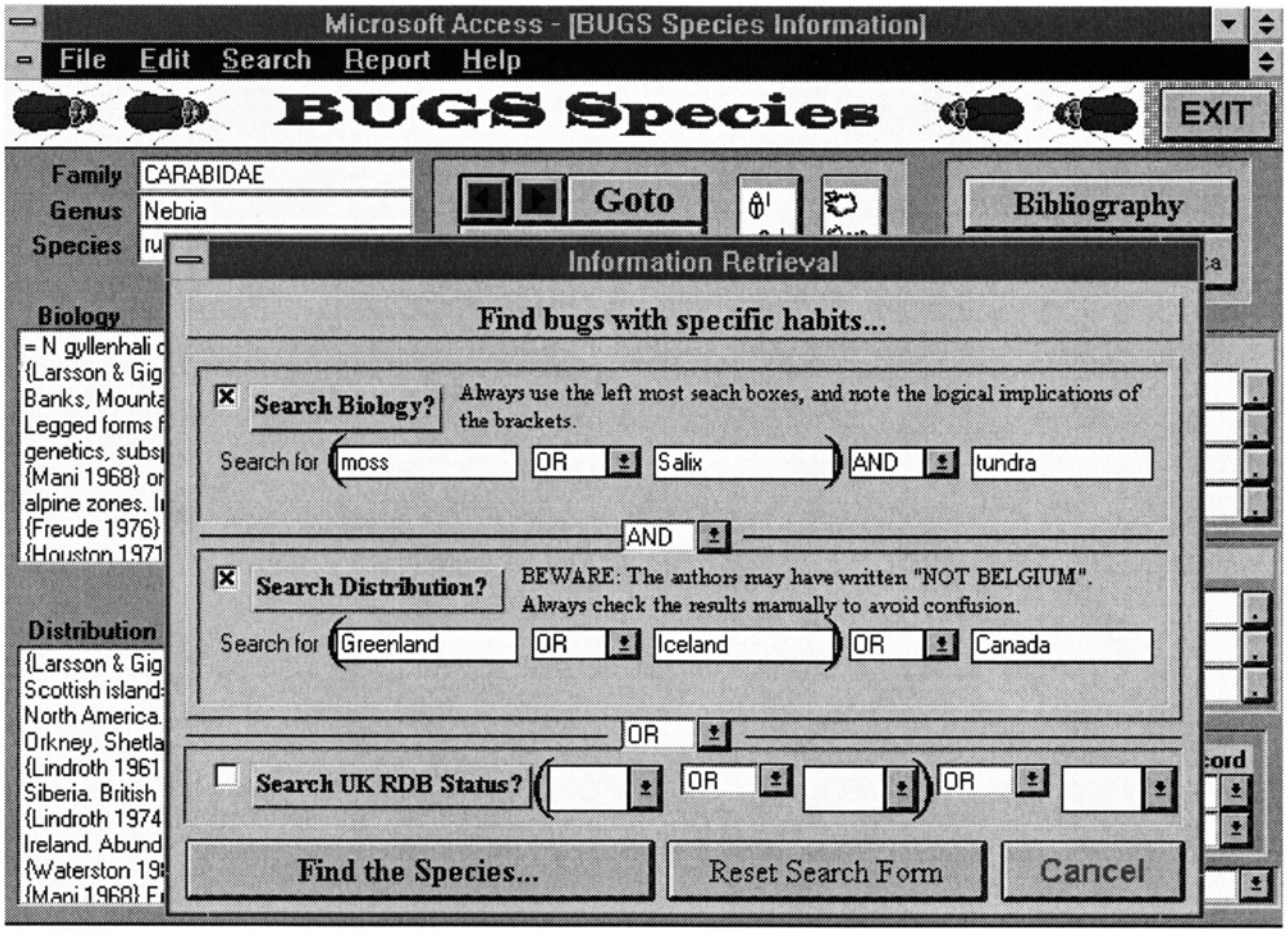


Figure 4. Finding species by their habitats.

Table 1. Ecocodes used in BUGS.

Group	Set	Description
Habitat & Ecology	Range	Distribution throughout habitat zones.
	Ecology	Specific ecological requirements, eg. (ri) by river banks.
	Type	General habitat characteristics, eg. (ht) salty environments.
	Seasonal	Fluctuations in habitat usage.
Food	Range	Broad food requirements.
	Type	Specific food sources, eg. (Pn) on pollen.
	Condition	Preferred state of food, eg. (Ol) starting to rot.

These codes are shown in boxes down the right hand side of the main screen (Fig.1). The button to the right of each box brings up a dictionary for each set, which also shows a check next to the appropriate codes, and lists the words and descriptions for each code. By selecting "Output ECOCODE Dictionary" from the "Report" menu, one can print or save the dictionaries for external reference.

With a system of coding implemented, the obvious next stage is to apply existing, or devise new statistical methods to the datasets from sites. These may be considered in two parts: summary measures — descriptive statistics; and interpretative tools, or comparative statistics.

i) Summary Measures

Descriptive statistics are extremely useful when there is a need to express complex data in a universally understandable language. The most commonly used methods, although not necessarily in ecology, relate to averages. BUGS is designed such that simple statistics could be created easily, either within Access or Excel. This could be done by outputting another file, additional to the spreadsheet, containing this information. However, a tidier method would be to use a page of the spreadsheet for summary data, which can be appended every time data are altered, or on demand. Excel © incorporates the Visual BASIC © programming language, which is more appropriate for statistical manipulation than Access BASIC ©, and has far greater scope for the later addition and alteration of statistical methods. However the data are manipulated, it is evident that more use can be made of the summarised ecology of a species list beyond any direct palaeoecological study. For example, one can calculate the modal codes within a sample or site species list in order to rank sites for conservation purposes.

A bar graph of code counts provides a visual means of comparing sites. Visualising the data is an important step in any analysis and the more angles the data are viewed from, the greater the breadth of analysis. Such a code-based classification is an extremely powerful tool. It can be used to identify patterns which represent certain 'types' of sample. For example, one can collate those samples

dominated by decomposer species; samples where carrion feeders are most prominent; or samples which include a superabundant species (*cf.* Kenward 1978). In his recent PhD thesis, Skidmore (1996) uses the characteristics of temperature dependency in populations of Diptera to define periods of habitation and abandonment on Norse farm sites in Greenland. In this context, flies are classed as exophilous or endophilous — *i.e.* outdoor or indoor — the latter being absent from samples dating from after the farm ceased to be occupied. Such a situation is an excellent example of how the grouping of insects, as defined by their habitat requirements, can aid in the interpretation of archaeological contexts.

ii) Comparative Statistics

Whilst it is true that the comparison of summary measures can go a long way towards analysing a site, there is often a demand for more quantitative correlation and comparative techniques. There are various relatively simple methods, such as the Jaccard or Modified Sorensen's correlation coefficients (*cf.* Southwood 1978), which compare differences between the presence and absence, or abundance of species between samples. From this a trellis diagram may be produced, which illustrates similarities in the construction of populations. Kenward (1978) has advocated the use of Fisher's α in examining species diversity in fossil assemblages, and both Perry (1981; Perry *et al.* 1985), and Sadler (Buckland *et al.* 1992) have employed multivariate techniques. There are also methods available which relate directly to the ecology of the species in question, and it is here that the future of BUGS lies. In the development of comparative methods for insect ecology statistics much is to be said for the automation of palaeoenvironmental analysis. One of the problems encountered is that of the visualisation of multidimensional data. Various statistics programs, such as Decorana, use composite layering systems where datasets can be displayed in various combinations on a two dimensional plane. The graphing facilities of Excel make it possible to do this simply, and in addition allow the use of rotatable three dimensional plots, which, although often difficult to extract detailed data from, are excellent for transmitting information quickly. It may be necessary to create further groupings and orderings within the ecocoding system, so that, for example, graphs can have axes where diametrically opposite environmental components are represented as the extremes. Samples could then be shown in three dimensional space, as defined by their assemblage's ecology construction. The use of computers inevitably increases the speed with which graphics can be produced, and hence allows for more experimentation, and therefore innovation. This is how new methods can emerge.

iii) Simple Reports

Whereas the DOS based version of BUGS relied heavily on the separation of specific data files, which greatly influenced the design of the interface, the new BUGS is designed to reflect the probable needs of the user, rather than the data structure. The ease with which display boxes and windows can be created gives far greater freedom of expression within programming. It would seem that the only limits to display are aesthetic ones. The same would appear true for report creation. At present BUGS facilitates three primary report types :

1) The complete information for any one species — to file or printer.
2) A report which extracts the site data, species list and the biology, distribution and ecology codes relating to that list.
3) A customisable report which outputs the result of a habitat/distribution query.

Whilst in no way can these replace the structured final report of any investigator, it at least allows researchers and students alike to obtain quickly a transcript of specific species information. As a result, one can concentrate on detailed interpretation, rather than on another intensive initial literature search, although the bibliography in BUGS can always be improved upon, the additional data being added to the relevant fields.

Databases and Developments

Another improvement, one enabled by the development of faster and larger computers, is the use of words where there were previously codes. For example, instead of "LL", as the DOS BUGS would show, the full term "Loch Lomond" can be selected from a list of full fossil record terms. The apparently paradoxical use of codes for words within the ecology output is a function of speed for statistical manipulation and ease of data display — you can see more codes than words per unit screen. In the past, storage space limited the size of data fields — the columns in our tables of virtual information — but it is now evident that storage space is unlikely to be a problem. Despite this optimism for the future of mass storage, we are still faced with the problem that uncompressed bitmap images of every species in BUGS could amount to over 1 Gigabyte. It is perhaps inevitable that the BUGS of the future will have to be produced in regional editions.

As part of a National Science Foundation (USA) Office of Polar Programs project through NABO, the North Atlantic Biocultural Organisation, BUGS has been made available over the Internet from NOAA, the World Data Centre — A for Palaeoclimatology. To download the data, the site is :

http://www.ngdc.noaa.gov/paleo/insect.html

It is envisaged that within the next year BUGS will also be available from web sites in New York, Edinburgh and Umeå, Sweden, as part of the NABO project. Further information may be obtained from the first author of this paper at phpbud@student.umu.se, from whom copies may be obtained at cost.

Acknowledgements

Initial funding for the DOS version of BUGS was provided by a NERC CASE studentship, with the Scottish Development Department, to Jon P. Sadler, who worked with Mike Raines and Paul C. Buckland on program design and data entry. Additional funding was provided by the Curriculum Development Fund of the University of Sheffield to enable BUGS to be employed in undergraduate teaching modules. The basic Windows version in Access was prepared by Yuan Zhuo Don as part of his M.Sc. in

Computer Science at the University of Sheffield. Funding from the NABO project enabled Philip I. Buckland to expand the program and develop the connections to Excel.

References

BUCKLAND, P. C. & COOPE, G. R. 1991. *A Bibliography and Literature Review of Quaternary Entomology*. J. Collis Publications, University of Sheffield.

BUCKLAND, P.C., SADLER, J. P. & SVEINBJARNARDÓTTIR, G. 1992. Palaeoecological Investigations at Reykholt, Western Iceland. In, C. J. Morris & D. J. Rackman (eds.) *Norse and Later Settlement and Subsistence in the North Atlantic* , 149–168. Dept. of Archaeology, University of Glasgow.

HYMAN, P. S. 1992; 1994. A review of the scarce and threatened Coleoptera of Great Britain, Parts 1 & 2. (Revised & updated by M.S.Parsons). UK Joint Nature Conservation Committee, Peterborough.

KENWARD, H. K. 1978. *The Analysis of Archaeological Insect Assemblages : a New Approach*. Archaeology of York, 19/1. Council for British Archaeology for York Archaeological Trust.

KOCH. K. 1989; 1990. *Ökologie*, 1–3. *Die Käfer Mitteleuropas*. Goecke & Evers, Krefeld.

LUCHT, W. H. 1987. *Katalog. Die Käfer Mitteleuropas*. Goecke & Evers, Krefeld.

PERRY, D. W. 1981. Cluster analysis of the insect remains beneath the "Raft". In S. McGrail (ed.) *The Brigg "Raft" and her Prehistoric Environment*, 176–182. British Archaeological Reports 89, Oxford.

PERRY, D. W., BUCKLAND, P. C. & SNÆSDÓTTIR, M. 1985. The Application of Numerical Techniques to Insect Assemblages from the Site of Stóraborg, Iceland. *Journal of Archaeological Science* 12, 335–345.

SADLER, J. P., BUCKLAND, P. C. & RAINES, M. 1992. BUGS: an entomological database. *Antenna* 16, 158–166.

SKIDMORE, P. (1996) *A Dipterological Perspective on the Holocene History of the North Atlantic Area*. Unpubl. Ph.D., University of Sheffield.

SOUTHWOOD, T. R. E. (1978) *Ecological Methods with particular reference to insects*. Methuen, London.