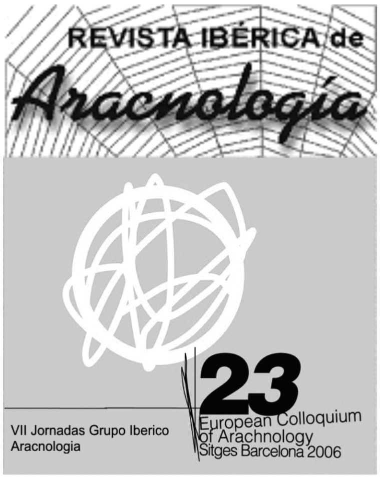
REVISTA IBÉRICA de
Aracnología
23
VII Jornadas Grupo Iberico Aracnologia
European Colloquium of Arachnology
Sitges Barcelona 2006

ARTÍCULO:

An 'Expert System' approach to digitising ecological information on spiders for habitat assessment.

Myles Nolan

48 Rathmines Road Upper,
Rathmines, Dublin 6, Ireland
mylesnolan@hotmail.com
e-mail: mylesnolan@hotmail.com
spidermanmyles@yahoo.ie

Revista Ibérica de Aracnología
ISSN: 1576 - 9518.
Dep. Legal: Z-2656-2000.
Vol. **15**, 30-VI-2007
Sección: Artículos y Notas.
Pp: 137 − 141.
Fecha publicación: 30 Abril 2008

Edita:
Grupo Ibérico de Aracnología (GIA)
Grupo de trabajo en Aracnología
de la Sociedad Entomológica
Aragonesa (SEA)
Avda. Radio Juventud, 37
50012 Zaragoza (ESPAÑA)
Tef. 976 324415
Fax. 976 535697
C-elect.: amelic@telefonica.net

Director: Carles Ribera
C-elect.: cribera@ub.edu

Indice, resúmenes, abstracts
vols. publicados:
**http://entomologia.rediris.es/sea/
publicaciones/ria/index.htm**

Página web GIA:
http://entomologia.rediris.es/gia

Página web SEA:
http://entomologia.rediris.es/sea

An 'Expert System' approach to digitising ecological information on spiders for habitat assessment.

Myles Nolan

Abstract

This paper introduces a database system, currently under development, which allows for the prediction of the spider fauna that one should expect in a given habitat of good quality. With Excel spreadsheets and supplementary Word files forming the basic structure of the system, hierarchically stacked categories are scored for maximal and minimal associations between species and macrohabitats, microsites and traits using a fuzzy coding method. The compilation of species accounts and subsequent coding of files will continue through 2007, by end of which it is hoped to have data on three-quarters of the Irish fauna (about 300 species) encoded. This should be sufficient to allow for field testing.

Keywords: database, spiders, habitat assessment, Irish fauna.

Aproximación a un "sistema experto" de digitalización de la información ecológica en arañas para la evaluación del hábitat.

Resumen

Este artículo propone un sistema de bases de datos, actualmente en desarrollo, que permite la predicción de la fauna de arañas esperada en un hábitat determinado. La estructura base del sistema está formada por archivos Excel y otros archivos suplementarios de Word. Las categorías jerárquicas acumuladas se valoran para las asociaciones máximas y mínimas entre las especies y los macrohábitats, los microhábitats y características, utilizando un método de codificación "fuzzy". La compilación de la información de las especies y la subsecuente codificación de archivos continuará en el 2007, al final del cual se espera tener codificados los datos sobre las tres cuartas partes de la fauna de arañas de Irlanda (cerca de 300 especies). En la práctica esto debería ser suficiente par probar el sistema.

Palabras clave: base de datos, arañas, evaluación del hábitat, fauna de Irlanda.

Introduction

Spiders are sufficiently diverse as a group to reflect well the structural and environmental complexity present in a chosen habitat (e.g. Bell et al. 1998; Bonte et al., 2002). However the compilation of invertebrate inventories and subsequent assessment of site quality is beset by a significant difficulty which is the inability to predict those species that should occur in a given habitat. With respect to spiders, only occasionally is a site or habitat assessed where the spider fauna can be stated to be very well understood in advance of survey (Scharff et al., 2003). Yet it is well known that a large range of spider species have strong associations with specific habitats (Hänggi et al., 1995). And it is certainly the case that an individual with experience of a country's spider fauna, even if working in an unfamiliar habitat, will have some idea in advance of survey, as to the range of species they may expect to find. There is not available however, a resource which allows a list of species, known to have an association with a particular habitat, to be derived quickly and thus provide a hypothetical target against which the recorded fauna from a habitat may be assessed. It is desirable as a goal of biodiversity thinking (admitting that the term encompasses a broad spectrum of varying definitions and targets) to be able to hypothesise about what is missing from a habitat of diminished quality and relate this to specific factors within that habitat. Survey work (limited by funding, time and manpower constraints) is generally discontinuous and surveys of similar habitats are frequently difficult to compare as a result of differing trapping methodologies and varying statistical approaches to rendering data. The database system introduced here relies on presence/absence of species to interpret habitat quality, with the understanding that the trapping methodology utilised must be able to accurately reflect the local fauna.

If the information found in published literature and, the knowledge of experts can be pooled into a single electronic resource in such a way as to take notice of degrees of affiliation between spider species and habitats (and a range of other factors), a vast amount of data would be readily available for usage in habitat assessment.

Since this system is, with respect to spiders, in the early stages of development the statement given here should be taken as one of intent. Much of the information offered here is based on the experience of those who constructed and now utilise a database which was developed for the Hoverflies (Syrphidae: Diptera) (Speight et al., 2006) and has served as the model for developing the spider version.

System structure

The database is comprised of a series of sheets within an Excel file into which information is categorised and coded. These sheets are divided under the following headings: Range and Status, Forest Macrohabitats, Open Ground & Wetland Macrohabitats, Microsites and, Traits. The Range and Status sheet presents information on the geographical range of a species (encompassing every European country and the major global biogeographic divisions) and a comment on the taxonomic status of species, represented by a coded indication as to how well a particular species is understood in the available literature. The Forest and Open Ground & Wetland Macrohabitats comprise a breakdown of the wide range of habitats governed by these umbrella terms. Table I illustrates how part of a coastal system can be broken down into macrohabitats, with Open Ground & Wetland Habitats functioning as the primary category.

Hierarchical subdivision into macrohabitats of the beach/dune section of a coastal system.
Taken from Speight *et al.* (2006)

Secondary	Coastal beaches & dunes (gen.)							
Tertiary		Coastal beaches (gen.)		Coastal dunes (gen.)				
Quaternary			shingle		*Ammophila* dunes	grey dunes/ dune scrub	Machair	dune slacks

The Microsites sheet is categorised and stacked in similar fashion but, since a microsite may occur in a range of habitats, functions independently of the Macrohabitats sheets. The designation of a microsite can be difficult for a range of reasons that are beyond the scope of this introduction to the system. Briefly however, it is well understood that a single species of spider may make use of a range of microsites throughout its lifecycle e.g. adult web location, preferred egg-sac location, over-wintering location. In the StN database (Speight et al., 2006) categories within the Microsites sheet were coded only for breeding sites of syrphid larvae, the reasoning being that the presence in a habitat of appropriate breeding sites was the only guarantor of continued breeding success; the absence of a breeding microsite has a negative impact on a species and the presence of same conversely has a positive effect. Because of the difficulties (sometimes insuperable) inherent in identifying either spider egg-sacs or juveniles to species, without resorting to breeding programmes, it is probably not possible to construct a microsites file for spiders solely with respect to the locations where these early stages occur with greatest frequency. As such, microsites in the spider database pertain to adults and to a lesser extent to other phases of the life cycle.

The Traits sheet contains a somewhat heterogeneous dataset including information on behaviour, responsiveness to trap methods and phenological data.

The Excel spreadsheets are accompanied by two Word files; Species Accounts and Contents & Glossary. The first of these contains relatively brief but comprehensive accounts of individual species that summarise information found in the database, outlining a species' distribution, life-cycle, preferred environment and containing a comment when necessary as to their conservation status. The other file is an adjunct of the database proper, containing amongst other informations, tabulated versions of the categorised hierarchies (retaining the hierarchical structure) contained in each file and glossaries of definitions appertaining to each file – each category having its own definition.

System anatomy

It would be correct to infer from Table I that within the Excel sheets categories are stacked hierarchically along the top rows and thus species will be listed in the left-most column. The hierarchical arrangement means that data coded into a specific cell may pertain to a primary, secondary, tertiary or quaternary category, the quaternary representing an entity the further division of which should be unnecessary. An essential aspect of the system is that every category in the various files is defined unless absolutely unnecessary. As well as being listed in the glossaries of definitions in a text file, definitions can be found in the relevant sheet as a comment attached to the cell in which the category is located, thus allowing them to be checked while the spreadsheet is in use. The purpose of the definition is to explain the category as clearly as possible to an individual using the system.

The microsites favoured by larval syrphids are for obvious reasons going to differ from those relevant to spiders and thus it has been necessary to define a series of microsites specifically for the spider version of the system. This is not the case however for the Macrohabitats sheets, which have been adopted into the spider database and are further refined only when this proves necessary. Recognisant of the fact that the CORINE biotopes manual (1991) represented the only officially endorsed and widely available system which attempted to distinguish habitats on a European scale, the definitions therein were used to assist in the construction of the original StN Macrohabitats files. However, it was found that because the CORINE definitions relied usually on botanical information, they sometimes did not adequately represent what was important to the Syrphidae at habitat level. Consequently refinements were added to some definitions while other ‘macrohabitats’ had to be recognised as such and a definition coined for them. Because the Macrohabitats file from the StN database has served as the model for the spider database, the latter also at present includes the CORINE definitions. However, a more recent classification of european habitat types, EUNIS (http://eunis.eea. europa.eu/habitats.jsp), is being studied in relation to the CORINE system and may also be acknowledged.

Coding the spreadsheets

Data is coded into the Macrohabitats, Microsites and Traits sheets using fuzzy coding. Fuzzy coding is obviously associated with fuzzy logic, a branch of set theory which deals with approximate reasoning. For the purpose of this brief account information was derived from Wikipedia (http://en.wikipedia.org/wiki/Fuzzy_logic). Fuzzy coding is felt to be useful in the context of the database because it can express a species’ relation to a category as a partial membership rather than as a probability. The scoring in the cells indicates the degree of affiliation between a species and category with a blank cell suggesting zero affiliation and a score of 3 a maximal affiliation (see Table II). This system of coding allows for the indication of approximate relationships between a species and a category and thus recognises the elasticity often characteristic of interactions between invertebrates and habitats.

Naturally, it must be pointed out that the quality of information available on particular species is highly variable and for many species there is a distinct lack of information. One of the inevitable consequences of the databasing system is the detection of such deficiencies. It is integral to the ‘expert system’ approach that such deficiencies may be addressed through consultation with experts who may be willing to offer relevant information.

The Range & Status file is coded differently from the other files. Occurrence in a country or geographical region is coded with a categorical 1 or 0 (blank cell). In order to comment on the taxonomic status of a species it is necessary to alter the scoring slightly, using 1, 2, 3, and 4 and retaining no blank cells. This is necessary because if a species has been named it has some taxonomic status, regardless of how poorly it may have been described. A blank cell would indicate the non-existence of a taxon. As such a score of 1 would indicate that a species is very poorly understood or that there is considerable confusion surrounding its status as a species. A species that is very well understood e.g. where a reliable recent revision of the species (and ideally of its congeners) is available would get a score of 4.

Table II
Fuzzy coding as used in the Macrohabitats, Microsites and Traits files
indicating the degree of association between species and categories.

Coding	Category: Macrohabitat, Microsite, Trait
0 / (Blank cell)	do not expect this species / no affiliation
1	do not expect this species but it may occur / some affiliation
2	expect this species but it may not occur / good affiliation
3	expect this species / strong affiliation

Using the database

The path by which the database is to be used in accompaniment with habitat survey involves following a series of steps which can be briefly summarised thus: site assessment, generation of a predicted list for the habitats represented, survey of those habitats and, comparison of predicted / observed results.

Once one is certain which macrohabitats are going to form the basis of the survey, and these have been identified within the database then it is possible to generate a predicted list. Briefly this involves selecting the appropriate macrohabitat columns from the Macrohabitats sheets, dropping them into a new Excel file with the species list and deleting those species scored with a blank cell for the macrohabitats in question.

The StN database, since it encompasses a large area of Europe, contains information on a large number of species not represented in all European countries. As such it is necessary to delete all those species not known to occur in the country wherein the survey will be located. Having generated this first list of predicted species it is then necessary to further refine it and this is done ideally by using a regional list; in the case of Ireland this would be a county list. Even on a relatively small body of land such as Ireland there can be considerable differences between the fauna of the northern and southern extremes of the country. Predicting for a specific site within a country and region thereof thus takes a realistic view of what can reasonably be expected to occur, on the basis of previous survey and collecting effort. The more comprehensive the local faunal lists, the greater coincidence there will be between species predicted and those recorded. Even if a local list is poor however, information from survey utilising the predictive system can be readily absorbed into the local list and thus improve the prediction quality on subsequent survey. Regional lists are not encompassed within the database, usually being maintained, if at all, by a small number of individuals within a given country.

Having generated a predicted list appropriate to region and habitat the next stage is to carry out the survey. While a standardised methodology should be used, the question of what constitutes such with respect to spiders is very much part of an ongoing debate (Scharff et al. 2003) and well beyond the scope of this introduction. Broadly speaking however it is assumed that assessing habitat quality should mean attempting to record as many species as possible from the habitat (despite the usually limited survey time available) and with respect to spiders this means deploying a range of capture methods, especially in the case of habitats with tall vegetation or trees. Having completed the survey and identified all species recorded it is then possible to compare the predicted list with the recorded.

Initial comparison of the predicted and observed fauna generates three lists: species predicted but not observed, species predicted and observed and, species not predicted but observed. At this point it is appropriate to return to the database and, using the Species Accounts and Traits and Microsites files, attempt to explain the absence of those species which were predicted but not recorded and the presence of those species which were not predicted but were recorded.

A compendium of expert knowledge

The database as designed is founded fundamentally on an understanding of a species as a complex organism which has a minimum set of requirements the satisfaction of which will allow continuity of generations.

While the system has been described throughout as a database this term is used more loosely today than when it was originally coined within the computer industry. A possible definition suggests that a database is "a collection of records stored in a computer in a systematic way, so that a computer program can consult it to answer questions." (http://en.wikipedia.org/wiki/Database). The spider database cannot be consulted in this particular manner and thus could perhaps be better described as a compendium of expert knowledge. The term compendium is appropriate since the system presents "a concise yet comprehensive compilation of a body of knowledge." (http://en.wikipedia.org/wiki/Compendium). The term 'expert' is recognisant of the fact that the data entered into the system files is the product of research by individual experts, regardless of the form in which this data exists or the manner of its transmission into the database. The term 'expert system' originated with the practise of transforming knowledge based systems into computer programs. Put simply, knowledge – 'real world values' – is input as data which a computer program can then rapidly manipulate.

The spider database can however only properly embrace this 'expert' quality once it has absorbed significant information from published sources and it is felt that the various files are reliable, comprehensive and well founded. It is hoped that once the spider database reaches this stage, individual experts may be willing to assist in offering information on species which is presently unavailable, may not be readily published and may otherwise never see light of day.

Development of the spider database

As was stated earlier the database system was first developed for the Hoverflies and has expanded greatly over the last fifteen years, with the most recent version providing coverage of over 600 species in a very large area of Europe outside of Russia (Speight et al., 2006). The system is available for usage upon signing of a user agreement. The spider database was initiated through a short feasibility study carried out in 2005 for the National Parks and Wildlife Services, Ireland, which looked broadly at whether the system could be developed effectively to accomodate spiders. On the basis of that study a contract was agreed for 2006 which allowed a significant effort to be made in the development of the various files and the compilation of species accounts. This contract has been continued into 2007 the aim being to have 300 species found in Ireland represented in the database by end of that year. This number would account for a little under three-quarters of the presently known Irish spider fauna and would be sufficient to allow for testing of the system. At present this database is restricted to the Irish fauna but information on species is drawn from a wide range of European literature and it is hoped that at some stage it will be possible to extend the project to cover by degrees a greater part of the European spider fauna.

Acknowledgements

I am very grateful to the National Parks and Wildlife Services, Ireland, who are funding the development of the database and also funded my attendance at the 23rd European Colloquium of Arachnology in Spain. I am equally grateful to my principal point of contact in the NPWS, Martin Speight, who was instrumental in initiating the feasibility study in 2005 and overseeing its continuation thereafter.

References

Bell, J.R., Haughton, A.J., Cullen, W.R. and Wheater, C.P. 1998. The zonation and ecology of a sand-dune spider community. In Selden, P. A. (ed.). *Proceedings of the 17th European Colloquium of Arachnology*, Edinburgh 1997, 267–72. British Arachnological Society, Burnham Beeches, Bucks.

Bonte, D., Baert, L. and Maelfait, J.-P. 2002. Spider assemblage structure and stability in a heterogeneous coastal dune system (Belgium). *Journal of Arachnology* **30**: 331-343.

Corine Biotopes Manual, 1991. Office for Official publications of the European Communities, Luxembourg.

Falkner, G., Obrdlik, P., Castella, E., and Speight, M.C.D. 2001. *Shelled Gastropoda of Western Europe* (Friedrich-Held-Gesellschaft: München)

Hänggi, A., Stöckli, E. and Nentwig, W. (1995) Lebensräume Mitteleuropäischer Spinnen. Habitats of central European spiders. *Miscellanea Faunistica Helvetiae* **4**. Centre Suisse de cartographie de la faune.

Scharff, N., Coddington, J. A., Griswold C. E., Hormiga, G. and Place Bjorn, P. de. 2003. When to quit? Estimating spider species richness in a Northern European deciduous forest. *Journal of Arachnology* **31**: 246-273

Speight, M.C.D., Monteil, C., Castella, E. & Sarthou, J.-P. 2006. StN Ferrara 2006. In: Speight, M.C.D., Castella, E., Sarthou, J.-P. & Monteil, C. (eds). Syrph the Net on CD, Issue 5. *The database of European Syrphidae*. ISSN 1649-1917. Syrph the Net Publications, Dublin.

Speight, M.C.D., Castella, E., Obrdlik, P. & Schneider, E. 1997. *Are CORINE habitats invertebrate habitats?* In: Schreiber, H. (ed.)