Behav Ecol Sociobiol (2009) 63:765–778
DOI 10.1007/s00265-008-0697-y

METHODS

SOCPROG programs: analysing animal social structures

Hal Whitehead

Received: 11 June 2008 / Revised: 24 September 2008 / Accepted: 9 December 2008 / Published online: 16 January 2009

Abstract SOCPROG is a set of programs which analyses data on animal associations. Data usually come from observations of the social behaviour of individually identifiable animals. Associations among animals, sampling periods, restrictions on the data and association indices can be defined very flexibly. SOCPROG can analyse data sets including 1,000 or more individuals. Association matrices are displayed using sociograms, principal coordinates analysis, multidimensional scaling and cluster analyses. Permutation tests, Mantel and related tests and matrix correlation methods examine hypotheses about preferred associations among individuals and classes of individual. Weighted network statistics are calculated and can be tested against null hypotheses. Temporal analyses include displays of lagged association rates (rates of reassociation following an association). Models can be fitted to lagged association rates. Multiple association measures, including measures produced by other programs such as genetic or range use data, may be analysed using Mantel tests and principal components analysis. SOCPROG also performs mark-recapture population analyses and movement analyses. SOCPROG is written in the programming language MATLAB and may be downloaded free from the World Wide Web.

Keywords Social analysis · Software · Association

Communicated by L. Z. Garamszegi.

H. Whitehead (✉)
Department of Biology, Dalhousie University,
Halifax, Nova Scotia, Canada B3H 4J1
e-mail: hwhitehe@dal.ca

Introduction

One of the most important attributes of any animal population is its social structure. Social structure (here synonymous with "social organisation") embodies a significant segment of interactions among organisms: those among nearby conspecifics. Social structure can affect population growth rates, dispersal and gene flow (e.g. Strier 1997) and is often an important factor in management and conservation (Sutherland 1998).

Hinde (1976) defines the social structure of a population to be the content, quality and patterning of the relationships among its members, where a relationship between a pair of animals is defined by the content, quality and temporal patterning of the interactions between them. Therefore, in order to study the social structure of a population, we need to gather data on interactions between identified individuals, assemble these to depict relationships and then synthesise the measures of relationship into a model of social structure (Whitehead 1997; Whitehead 2008b). However, interactions (actions directed towards, or affecting, the behaviour of another animal) may be hard to observe in many circumstances. For this and other reasons, associations (dyadic states) are often used in place of interactions (dyadic events) as the foundation of studies of social structure. This is a defensible simplification if "association" is defined so that the majority of interactions take place between associated individuals (Whitehead 1997).

A number of authors, including Altmann (1974), Lehner (1998) and Martin and Bateson (2007), have described methods of collecting interaction or association data in an unbiased manner, and there are computer programs, such as "JWatcher" (Blumstein and Daniel 2007) and the "Noldus Observer" (Visser 1993), which facilitate this. Using

Hinde's (1976) conceptual framework, Whitehead (1997, 2008b) describes an analytical framework for the next steps in the analysis of social structure, building relationship measures and synthesising them into displays and models of social structure. The computer program, SOCPROG, which is described in this paper, provides a coherent tool for the numerical analyses which make up this framework. Some of these analyses (including cluster analyses and multidimensional scaling) can be carried out using standard statistical packages, such as SAS, SPSS or R. Specialised packages, such as MatMan (Noldus Information Technologies 2003) and UCINET (Borgatti et al. 1999), can help with other parts of the analysis, such as Mantel tests and network analyses, respectively. Other informative analyses, such as permutation tests for preferred companionship (Bejder et al. 1998) and lagged association rates (Whitehead 1995), are not implemented as part of a package, other than in SOCPROG.

In addition to the benefits of a coherent package, another factor provided impetus for the development of the SOCPROG programs: large data sets. Many of the first attempts to use Hinde's (1976) conceptual framework to model social organisation referred to communities of primates (Cheney et al. 1987). These usually contained on the order of ten to 30 individuals (Whitehead and Dufault 1999). More recently, similar methods have been attempted with populations of large ungulates or cetaceans containing hundreds, and occasionally over one thousand, identified individuals (Whitehead and Dufault 1999). If symmetric associations are measured between n individuals over m sampling periods, then this amounts to $(n-1)\times(n-2)\times m/2$ data elements. With 100 sampling periods and 20 individuals, the association data constitute a manageable array of size 17,100 elements. However, with 1,000 individuals, it rises to about 5.10^7 elements, a computational challenge.

SOCPROG is designed to provide flexible and fairly comprehensive analyses of social structure using data on the associations or interactions of identified individuals, which can number in the thousands. It uses graphical user interfaces, so that operation is straightforward, but the programs can also be adapted quite easily to carry out new analyses. It is designed so that analyses can be repeated easily after changes in attributes such as the length of the sampling period or the definition of association or restrictions (such as to sex or age of the individuals or season of year) are applied to the data. Such modified data sets can also be saved for future analyses. SOCPROG is written in the language MATLAB, a high-performance language for technical computing which is heavily used throughout the sciences and engineering. It has a wide range of matrix manipulation and other functions as well as excellent, and highly flexible, graphics capabilities, including graphical user interfaces (e.g. Fig. 1). This makes it possible to design easy-to-use, but adaptable and fast, routines for manipulating large amounts of data. The extensions in the MATLAB statistics toolbox are useful in a number of ways, but especially in the provision of cumulative probability distribution functions, directly giving the probability of values of test statistics (or more extreme values) under many null hypotheses and thus the

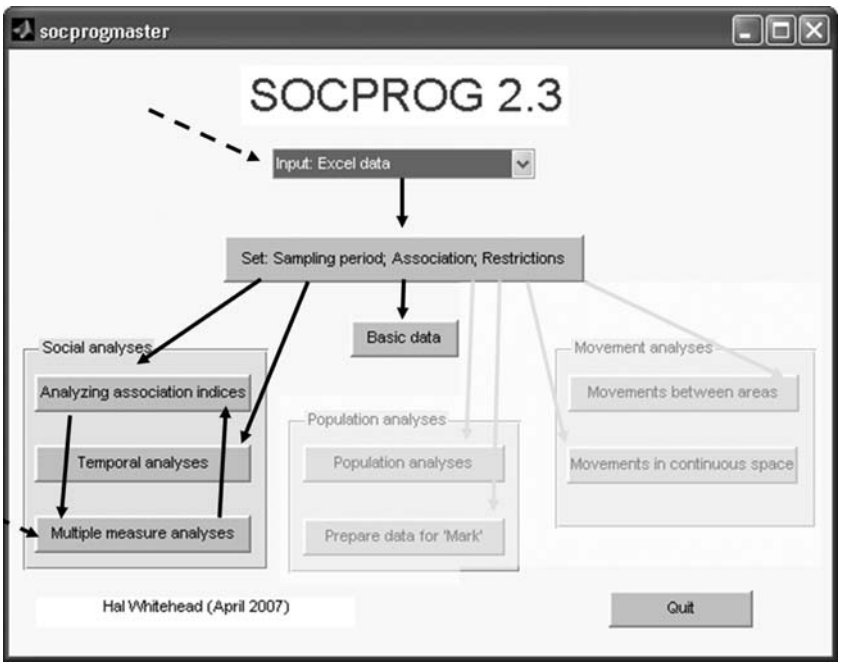

Fig. 1 Master graphical user interface of SOCPROG, showing the modules and the routes by which data are input, and flow between modules. The social analysis modules (*not faded*) are those described in this paper. Clicking on any of the pushbuttons starts that module

significance levels of hypothesis tests. A particularly important attribute of MATLAB for the SOCPROG programs is the availability of sparse matrix routines which allow efficient storage and manipulation of large matrices consisting largely of zeros, as is often the case with association data.

The major elements of SOCPROG are shown in Fig. 1. The structure of the package is based on the analytical framework described by Whitehead (1997, 2008b); please consult these publications for advice on analytical methods as well as their limitations. In this paper, I describe the use of the SOCPROG programs (version 2.3) for social analyses, illustrating some of the output of SOCPROG using a data set which is provided with the package. SOCPROG also carries out some mark-recapture population analyses (and prepares data for the program MARK; White and Burnham 1999) as well as analyses of movements of identified individuals either in continuous space or between discrete areas (described in Whitehead 2001).

SOCPROG has been used in 48 papers published in refereed journals (Table 1), primarily for social analysis, and primarily for cetaceans. Thirty-nine per cent of the papers describing social analyses were published in the major animal behaviour journals (Animal Behaviour, Applied Animal Behaviour Science, Behavioral Ecology, Behavioural Ecology and Sociobiology) and an additional 29% in general ecological or biological journals (Canadian Journal of Zoology, Journal of Animal Ecology, Proceedings of the Royal Society B, Research Letters Ecology). The use of SOCPROG appears to be increasing (Table 1). However, to date, published analyses of social structure using SOCPROG only refer to mammals, and largely to cetaceans (for which association data are particularly important), although I am aware of its use (unpublished, in 2008) for studying social structure in non-mammal taxa.

SOCPROG (including its MATLAB code) can be downloaded free from the World Wide Web (see Appendix). Whilst the standard version of SOCPROG requires MATLAB plus its statistics toolbox (version 7.4), which must be purchased, a compiled version of SOCPROG, which does not need MATLAB (but possesses some limitations), is also available.

The following sections are indexed by the modules of SOCPROG that process and analyse social data, as indicated in Fig. 1. Arrows in Fig. 1 show the flow of data between modules.

Data input

Input of observation data from Excel worksheets

The principal source of the input data used by SOCPROG is the Excel worksheet. The "primary data file" contains lines, or records, each of which corresponds to an observation, either of an individual (linear mode) or a group (group mode). Each record contains an observation of an individual or group, usually with a variety of other information (e.g. date, time, position, behaviour, group identifier, quality of identification; see Tables 2 and 3). The identification code(s) of the individuals, which can be numeric ("13" or "19"; Table 3) or alphanumeric ("A1" or "N14"; Table 2) are in the final field (column). It is not necessary that all members of each group are recorded, although a few analyses (typical group sizes, lagged association rates) will not make sense if this is not the case. SOCPROG can convert group mode data to linear mode data, but not vice versa.

Input of supplemental data from Excel worksheets

Analyses are richer with supplemental data about individuals, such as sex, age or genetic data (e.g. haplotype). These can be input from another Excel worksheet (e.g. Table 4).

Table 1 Use of SOCPROG in papers published in journals with referee system (found using Google Scholar search on 14 Sept 2008) by type of analysis, and, for social analyses, by taxon, year and journal

All papers	48
Movement analyses[a]	5
Population analyses[a]	4
Social analyses[a]	41

Taxon: bat (3); buffalo (1); cetacean (25); elephant (1); pig (1); primate (1); wombat (3); theoretical/technical (6)

Year: 1999 (1); 2001 (4); 2002 (2); 2003 (6); 2004 (3); 2005 (5); 2006 (3); 2007 (6); 2008 (January–September, 11)

Journal: *American Journal of Primatology* (1); *Animal Behaviour* (11); *Applied Animal Behaviour Science* (1); *Behavioral Ecology* (1); *Behavioural Ecology and Sociobiology* (3); *Canadian Journal of Zoology* (6); *Communications in Statistics* (1); *Journal of Heredity* (1); *Journal of Mammalogy* (2); *Journal of the Marine Biology Association, UK* (1); *Journal of Animal Ecology* (1); *Marine Mammal Science* (4); *Molecular Ecology* (3); *Proceedings of the Royal Society B* (4); *Research Letters Ecology* (1)

[a] A few papers used SOCPROG for more than one type of analysis

Table 2 Primary data in linear mode as encoded in an Excel worksheet, with fields giving date, position (along a linear transect), behaviour and alphanumeric individual identities

Date	Position	Behaviour	ID
1/1/00 9:00	279.9	X	A1
1/1/00 9:00	279.7	X	I9
1/1/00 9:00	278.2	Y	N14
1/1/00 9:00	280	X	O15
1/1/00 12:00	42.6	Y	H8
1/1/00 12:00	40.3	Y	K11
1/1/00 12:00	42	Y	M13
1/1/00 12:00	41.1	Z	T20
1/1/00 15:00	664	X	D4
1/1/00 15:00	663.6	X	G7
1/1/00 15:00	664	X	L12
1/1/00 15:00	664.8	Y	Q17
1/1/00 15:00	663.6	Y	S19
1/2/00 9:00	325	Z	A1
1/2/00 9:00	325.9	Z	I9

Other input formats and structures

Although Excel worksheets are the recommended format for primary and supplemental data, they can also be entered as ASCII files.

A primary output of SOCPROG is the matrix of association indices or interaction rates (as described later) which is a square matrix describing the relationships between the individuals in the study population. However, these matrices of "relationship measures" can also be input directly (either initially, or straight into the "Multiple measure analyses" module; Fig. 1) from Excel worksheets or as ASCII files (e.g. Table 5) if calculated by hand or output by another computer package. For instance, estimates of genetic relatedness among members of a population that are produced by the KINSHIP software (Queller and Goodnight 1989) using molecular genetic data are of this format and can be input into SOCPROG.

Table 3 Primary data in group mode as encoded in an Excel worksheet, with fields giving date, location (three subareas), behaviour (five categories) and numeric individual identities of group members

Date	Location	Behaviour	Group
1/1/00 9:49	A	2	8 11 13 20
1/1/00 14:54	A	1	1 9 14 15
1/1/00 15:41	A	2	4 7 12 17 19
1/2/00 9:11	B	1	4 7 12 17 19 20
1/2/00 9:41	B	1	2 10 18
1/2/00 10:09	A	3	3 5 6 16
1/3/00 10:35	A	5	2 10 18
1/3/00 11:03	A	3	4 7 12 17 19 20
1/3/00 14:32	A	2	5 6 16
1/3/00 17:40	A	1	1 9 14 15 8
1/4/00 7:16	A	2	4 7 12 17 19 20
1/4/00 13:17	A	2	11 13 3
1/5/00 6:00	A	2	1 9 14 15 8
1/5/00 15:57	B	2	5 6 16 10
1/5/00 17:55	B	2	11 13 3
1/11/00 7:19	C	2	2 18 12
1/11/00 10:09	A	4	1 9 14 15 8 4
1/12/00 7:14	C	2	1 14 15 8 4
1/12/00 9:01	B	2	5 6 16 10

Viewing and saving the data

Once primary, and optionally supplemental, data have been entered, a window appears with summary information about the file: primary data file name; data mode: linear or group; number of records in the primary data file; list of primary data file fields; number of individuals; list of supplemental data file fields, if entered. For each primary and supplemental field, clicking on a pushbutton gives the levels of fields (integer values for numeric fields; days for dates) and number of records or individuals with each level. These can be used to check that the data have been read in correctly or for other purposes (e.g. assessing the overall proportion of observations of each behavioural type).

The data can then be saved as a single "SOCPROG data file" which contains all primary and supplemental data plus any processing (see next section). This makes reusing the data particularly easy and efficient, as loading very large

Table 4 Supplemental data from an Excel worksheet with numeric individual identities

ID	Sex	Age	Haplotype
1	M	15.5	A
2	M	2.7	A
3	F	5.8	B
4	M	14.5	G
5	M	20.8	D
6	F	9.7	D
7	F	7.4	E
8	F	24.6	A
9	M	6.1	A
10	F	17.2	B
11	M	11.7	G
12	M	17.7	F
13	F	11.7	F
14	M	4.0	A
15	M	15.7	A
16	F	0.3	A
17	F	2.7	C
18	F	15.4	G
19	F	18.1	F
20	M	19.2	F

Table 5 Example matrix containing relationship measures, such as association indices, which can be input from an Excel worksheet or ASCII file and is an output of SOCPROG

	Andy	Bert	Charlie	Deb	Elen	Fran	George	Harry
Andy	1.0	0.2	0.4	0.7	0.0	0.0	0.6	0.2
Bert	0.2	1.0	0.4	0.1	0.2	0.1	0.3	0.0
Charlie	0.4	0.4	1.0	0.3	0.1	0.4	0.1	0.2
Deb	0.7	0.1	0.3	1.0	0.4	0.4	0.1	0.1
Elen	0.0	0.2	0.1	0.4	1.0	0.0	0.2	0.2
Fran	0.0	0.1	0.4	0.4	0.0	1.0	0.5	0.3
George	0.6	0.3	0.1	0.1	0.2	0.5	1.0	0.5
Harry	0.2	0.0	0.2	0.1	0.2	0.3	0.5	1.0

data sets and supplementary files can be quite computationally intensive.

Processing data

Sampling period

As a first step in the analysis, the behavioural data are divided into "sampling periods" using information from the data files (see Whitehead 2008b for advice on choosing sampling periods). For instance, if the association data are given together with date and time, sampling periods can be defined as virtually any time period, including decades, years, days or 3-h periods. Sampling periods can also be defined using input variables that are not linear measures of time, for instance "surveys", if the necessary information is coded into the primary data worksheet.

Association measures

The program next needs criteria to determine associations or interactions among pairs of individuals within each sampling period. There are several possibilities for defining the association between two individuals in a sampling period:

1. Grouped. Pairs are associated if they were observed in the same group during the sampling period. In group mode, this type of association is already defined. In linear mode, groups can be defined using input variables, such as "groupnumber" (numbered groups) or "hour" (animals sighted in each hour of the day are considered grouped).
2. Number of groups each pair was observed in together during each sampling period.
3. Number of groups each pair was observed in together during each sampling period, with groups being weighted, for instance by the duration of observation (just available in group mode).
4. Individuals are associated in a sampling period if the difference between the values of some attribute (such as the time of observation, or measure of location or combinations of these) is less than some minimum value for a pair of observations in the sampling period (available only in individual mode). For instance, two individuals could be defined as associated if they were observed at least once within 2 h of each other during the sampling period.
5. More complex associations (e.g. "nearest neighbour") can also be defined by referring to MATLAB script files.

Methods 1 and 4 give 1:0 (associated/not associated) measures of association in each sampling period, whereas methods 2 and 3 do not. Association measures are usually symmetrical (if A is associated with B, then B is associated with A), but this need not be the case (e.g. with nearest neighbours). Although SOCPROG was not originally designed with interaction data in mind, by coding the data carefully, if the data are input in a suitable format, then method 3 can be used to obtain counts of interactions between each dyad (which may be asymmetric, as with, for instance, observations of grooming) in each sampling period.

Restrictions

SOCPROG allows analyses to be carried out on only certain portions of the data set. Restrictions can be made on groups (group mode), records (linear mode), individuals or combinations of these. For instance, if appropriate variables are input, it is possible to restrict analyses to: data collected after a certain date; data only collected in certain months of the year; one gender only; certain behaviour types only; only data concerning individuals when they have passed a certain age; those animals observed more than some minimum number of times or only those individuals whose identification code contains a certain character.

Basic data

Once the sampling period and any restrictions have been set and association defined, SOCPROG can provide a summary of the data, including the number of individuals, total number of identifications and sampling periods, the number of individuals identified in each sampling period and "discovery" curves plotting the cumulative number of individuals identified against the cumulative number of identifications or the sampling period. These indicate the rates that previously unknown individuals enter the data set and thus what proportion of the population has been identified. The "Basic data" module also gives the mean number of associations per dyad (i.e. the number of sampling periods in which a dyad was associated, averaged over dyads) and per individual as well as the "social differentiation". This is an estimate of the coefficient of variation of the true association indices, the proportion of sampling periods dyads spend together, calculated by removing an estimate of the sampling variance from the coefficient of variation of the estimated association indices (Whitehead 2008b). Social differentiation is a measure of how varied the social system is, with social differentiations less than about 0.3 indicating rather homogeneous societies, greater than about 0.5 well socially differentiated populations and greater than about 2.0 a population with extreme social differentiation (i.e. generally weak relationships with a few very strong relationships; Whitehead 2008a). SOCPROG also gives an estimate of the correlation coefficient between the true association indices (the actual proportion of time pairs of individuals spend associated) and the calculated association indices (estimates of these proportions), an indicator of the power of the analysis to detect the true social system (Whitehead 2008a).

Analysing association indices

Association indices

The associations between two individuals over a number of sampling intervals are usually summarised into a single symmetric "association index" which is generally an estimate of the proportion of time that two individuals spend together. A number of indices have been proposed and used. Their relative merits are discussed by Cairns and Schwager (1987), Ginsberg and Young (1992) and Whitehead (2008b). Options available in SOCPROG are: simple ratio (preferred by Ginsberg and Young (1992) and the default when association is defined by presence in the same group); half-weight (the default when association is defined by number of groups or weighted groups that the pair was observed in); twice-weight; square root; both identified (the proportion of those sampling intervals in which both members of a dyad were identified that they were associated) and the sum of associations over all sampling periods (useful for interaction data). Additionally, the user can define an association index in a MATLAB script file.

Labels and class variable

In the output from this module, including displays of association indices (e.g. Figs. 2 and 3), individuals are labelled usually by means of the identification code in the original data file. However, it is possible to change this labelling. For instance, labels can just give gender or a combination of gender and identification code [e.g. Q405 (M)], or can be suppressed.

Many of the analyses are richer if individuals can be divided into "classes" defined by variables such as gender or genetic haplotype, and this is also possible with SOCPROG.

Listing and saving association indices

SOCPROG can output a calculated association matrix between identified individuals (e.g. Table 5) as well as estimated standard errors calculated either analytically or using the nonparametric bootstrap (Whitehead 2008a). It can save the association matrix as a SOCPROG MATLAB file (useful in the "Multiple measures analysis" module), an ASCII text file (which can be input into Excel or other programs) or a .vna file (used by network analysis programs such as UCINET).

It can also provide a summary of the association indices. For each individual, it will list the mean association index with all other individuals (the "gregariousness" as defined by Pepper et al. 1999), the maximum association index with all other individuals and an estimate of the typical group size (the sum of the association indices with all other individuals, plus one; see Jarman 1974). Also listed are the means of these measures for all individuals. If a class variable (such as gender) has been set, then these measures are presented for all combinations of classes (in the case of gender: males with males; females with females; males with females; and females with males), as are the results of a Mantel test testing for differences in the associations between, as opposed to within, classes (Schnell et al. 1985). Distributions of association indices, maximum association indices per individual, and typical group sizes, can also be plotted in a variety of ways (e.g. Fig. 2a).

Sociogram

As implemented by SOCPROG, a sociogram is a display in which individuals are represented by points arranged

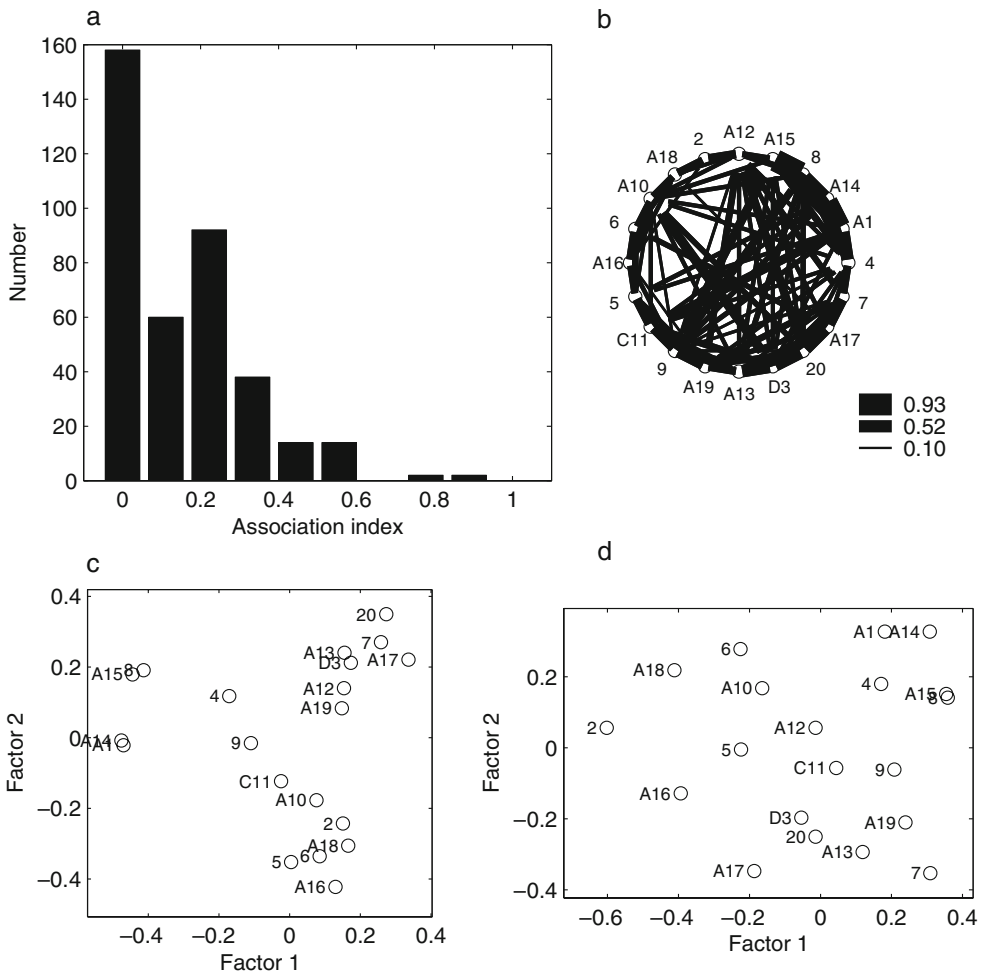

Fig. 2 Displays of an association matrix of 20 individuals from SOCPROG (example data set simgrpa.xls): **a** histogram of association indices; **b** sociogram; **c** principal coordinates analysis (first two dimensions contain 29% of variance); **d** non-metric multidimensional scaling (stress=0.21)

around the perimeter of a circle. The points are equally spaced and are linked by lines whose thickness is proportional to their association index (Fig. 2b). The minimum association at which linking lines are drawn can be set to change the level of the information/clarity trade-off. However, with more than about 25 individuals, these sociograms become cluttered. The program NetDraw (www.analytictech.com/netdraw.htm) produces much better sociograms with many more options, can use association matrices exported from SOCPROG using the .vna format and is particularly recommended when there are many individuals.

Principal coordinates analysis and non-metric multidimensional scaling

These options produce plots in which points represent individuals so that those closer together are generally more associated (e.g. Morgan et al. 1976). In principal coordinates analysis (Digby and Kempton 1987), which is sometimes called "metric scaling", the distance between two individuals is ideally inversely proportional to the square root of their association index. In non-metric multidimensional scaling, the ideal is less stringent, a monotonic arrangement in which more associated dyads are plotted more closely together than less associated ones. The number of dimensions to be used is chosen by the user. Non-metric multidimensional scaling, having less stringent ideals, usually plots satisfactorily in fewer dimensions than principal coordinates. In principal coordinates, SOCPROG gives the percent of variance accounted for, the cumulative percent of variance accounted for by a given number of dimensions and the eigenvalue and can plot any dimension against any other (Fig. 2c). For non-metric multidimensional scaling, all chosen dimensions are plotted against each other (Fig. 2d), and the stress criterion is output (with stress less than ~0.2 suggesting a useful display). Because principal coordinates is a metric display, very close associations (pairs of individuals with high association indices) are usually represented by relatively shorter distances than in non-metric scaling. The orientation of the displays is

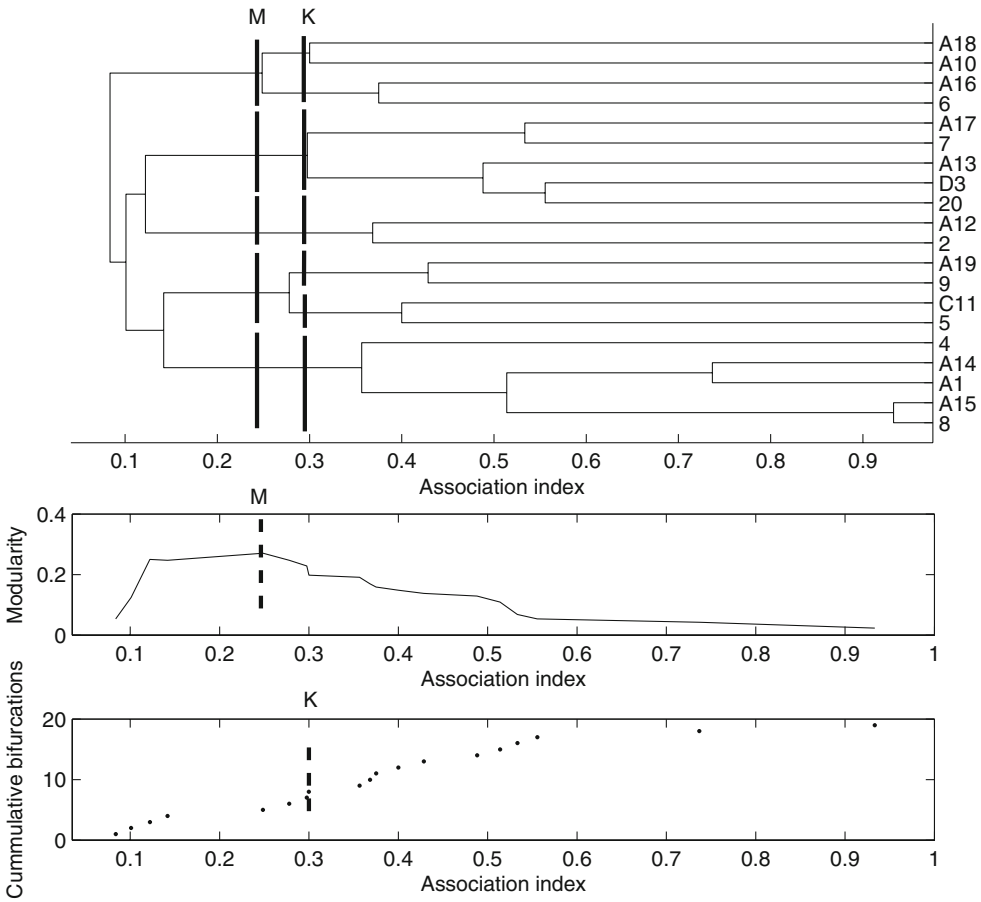

Fig. 3 Hierarchical cluster analysis of 20 individuals from SOCPROG (example data set simgrpa.xls, as in Fig. 2) from SOCPROG. Beneath the dendrogram is shown the modularity for cutoffs at different values across the dendrogram (with the maximum value, "*M*", and thus recommended cutoff value for forming groups, shown by a *dashed line*), and the knot diagram, with a "knot" (*K*), and thus potential cutoff value, indicated by the *dashed line*. The groups formed using these two cutoffs are indicated in the dendrogram above

arbitrary. For instance, Fig. 2c is quite similar to Fig. 2d with left–right and up–down rotations, although points are more clustered in the principal coordinates (Fig. 2c) ordination.

Cluster analysis

SOCPROG can make a hierarchical cluster analysis of the association data (Fig. 3). Several options are available for how clusters should be linked (average-linkage, single-linkage, complete-linkage and Ward's), but the default is average-linkage which is often optimal (Milligan and Cooper 1987). In the output dendrogram, the individuals are arranged on one axis and their degree of association on the other. The tree shows the associations among hierarchically formed clusters. However, dendrograms can be misleading: even random data can produce apparently informative dendrograms. The cophenetic correlation coefficient (calculated by SOCPROG), the correlation between the actual association indices and the levels of clustering in the diagram, indicates the effectiveness of a hierarchical cluster analysis, with values of above about 0.8 being considered to indicate an effective representation (Bridge 1993). The cophenetic correlation coefficient can also be used to select the most appropriate linkage method.

With a reasonably representative dendrogram, we may wish to define a cutoff such that clusters formed at association indices larger than the cutoff are considered "groups". SOCPROG allows this and can save the group identifiers as a supplemental variable, so that, for instance, additional analyses can be restricted to particular groups. It also provides objective means of identifying the cutoff: maximum modularity (Newman 2004) and the "knots" of a knot diagram (Wittemyer et al. 2005; Fig. 3).

Hierarchical cluster analysis is particularly suitable when a social organisation can be resolved into a structure of imbedded hierarchical levels, such as "families" within "herds" within "communities". Such a structure can be represented unambiguously by a dendrogram.

Community division using network analysis

Some societies can be usefully divided into communities such that there is little association between individuals of different communities. Hierarchical cluster analysis, as outlined above, is one way to assess the potential for such divisions and to allocate individuals among them. However, it is not necessarily the most efficient method. Network analysts have examined the problem of community division in some detail and come up with a number of algorithms.

Perhaps the most efficient is the eigenvector modularity method of Newman (2006). SOCPROG uses this technique and will output the modularity of the optimal division. Modularities greater than ~0.3 are often considered to represent useful community divisions (Newman 2004), whereas those less than ~0.3, should probably be ignored. The program also gives cluster identifiers for each individual in the optimal division, which can be saved as a supplemental variable and then used in other parts of SOCPROG.

Network analyses

The suite of techniques termed network analysis is developing rapidly and has, in the past few years, started to be applied to animal social systems (Croft et al. 2008; Krause et al. 2009; Wey et al. 2008). Whilst there are several useful specialised network analysis programs available (e.g. UCINET and NetDraw) and SOCPROG can export association matrices in .vna format which is suitable for these programs, Lusseau et al. (2008) argue that animal social network data possess attributes which mean that the more standard methods of network analysis may not be suitable. In particular, binary (1:0) representations are neither optimal nor, in many cases, appropriate, and observational error, which may be considerable, must be considered. For these reasons, SOCPROG now calculates some weighted (i.e. not 1:0) network measures from association matrices. The network measures for individuals are (Whitehead 2008b):

- strength. This is simply the sum of association indices of any individual with all other individuals and is closely related to typical group size. High strength indicates that an individual has strong associations with other individuals, many associations with other individuals or both.
- eigenvector centrality. This is given by the first eigenvector of the matrix of association indices (Newman 2004) and is a measure not only of how well an individual is associated to other individuals but also how well its close associates are themselves associated.
- reach. The reach of an individual is a measure of how well an individual is indirectly connected to others in the population and is a useful concept in a society that possesses behavioural contagion, so that the behaviour of A towards B may influence the behaviour of B towards C (Flack et al. 2006). SOCPROG calculates the reach of A as the sum, over other individuals B, of the sum of the products of all pairs of association indices linking A and B through another individual C.
- clustering coefficient. The clustering coefficient is a measure of how well the associates of an individual are themselves associated. SOCPROG uses the matrix definition of clustering coefficient for weighted networks of Holme et al. (2007).
- affinity. The affinity of an individual is the weighted average strength of its associates, weighting the strength of an associate by the association index between the individual and its associate. Therefore, the principal associates of an individual with high affinity tend to have high strength.

SOCPROG presents, for each network measure, the average and standard deviation for the whole population, for each class (if classes, such as gender, are defined), as well as, optionally, for each individual. Bootstrap SEs are available for all measures, and the network measures can be tested against null hypotheses by comparing the actual values of the network statistics with those calculated from random networks produced by the permutation test algorithms described in the next subsection.

Tests for preferred/avoided associations among individuals

In these Monte Carlo permutation tests, the null hypothesis is that individuals associate at random with other members of the population, given certain constraints. They are based upon the procedure outlined by Bejder et al. (1998), although they contain extensions and modifications to deal with issues that have come to light in recent years (Miklós and Podani 2004; Whitehead 1999; Whitehead et al. 2005). SOCPROG implements three alternative sets of constraints in these tests:

(a) the number of groups each animal was observed in during each sampling period and the number of animals in each observed group are kept constant (as in Whitehead 1999);
(b) the total number of groups each animal was observed in during the study and the number of animals in each observed group are kept constant (as in Bejder et al. 1998);
(c) the number of associations of each animal in each sampling period is kept constant (as in Whitehead 1999).

These tests are only possible if associations are defined in a 1:0 manner within each sampling period, and those of types (a) and (b) are restricted to situations in which groups are defined. Method (b) may reject the null hypothesis in cases when associations among individuals in the study area are random, but not all animals are in the study area during all sampling periods (perhaps because of birth, death or migration) or when there is autocorrelation (animals in the same group in consecutive sampling periods). It therefore is not recommended except in those situations

when the study area can be considered closed during the entire study and data from consecutive sampling periods are independent. Method (a) corrects for these problems, but may reject the null hypothesis when there are differences in gregariousness between individuals. Method (c) makes the fewest assumptions and is the most generally applicable (Whitehead 2008b).

The routines give several possible test statistics including the mean, SD and coefficient of variation (CV) of the association indices and the proportion of non-zero association indices. Usually, the significance levels of these measures are well correlated. However, they are testing different aspects of the data. For instance, if some individuals preferentially associate with other individuals, then the SD of association indices should be higher in the real data set than the random ones, but the means may be the same; and if some individuals avoid others, the proportion of non-zero association indices should be higher in the real data than in random data. The procedure has the option of finding dyads that have significantly large or small associations (as in Bejder et al. 1998). The permutation tests can also look for significantly different gregariousness among individuals, where gregariousness is an individual's tendency to form groups or associations. All these tests can be carried out within and between classes, allowing tests of null hypotheses such as: "individual males have no preferential or avoided associations with particular females".

These permutation routines produce random data sets of association indices, one of which SOCPROG saves. These random data sets can be accessed and compared with the real data when examining the distribution of association indices or constructing lagged association rates (see below).

There are a number of technical issues which need careful consideration when carrying out these permutation tests. Please see the SOCPROG manual or Whitehead (2008b) for advice.

Tests for reciprocity/unidirectionality

These tests, which were introduced by Hemelrijk (1990a, b), investigate the hypothesis that an asymmetric interaction measure is reciprocal, in other words, that individuals direct more acts to those group members from whom they receive more. If this correlation is negative (i.e. individuals direct more acts to those from whom they receive less), the behaviour is said to be "unidirectional". The analysis works by performing a Mantel test, or non-parametric variants of the Mantel test, between a matrix of interaction rates and its transpose (so that receivers become actors and vice versa). These tests can also be carried out within and between classes of individual, such as genders.

Temporal analyses

One of the most important but neglected elements of Hinde's (1976) conceptual framework of social organisation is the temporal patterning of relationships. The principal method that SOCPROG uses to examine temporal patterning in social relationships is the "lagged association rate" (Whitehead 1995). This is a function which traces changes in the association between two animals with the time lag after a sampling period in which they were associated. Given that two animals are associated now, it estimates the probability that they will also be associated τ time units in the future.

To place the lagged association rates in perspective, it can help to calculate the null association rate. This is the expected value of the lagged association rate if there is no preferred association among pairs given the number of associations of each individual in each sampling period. "No preferred association" means that the probability that A and B associate is independent of whether they have associated before. The null association rate will generally be less than or equal to the lagged association rate. When the lagged association rate equals the null association rate over a range of time lags (e.g. right-hand side of Fig. 4), this indicates no preferred associations among individuals over this time period.

Standardised lagged or null association rates should be used in situations when not all true associates of an individual are recorded during a sampling period (Whitehead 1995). Standardised rates consider effort (in the sense of the number of associates observed with each individual).

In SOCPROG, lagged association rates (and null association rates) are plotted continuously against time lag using a moving average method (Fig. 4). Lagged association rates can also be calculated for random data produced during the permutation test analyses (described above). Comparing lagged association rates for real and random data can help disentangle demographic (e.g. animals moving into and out of the study area together) and social (animals drawn to one another) causes of association.

If a class variable has been entered, then the temporal analyses can be carried out between or among the classes (e.g. lagged association rates between males and females).

It can be instructive to fit mathematical models to lagged association and standardised lagged rates (Fig. 4). SOCPROG contains a range of negative exponential models which are appropriate for fitting to fission–fusion type social structures (Whitehead 2008b), but custom models can also be fitted. SOCPROG fits the models using summed maximum likelihood and binomial loss to the full data set. Because of a lack of independence, the summed log-likelihoods resulting from fitting two different models to a data set cannot be used for formal likelihood ratio tests. However,

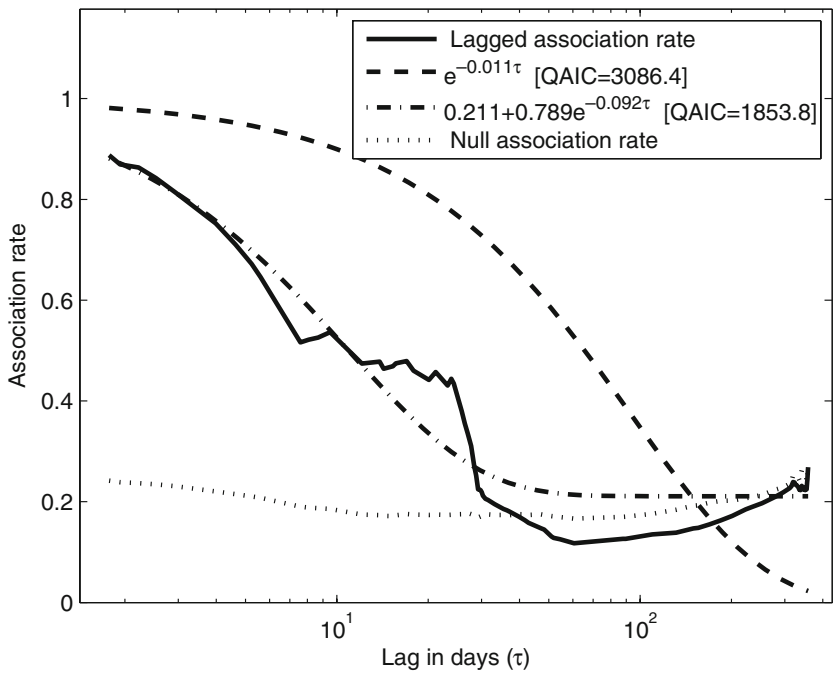

Fig. 4 Lagged association rates and null association rates from SOCPROG (example data set simgrpa.xls, as in Figs. 2 and 3) with two fitted models of exponential decay in association and QAIC. The lower QAIC indicates the model with best support

the quasi Akaike information criterion (QAIC; see Burnham and Anderson 2002) seems to be moderately effective in selecting the best model of lagged association rates (Whitehead 2007), and this is given by SOCPROG (see Fig. 4). The parameters of the models, or functions of them, can sometimes be interpreted in measures such as typical group size, size of permanent social units and rate of disassociation (Whitehead 2008b).

In order to obtain estimates of precision for lagged association rates as well as the parameters of fitted models, SOCPROG uses the jackknife procedure in which the analysis is run several times omitting one sampling period or set of sampling periods (e.g. those in a particular calendar year) each time (Efron and Stein 1981). This jackknife procedure is quite approximate.

Multiple measures analysis

This module of SOCPROG is designed for analysing two or more measures of relationship between pairs of individuals. These can either be association indices calculated by SOCPROG as described earlier, measures calculated using the supplemental data (such as age difference, or 1:0 same/different sex or haplotype) or externally input matrices (such as home range overlaps from GIS analysis or estimates of relatedness from molecular genetic analysis) in SOCPROG MATLAB, Excel or ASCII formats. The analyses in this module allow consideration of questions such as: "Do dyads have different patterns of association in different behavioural states (such as feeding or resting)?"; "Do individuals that are closely genetically related spend more time in the same group than individuals which are not so closely related?"; "Are rates at which individuals groom one another positively related to the proportion of time they spend in the same groups?"; "How can we best summarise several interaction/association measures to produce an overall relationship measure?".

The utilities in this module allow association measures to be saved (singly or as a set of measures), displayed, transformed (e.g. logged, or made into a 1:0 binary matrix based on whether the association is above some cutoff) or restricted to certain individuals based upon values of supplemental data. Any association measure, for instance one input externally, can be subjected to many of the analyses in the "Analysing association indices" module (such as cluster analyses or the production of network statistics).

However, the major goal of this module is to allow two or more association measures to be analysed together. Measures can be plotted against each other with each point representing one dyad, or several measures can be summarised using principal components analysis. SOCPROG can calculate matrix correlations that indicate the relationship between two association measures and test their significance using the Mantel test and its non-parametric variants. Also possible are partial, or semi-partial, Mantel tests in which one or both of the measures are controlled for a third (Smouse et al. 1986), so addressing questions such as "Do more closely related animals spend more time together, given their home range overlaps?"

Analysis of example data set

To illustrate some of the more fundamental functions of SOCPROG, Figs. 2, 3 and 4 use a fabricated data set provided with the package (simgrpa.xls). This data set contains 70 group mode records of a population of 20 individuals. With sampling period defined by days, there are 25 sampling periods, with an average of 60% of the individuals being identified on each day.

The histogram of simple ratio association indices (Fig. 2a) shows two very strong dyadic associations (A1 and A14, 8 and A15; Figs. 2c, d and 3) and a number of moderate associations (pairs spending 20–60% of their time together; Fig. 2a). The coefficient of variation of the association indices (CV=1.08) is significantly (P=0.0001) greater than that expected from permuting (1,000 permutations) the associations within each sampling period (CV= 0.72). Thus, it seems that within this population, there are some preferred and/or avoided relationships. The cophenetic correlation coefficient of the average-linkage cluster analysis shown in Fig. 3 is 0.79, indicating its marginal utility. Newman's (2006) eigenvector method for community division produced communities very similar, but not identical, to those suggested using cluster analysis and either the maximum modularity or knot method (Fig. 3). The modularity of the optimal arrangement is 0.29, indicating marginal utility of the community division. Thus, there is some, but rather inconclusive, evidence that this population is divided into communities.

The temporal analysis (Fig. 4) clarifies these results. Lagged association rates decline over lags from 1 day to about 1 month and then remain close to the null association rate. This indicates that preferred associations among animals within this population are not long term. An estimate of the mean duration of association is 10.9 days (the inverse of the coefficient of the exponent of the best fitting model, 0.092). The lack of long-term associations can explain the inconclusive results of the cluster analysis and community division.

Discussion

I wrote these programs to aid scientists studying animal social organisations, and in particular to aid those with large data sets on associations among identified individuals, a common situation with wild vertebrates. Among the positive features of SOCPROG are the ability to analyse data sets containing large numbers of individuals and the use of graphical user interfaces (Fig. 1) so that users can carry out many analyses of their data with little or no knowledge of the MATLAB language or when using the compiled version of the programs. With access to, and some knowledge of, MATLAB, many other types of analysis, as well as variants on the standard forms, are possible. These include using custom options for association indices; altering output such as plots; additional user-commanded analyses or plots of stored data and changing the MATLAB script or function files to carry out new analyses.

These advantages of breadth, flexibility, ease of use, diversity and extensibility carry some costs. SOCPROG is a quite complex set of programs (containing 34 script and function files, some several hundred lines long) with many options, not all combinations of which have been tested. Therefore, whenever possible, users should check that their results make sense. Another disadvantage of the programs is that little effort has been put into having the program explain errors. Thus, when the programs fail, users are usually confronted with MATLAB error messages whose meaning may be obscure.

In common with all statistical analysis packages, SOCPROG can be misused. I have seen cases in which SOCPROG results are overinterpreted (for instance constructing an elaborate social scenario based upon a cluster analysis that does not have much support). A great deal of data are required to describe and analyse almost any social system at an acceptable level of precision (Whitehead 2008a). Other misuses include the misinterpretation of lagged association rates and the results of hypothesis tests for preferred/avoided companionship. In Whitehead (2008b), I give advice on appropriate use of these and other techniques. More generally, without appropriate consideration of the analyses and their application to a particular dataset, nonsensical or misleading interpretations can result.

SOCPROG was designed for the analysis of association data. Whilst it can be used for interaction data, this is currently unwieldy. In upcoming developments, I hope to improve the ability of SOCPROG to read, process and analyse interaction, and especially asymmetric interaction, data.

Acknowledgements Thanks to Robin Baird, Jenny Christal, Susan Dufault, Shane Gero, Shannon Gowans, Andrea Ottensmeyer and, especially, David Lusseau for ideas and testing of the programs. Susan Dufault put SOCPROG on the World Wide Web. The research was funded by the Natural Sciences and Engineering Research Council of Canada. I am grateful for the constructive and detailed comments of two anonymous reviewers.

Appendix

Some technical details

The SOCPROG programs are in the programming language MATLAB (The MathWorks, Inc., 24 Prime Park Way,

Natick, Massachusetts, USA 01760-1500; www.mathworks.com). Programs were originally (1997) written in MATLAB4.2 plus the Statistics Toolbox, but the current version (SOCPROG 2.3) uses MATLAB7.4 plus the Statistics Toolbox. It should be very largely compatible with any other MATLAB7 version.

The standard version of the programs (which needs MATLAB plus the Statistics toolbox) and the compiled version (which does not need MATLAB, but possesses some limitations) can be downloaded free from the web site: http://myweb.dal.ca/hwhitehe/social.htm. There is also an online version of the manual at this site.

The programs were developed on the Windows versions of MATLAB, but they are known to work reasonably well on the UNIX, LINUX and Macintosh versions. The compiled version of SOCPROG can only be used on Windows platforms.

After downloading the uncompiled version, the user receives a .zip file containing the following: a .pdf version of the manual; SOCPROG MATLAB script and function files; a .pdf list of the MATLAB script and function files, what they do, and some important variables (useful for those with MATLAB experience, and especially those who wish to alter the programs for their own purposes); and simulated data sets which can be used to explore the programs.

The compiled version contains: a .pdf version of the manual; MCRInstaller.exe (which installs a program to run compiled MATLAB code); files with compiled SOCPROG code and simulated data sets which can be used to explore the programs. The fundamental drawback of using the compiled version of MATLAB is a loss of flexibility: The code cannot be changed to perform the exact analysis desired; further analysis of the results in MATLAB is impossible. Perhaps the biggest disadvantage for most users is that the figures and graphs produced by SOCPROG cannot be modified easily (in the uncompiled version they can be altered easily in a huge range of ways). However, the compiled figures can be exported (e.g. as .emf files) which can be edited by other programs.

Support can usually be obtained (often after a delay of some days) by emailing me at hwhitehe@dal.ca.

References

Altmann J (1974) Observational study of behavior: sampling methods. Behaviour 49:227–267

Bejder L, Fletcher D, Bräger S (1998) A method for testing association patterns of social animals. Anim Behav 56:719–725

Blumstein DT, Daniel JC (2007) Quantifying behavior the JWatcher way. Sinauer, Sunderland, MA

Borgatti SP, Everett MG, Freeman LC (1999) UCINET 6.0 Version 1.00. Analytic Technologies, Natick, MA

Bridge PD (1993) Classification. In: Fry JC (ed) Biological data analysis. Oxford University Press, Oxford, UK, pp 219–242

Burnham KP, Anderson DR (2002) Model selection and multimodel inference: a practical information-theoretic approach. Springer, New York

Cairns SJ, Schwager SJ (1987) A comparison of association indices. Anim Behav 35:1454–1469

Cheney DL, Seyfarth RM, Smuts BB, Wrangham RW (1987) The study of primate societies. In: Smuts BB, Cheney DL, Seyfarth RM, Wrangham RW, Struhsaker TT (eds) Primate societies. Chicago University Press, Chicago, pp 1–8

Croft DP, James R, Krause J (2008) Exploring animal social networks. Princeton University Press, Princeton, NJ

Digby PGN, Kempton RA (1987) Multivariate analysis of ecological communities. Chapman and Hall, London

Efron B, Stein C (1981) The jackknife estimate of variance. Ann Stat 9:586–596

Flack JC, Girvan M, de Waal FBM, Krakauer DC (2006) Policing stabilizes construction of social niches in primates. Nature 439:426–429

Ginsberg JR, Young TP (1992) Measuring association between individuals or groups in behavioural studies. Anim Behav 44:377–379

Hemelrijk CK (1990a) A matrix partial correlation test used in investigations of reciprocity and other social interaction patterns at group level. J Theor Biol 143:405–420

Hemelrijk CK (1990b) Models of, and tests for, reciprocity, unidirectionality and other social interaction patterns at a group level. Anim Behav 39:1013–1029

Hinde RA (1976) Interactions, relationships and social structure. Man 11:1–17

Holme P, Park SM, Kim BJ, Edling CR (2007) Korean university life in a network perspective: dynamics of a large affiliation network. Physica A 373:821–830

Jarman PJ (1974) The social organization of antelope in relation to their ecology. Behaviour 48:215–267

Krause J, Lusseau D, James R (2009) Introduction to animal social networks. Behav Ecol Sociobiol (in press)

Lehner PN (1998) Handbook of ethological methods. Cambridge University Press, Cambridge, UK

Lusseau D, Whitehead H, Gero S (2008) Incorporating uncertainty into the study of animal social networks. Anim Behav 75:1809–1815

Martin P, Bateson P (2007) Measuring behaviour: an introductory guide, 3rd edn. Cambridge University Press, Cambridge, UK

Miklós I, Podani J (2004) Randomization of presence–absence matrices: comments and new algorithms. Ecology 85:86–92

Milligan GW, Cooper MC (1987) Methodology review: clustering methods. Appl Psychol Meas 11:329–354

Morgan BJT, Simpson MJA, Hanby JP, Hall-Craggs J (1976) Visualizing interaction and sequential data in animal behaviour: theory and application of cluster-analysis methods. Behaviour 56:1–43

Newman MEJ (2004) Analysis of weighted networks. Phys Rev E 70:056131

Newman MEJ (2006) Modularity and community structure in networks. Proc Natl Acad Sci USA 103:8577–8582

Noldus Information Technologies (2003) MatMan, reference manual, version 1.1. Noldus, Wageningen, The Netherlands

Pepper JW, Mitani JC, Watts DP (1999) General gregariousness and specific social preferences among wild chimpanzees. Int J Primatol 20:613–632

Queller DC, Goodnight KF (1989) Estimating relatedness using genetic markers. Evolution 43:258–275

Schnell GD, Watt DJ, Douglas ME (1985) Statistical comparison of proximity matrices: applications in animal behaviour. Anim Behav 33:239–253

Smouse PE, Long JC, Sokal RR (1986) Multiple regression and correlation extensions of the Mantel test of matrix correspondence. Syst Zool 35:627–632

Strier KB (1997) Behavioral ecology and conservation biology of primates and other animals. Adv Study Behav 26:101–158

Sutherland WJ (1998) The importance of behavioural studies in conservation biology. Anim Behav 56:801–809

Visser ME (1993) The Observer, a software package for behavioural observations. Anim Behav 45:1045

Wey T, Blumstein DT, Shen W, Jordán F (2008) Social network analysis of animal behaviour: a promising tool for the study of sociality. Anim Behav 75:333–344

White GC, Burnham KP (1999) Program MARK: survival estimation from populations of marked animals. Bird Study (Suppl) 46:120–138

Whitehead H (1995) Investigating structure and temporal scale in social organizations using identified individuals. Behav Ecol 6:199–208

Whitehead H (1997) Analyzing animal social structure. Anim Behav 53:1053–1067

Whitehead H (1999) Testing association patterns of social animals. Anim Behav 57:F26–F29

Whitehead H (2001) Analysis of animal movement using opportunistic individual-identifications: application to sperm whales. Ecology 82:1417–1432

Whitehead H (2007) Selection of models of lagged identification rates and lagged association rates using AIC and QAIC. Commun Stat Simul Comp 36:1233–1246

Whitehead H (2008a) Precision and power in the analysis of social structure using associations. Anim Behav 75:1093–1099

Whitehead H (2008b) Analyzing animal societies: quantitative methods for vertebrate social analysis. Chicago University Press, Chicago, IL

Whitehead H, Dufault S (1999) Techniques for analyzing vertebrate social structure using identified individuals: review and recommendations. Adv Study Behav 28:33–74

Whitehead H, Bejder L, Ottensmeyer AC (2005) Testing association patterns: issues arising and extensions. Anim Behav 69:e1–e6

Wittemyer G, Douglas-Hamilton I, Getz WM (2005) The socioecology of elephants: analysis of the processes creating multitiered social structures. Anim Behav 69:1357–1371