2 Data management and software

2.1 Introduction

This chapter reviews some statistical programmes with which we have experience and reinforces some ideas of good data management practice.

Although there is nothing more important than looking after your raw data, our experience shows this is often poorly implemented. Unfortunately, this criticism also often applies to the subsequent collation and management of data. We are sometimes given long-term datasets to analyse, where the data have been collected and stored in a different format for each year of the monitoring programme, different surveyors have collected data from a different range of variables at different levels of detail, some data 'have been lost' and some variables are labelled with cryptic codes whose meaning no one can remember. Although we are confident that 'our' readers will have an exemplary approach to data management, we also feel it is an important enough issue to use this chapter to (briefly) reinforce some key points of data management.

Also a word of warning about the reliability of raw data: The first author used to work at an institute that collected hydrological data used for a national monitoring network. The data were collected by farmers (among others) that provided the information for a small financial award. When employees of the institute asked one of the farmers to show them the location of the measurement apparatus, he could not find them. Hence, the farmer must have made up the data.

Choice of software, both for data management and analysis, is also something worth thinking about. Many large datasets are inappropriately managed in spreadsheet programmes when they should be managed in a database programme. A very wide range of statistical programmes are available, with many offering the same range of analytical techniques, and to a large extent the choice of programme is unimportant. However, it is important to be aware that different programmes often use slightly different algorithms and different default settings. It is therefore possible to get a different answer to the same analysis performed in two different programmes. This is one reason why the make of software as well as the procedures should stated whenever the results of an analysis are presented or published.

Excellent advice on statistical good practice is given in Stern et al. (2004), and for those who insist on using spreadsheets for data management, O'Beirne (2005) is essential reading.

2.2 Data management

Data management begins at the planning stage of a project. Even simple studies need established sampling protocols and standard recording methods. The latter take the form of a printed or electronic recording sheet. This goes some way to ensuring that each surveyor records the same variables to the same level of detail. Written field records should be transferred to a 'good copy' at the end of each day, and electronic records should be backed up to a CD and/or emailed to a second secure location.

Small datasets, where the number of observations is in the hundreds, can be stored in a spreadsheet. However, storing data in a spreadsheet is less secure than using a database programme, and even for small datasets, you should consider creating a proper database. Although increasing the time required to set up the project, your data will be safer and allow a more formal and consistent approach to data entry. Hernández (2003) and Whitehorn and Marklyn (2001) give good introductions to developing a relational database. Entering data into a spreadsheet or database is an obvious source of errors, but there are some tools available to reduce these errors. Some programmes allow a double entry system (e.g., GenStat and Epidata) where each entry is typed in twice, and the programme warns you if they do not match. The use of a drop-down 'pick list' is available in both spreadsheets and databases, which will keep entries consistent. You can also constrain the range of values that a spreadsheet cell or database field will accept, removing the risk of accidentally typing in a nonsensical value. For example if you are measuring pH, you can set up a constraint where the spreadsheet cell will only accept values between 1 and 14. Typing in data from paper records to a computer is also assisted if the format of the paper record mirrors the format of the electronic entry form. And one final comment on this, although it may seem to be an effective use of staff resources to use an admin member of staff to type in data, this can also result in serious problems. Whether this is a good idea or not will depend on how well you have set up your data management system. In one instance, bad handwriting from the ecologist led to records of several non-existent invertebrates appearing after the field notes were typed into the computer by an admin assistant.

Backing up is obviously a critical part of good data management and the low cost of CD or DVD burners, or even external hard drives, now makes this relatively easy and inexpensive. In our experience, the best back-up facility is online with software that makes a daily back up of the modified files. Obviously, this only works if you have a reasonably fast Internet connection. For example, this entire manuscript (with all its data files, graphs, word documents, and script codes) was backed up daily in this way; total size was about 2 Gigabyte but the incremental back up only took 5 minutes per day (as a background process). There is even software available that will automatically back up key directories at preset intervals or every time the computer is shut down (e.g., Second Copy, http://www.secondcopy.com). It is important to check the archival properties of your chosen back-up media and design a strategy to suit. In addition to daily back-up sets, a longer term approach to backing up and storing the project's raw data should be considered. For this set you should consider storing the data in ASCII and making sure a copy is stored in more than one location, preferably in a fireproof safe. Although, for some, this may seem excessive, many datasets are the result of many years of study and irreplaceable. Even small datasets may play a critical role in future meta-analysis studies, and therefore worth looking after.

A critical aspect of backing up information is that occasionally you inspect the back-up files, and that you ensure that you know how to retrieve them.

2.3 Data preparation

Before any analysis can be done, the data have to be prepared in a database or spreadsheet programme. The data preparation is one of the most important steps in the analysis, and we mention a few key considerations below.

Data structure

Most statistics programmes assume data are in the form of a matrix with the measured values for each variable in their own column, and each row storing all the results from a single sample or trial. Even though some programmes will allow you to analyse data with a different structure, a standardised data structure can still save time and reduce risk of errors, particularly if you hand over data for someone else to help with the analysis.

Variable names

Variable names should be kept as short as possible. Some statistical programmes still require short variable names and, on importing, will truncate the name if it is longer than 10 (or 8) characters. Depending on your naming convention, this can result in several variables ending up with identical names. In graphs (particularly biplots and triplots) where you want to show the variable names on the plot, long variable names increase the chances of them overlapping and becoming unreadable. Shorter names also allow more variables to be seen and identified on one computer screen without scrolling through the columns. It is also important to check for any naming conventions required by the statistics programme you are going to use; for example, you cannot use an underscore as part of a variable name in either SPLUS or R. Developing a formal naming convention at the beginning of the project is well worth the time; for example, you can use a prefix to each variable name that indicates to which group the variable belongs. An example of this is where the same variables are measured from several different transects: 'a.phos', 'a.pot', 'a.ph' and 'b.phos', 'b.pot', 'b.ph' where the 'a' and 'b' identify the transect. A dot has been used in this instance as this is acceptable to both SPLUS and R as part of a variable name. However, it may cause problems in other software packages.

Missing and censored data

Inevitably some datasets have missing data and different programmes treat missing data differently. With SPLUS the letters NA are used to identify cells with missing data, but with other programmes, an asterisk may be used instead (GenStat) and others expect a completely blank cell (Minitab) as an indicator of missing data. If you use NA as your code for no data you will find that some statistics programmes will simply refuse to import the data, as they do not allow alphanumeric characters. Others will convert all the values for that particular variable into factors, even the numeric ones, because they recognise the NA as an alphanumeric character and assume the data refer to a nominal or categorical variable. Some programmes allow you to define the missing data code, and the use of '9', '99' or '999' is a common convention depending on the length of the variable (Newton and Rudestam 1999). If you already know the programme you will use for analysis, then the choice of code for missing data will be obvious, but whichever approach is adopted, it should be consistent and prominently documented.

Censored data are common when measuring the chemical content of an environmental variable. They arise because the concentration is below the detection limit of the technique being used to measure them. This gives results such as <0.001 ppm, and a decision needs to be made on how to deal with this type of data. Several approaches are possible, and an overview is given in Manly (2001). Different protocols seem to exist for different disciplines, and software help is available from www.vims.edu/env/research/software/vims_software.html. However, as with dealing with missing data, the key is to use a well-documented and consistent approach.

Nominal data

Nominal variables are variables of the form: yes/no; or yellow/green/blue; or transect 1, transect 2, transect3; or observer 1, observer 2, observer 3. Other examples of nominal variables are month, gender, location, etc. Some programmes can work with alphanumerical values such as 'yes' and 'no', and others cannot. For those that cannot work with alphanumeric variables, these variables have to be converted into numbers. For example, a 'yes' can be converted into a 1, and 'no' into a 0. If you do this process in a spreadsheet, you end up with an extra column containing only zeros and ones. The same principle holds if the variable has three classes; use a 1 for yellow, 2 for green and 3 for blue (or 1 for January and 12 for December). Table 2.1 shows an example for colour. The first column contains the sample number, and the second the colours. Table 2.2 shows the conversion to numerical values. Note that all data are numeric. Removing the original alphanumerical values and importing only the numerical data might still cause problems for some programmes. Table 2.3 shows how to prepare the data so that it can be used in specialised multivariate programmes that cannot convert a nominal variable with multiple classes in 0-1 dummy variables (Chapter 5).

Table 2.1. Artificial data to illustrate coding of nominal variables.

Sample	Colour	Other variables
1	Green	.....
2	Yellow	.....
3	Blue	.....
4	Blue	.....
5	Green	.....
6	Yellow	.....

Table 2.2. Artificial data to illustrate coding of nominal variables. The numbers 1, 2 and 3 represent yellow, green and blue, respectively.

Sample	Colour	Other variables
1	2	.....
2	1	.....
3	3	.....
4	3	.....
5	2	.....
6	1	.....

Table 2.3. Artificial data to illustrate coding of nominal variables.

Sample	Yellow	Green	Blue	Other variables
1	0	1	0	.....
2	1	0	1	.....
3	0	0	1	.....
4	0	0	1	.....
5	0	1	0	.....
6	1	0	0	.....

Coding the underlying question

In Chapter 28, we analyse a zoobenthic dataset measured in salt marshes in Argentina. The dataset contains four zoobenthic species and four explanatory variables measured at three transects (10 observations per transect) in two seasons. One underlying question is whether there is a transect effect and a season effect. To quantify this information, two new columns were made and labelled 'Season' and 'Transect'. Table 2.4 shows how we prepared the spreadsheet. Data of species and sediment variables are in columns. The first 30 rows are from Autumn and the next 30 rows from Spring (and there are 60 rows in total). The first 10 rows are from transect A in the Autumn, the second 10 from transect B in the Autumn, etc. The season (Autumn and Spring) was quantified using 0 and 1, and transect by 1, 2 and 3.

Table 2.4 Illustration of the preparation of a spreadsheet. The 30 observations in the three transect are denoted by A_1,..,A_{10}, B_1,..,B_{10}, C_1,..,C_{10}. We created the nominal variables Season (1 = Autumn, 2 = Spring) and Transect (1 = transect A, 2 = transect B, 3 = transect C).

Site	Season	Transect	*Laeonereis acuta*	*Heteromastus similis*	....	Mud
A_1	1	1	10	4	...	2
A_2	1	1	22	21	...	3
...	...	...	...	...	...	4
A_{10}	1	1	21	55	...	32
B_1	1	2	5	78	...	2
B_2	1	2	6	3	...	1
...	...	...	...	...	...	...
B_{10}	1	2	...	...	...	...
C_1	1	3	...	...	...	...
...	...	...	...	...	...	...
C_{10}	1	3	...	...	...	...
A_1	2	1	...	...	...	...
A_2	2	1	...	...	...	...
...	...	...	...	...	...	...
A_{10}	2	1	...	...	...	...
B_1	2	2	...	...	...	...
B_2	2	2	...	...	...	...
...	...	...	...	...	...	...
B_{10}	2	2	...	...	...	...
C_1	2	3	...	...	...	...
...	...	...	70	101	...	8
C_{10}	2	3	88	265	...	5

The underlying hypothesis for a study can sometimes be formulated using so-called dummy variables that consist of zeros and ones (or even more levels if required). For example, in Chapter 20, we analyse decapod data from two different areas and two different years. The question of whether there is an area effect can be investigated by introducing a new variable called 'Area' with values zero and one to represent the two different areas. It can then be included in the models as a main term or as an interaction term with other variables. Another example is presented in Chapter 23 in which the main underlying question is whether there is an effect of straw management on bird abundance in rice fields. To quantify the six different types of straw management, a new variable was introduced that had six different values (1 to 6), with each value representing a particular straw management regime. In Chapter 27, we use a nominal variable 'week' and 'exposure' to investigate whether there are differences in benthic communities between (i) the four weeks and (ii) beaches with different exposure types. For the data in Table 2.4, the variables Transect and Season can be used to answer the underlying questions.

2.4 Statistical software

In this section we discuss some statistics programmes with which we have experience. The absence of a programme from this chapter is simply because we have no experience using it, and is not a judgement on its quality or suitability.

Brodgar

Most of this book is based on statistics courses run by its authors. During these courses we use Brodgar (www.brodgar.com). Written by one of the books author's, Brodgar has been developed around the needs of the ecologist and environmental scientist. It is a low-cost, user-friendly programme with a graphical user interface (GUI). It allows users to apply data exploration tools, univariate, multivariate, time series and spatial methods with a few mouse clicks. About half of its methods call routines from R (www.r-project.org). This gives users a GUI alternative to the R command line interface, and with three exceptions allows all the techniques used in the book to be run from a single programme. The exceptions are (i) the geometric morphometrics analysis of the landmark turtle data in Chapter 30, (ii) neural networks in Chapter 25, and (iii) kriging in Chapters 19 and 37. In these cases, we used specialised R routines.

SPLUS and R

Another programme we use extensively is SPLUS (http://www.insightful.com). SPLUS is a commercial implementation of the S language, which provides a user-friendly menu-driven interface. An alternative is the Open Source version of the S language called R (www.project.org). Its syntax is 95% identical to SPLUS, and an excellent range of books is available to help the beginning and the experienced user of R (and also SPLUS): for example, Chambers and Hastie (1992), Venables and Ripley (2002), Dalgaard (2002), Maindonald and Braun (2003), Crawley (2002, 2005) and Verzani (2005). If we could choose only one programme to use, then we would chose R as nearly every statistical technique available has been implemented in either the base programme or in one of its many add-in libraries. The only possible problem with R is the lack of a graphical user interface, and to make the most of R, the user needs to learn R programming. Some people would argue that this is good thing as it forces the user to know exactly what he or she is doing. However, for users concerned about this programming aspect, there are menu-based add-ins available, such as Rcmdr and Biodiversity.R. Rcmdr (http://socserv.mcmaster.ca/jfox/Misc/Rcmdr/) provides a menu system for general exploratory univariate and some multivariate statistics, whereas Biodiversity-R (www.worldagroforestry.org) provides a menu interface for a range of more specialised ecologically useful routines in R.

R is excellent for analysing data and, in our experience, can also be used for teaching basic statistics to undergraduate students. Things like means, medians, *p*-values, etc. are easily calculated, and most of the methods used in this book can be learned from a third-party book such as Dalgaard (2002) or Venables and Ripley (2002) in a few days of dedicated learning.

However, in our experience it is difficult to *teach* the statistical methods discussed in this book with R to a group of 10–80 undergraduate (or postgraduate) biology students who have had no previous experience of using R. A proportion of students simply refuse to learn a programming language when they know easy-to-use graphical user interface alternatives are available.

GenStat

In Chapter 5, we discuss linear regression, and extensions to linear regression such as generalised least squares and mixed modelling are explained in subsequent chapters. For these more advanced techniques we have found GenStat (www.vsn-intl.com/genstat/) to be one of the best programmes available, and for mixed modelling and related methods, it is probably even better than R (or SPLUS). GenStat also allows the user to apply generalised linear mixed modelling, a method not covered by this book. GenStat has roots in the applied biological sciences and has a user-friendly menu-driven interface making it a good teaching tool and a good choice for users looking for a high-quality menu-driven programme. As well as the friendly front end, GenStat has a powerful programming language comparable in power with SPLUS or R, which together with its capability with large datasets also makes it a serious tool for the advanced statistician. Instructors who are considering using GenStat for classroom teaching are advised to contact GenStat as temporary free classroom licences may be available.

Other programmes

We have also used other programmes such as Minitab, SYSTAT and SPSS for undergraduate statistics courses. Although these programmes can all apply basic statistical techniques, they have a limited range of exploratory graphics tools and are less suitable for specialised statistical techniques like generalised additive modelling, redundancy analysis, dynamic factor analysis, etc. (unless special add-on libraries are bought).

Specialised multivariate programmes

For multivariate analysis, three key programmes need to be mentioned: Canoco, PRIMER and PC-ORD. Canoco for Windows version 4.5 is a software tool for constrained and unconstrained ordination. The ordination methods are integrated with regression and permutation methodology, so as to allow sound statistical modelling of ecological data. Canoco contains both linear and unimodal methods, including DCA, PCA, CCA, RDA, db-RDA, or PCoA. Ordination diagrams can be displayed and exported in publication quality right after an analysis has been completed. Canoco is unique in its capability to account for background variation specified by covariables and in its extensive facilities for permutation tests, including tests of interaction effects. Canoco has been designed for ecologists, but it is also used in toxicology, soil science, geology, or public health research, to name a few.

PRIMER is popular in marine benthic fields, and its primary assets are measures of association combined with ANOSIM, BVSTEP and other methods that carry out permutation tests (Chapters 10 and 15). Primer has also been designed with ease of use in mind and only provides the more robust analytical tools, making it a good choice for the less experienced.

PC-ORD is a Windows programme that performs multivariate analysis of ecological data entered into spreadsheets. Its emphasis is on non-parametric tools, graphical representation, and randomization tests for analysis of community data. In addition to utilities for transforming data and managing files, PC-ORD offers many ordination and classification techniques not available in other major statistical packages. Very large datasets can be analyzed. Most operations accept a matrix up to 32,000 rows or 32,000 columns and up to 536,848,900 matrix elements, provided that you have adequate memory in your machine. The terminology is tailored for ecologists.

All three programmes are popular tools for multivariate analysis, but the user will need at least one other programme for data exploration and univariate methods.

Other multivariate programmes with a focus on ecological analysis are PATN, MVSP, CAP and ECOM. The last two are from Pisces Conservation Ltd. However, the authors have limited or no experience with these programmes.

Table 2.5 contains a list of all statistical methods discussed in this book and shows which programmes can be used for each method. It should be noted that several other programmes can do the majority of these statistical methods as well (e.g., SAS), but we are not familiar with them. The time series method, dynamic factor analysis (DFA) as in Zuur et al. (2003a) is, to the best of our knowledge, only available in Brodgar.

Table 2.5. Comparison of some statistical programmes. Note that several other excellent programmes are available (e.g., SAS), but they were not taken into account as the authors have not worked with them. The symbol 'X' is used to express that a method can easily be applied in the software package. In the column Brodgar, the symbol 'R' means that an interface to R is used, 'N' means native and 'NR' indicates both native and using R. The notation 'P_{simple}' means 'requires simple script programming' and 'P_{compl}' is ' requires complicated script programming'.

	Brodgar	Genstat	CANOCO	PRIMER	R	PC-Ord
Data exploration	NR	X			X	X
Linear regression	R	X	X		X	X
Partial linear regression	R	P_{simple}	X		P_{simple}	
GLM	R	X	X		X	
GAM	R	X	X		X	
Mixed modelling	R	X			X	
GLS	R	X			X	
Tree models	R	X			X	X
Neural networks	R	P			X	
Measures of association	NR	X	X	X	X	X
PCA	N	X	X	X	X	X
RDA	N		X		X	
Partial RDA	N		X		X	
CA	N	X	X	X	X	X
CCA	N		X		X	X
Partial CCA	N		X		X	
Discriminant analysis	N	X	X		X	X
NMDS	N	X	X	X	X	X
Geometric morphometric analysis		P_{compl}			P_{compl}	
De-seasonalising	N	X			X	
Repeated lowess smoothing	R	P_{compl}			P_{compl}	
MAFA	N	P_{compl}			P_{compl}	
DFA	N	P_{compl}			P_{compl}	
Chronological clustering	N	P_{compl}			P_{compl}	
Spatial statistics						
SAR	R				X	
SMA	R				X	
Variograms	R	X			X	
Surface variogram	R	X			X	
Kriging		X			X	