Chapter

1 *Complex ecological data sets*

1.0 Numerical analysis of ecological data

The foundation of a general methodology for analysing ecological data may be derived from the relationships that exist between the conditions surrounding ecological observations and their outcomes. In the physical sciences for example, there often are cause-to-effect relationships between the natural or experimental conditions and the outcomes of observations or experiments. This is to say that, given a certain set of conditions, the outcome may be exactly predicted. Such totally deterministic relationships are only characteristic of extremely simple ecological situations.

Probability

Generally in ecology, a number of different outcomes may follow from a given set of conditions because of the large number of influencing variables, of which many are not readily available to the observer. The inherent genetic variability of biological material is an important source of ecological variability. If the observations are repeated many times under similar conditions, the relative frequencies of the possible outcomes tend to stabilize at given values, called the *probabilities* of the outcomes. Following Cramér (1946: 148), it is possible to state that "whenever we say that the probability of an event with respect to an experiment [or an observation] is equal to P, the concrete meaning of this assertion will thus simply be the following: in a long series of repetitions of the experiment [or observation], it is practically certain that the [relative] frequency of the event will be approximately equal to P." This corresponds to the frequency theory of probability — excluding the Bayesian and likelihood approaches.

Probability distribution

In the first paragraph, the outcomes were recurring at the individual level whereas in the second, results were repeatable in terms of their probabilities. When each of several possible outcomes occurs with a given characteristic probability, the set of these probabilities is called a *probability distribution*. Assuming that the numerical value of each outcome E_i is y_i with corresponding probability p_i, a *random variable* (or

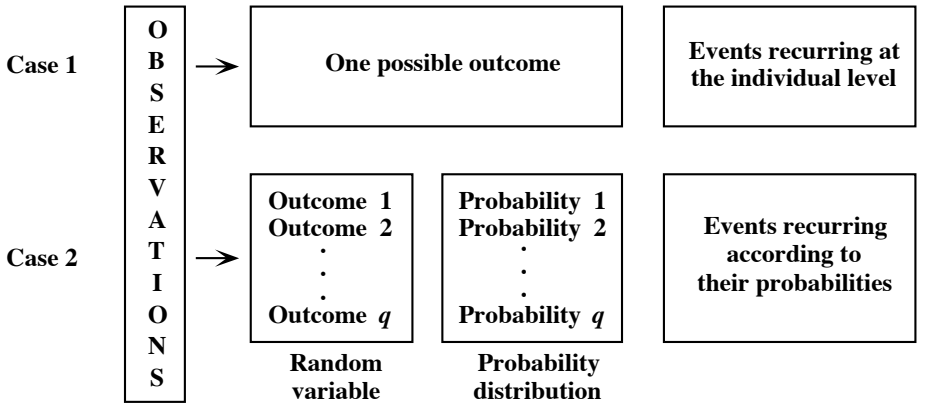


Figure 1.1 Two types of recurrence of the observations.

Random variable

variate) **y** is defined as that quantity which takes on the value y_i with probability p_i at each trial (Morrison, 1990). Figure 1.1 summarizes these basic ideas.

Of course, one can imagine other results to observations. For example, there may be *strategic* relationships between surrounding conditions and resulting events. This is the case when some action — or its expectation — triggers or modifies the reaction. Such strategic-type relationships, which are the object of *game theory*, may possibly explain ecological phenomena such as species succession or evolution (Margalef, 1968). Should this be the case, this type of relationship might become central to ecological research.

Another possible outcome is that observations bear some degree of *unpredictability*. Such data may be studied within the framework of chaos theory, which explains how deterministic processes can generate phenomena with a sensitive dependence on initial conditions that ensures dynamical behaviour with short-term predictability but long-term unpredictability (e.g. Ferriere *et al*., 1996). This is the famous "butterfly effect", whereby a butterfly flapping its wings somewhere on Earth could alter weather patterns somewhere else at a later time. The signature of chaos has been detected in a number of biological systems. For example, Beninca *et al*. (2008) used the data on a bacteria-phytoplankton-zooplankton food web that had been cultured for more than 2300 days under constant external conditions in a laboratory mesocosm to show that species interactions in that food web generated chaos. According to the authors, this result implies that the long-term prediction of species abundances could be fundamentally impossible. For an overview of chaos theory, interested readers can refer to Peitgen *et al*. (2004).

Methods of numerical analysis are determined by the four types of relationships that may be encountered between surrounding conditions and the outcome of observations (Table 1.1). The present text deals only with methods for analysing random response variables, which is the type ecologists most frequently encounter.

The numerical analysis of ecological data makes use of mathematical tools developed in many different disciplines. A formal presentation must rely on a unified approach. For ecologists, the most suitable and natural language — as will be shown in Chapter 2 — is that of *matrix algebra*. This approach is best adapted to the processing of data by computers; it is also simple, and it efficiently carries information, with the additional advantage of being familiar to many ecologists.

Other disciplines provide ecologists with powerful tools that are well adapted to the complexity of ecological data. From mathematical physics comes *dimensional analysis* (Chapter 3), which provides simple and elegant solutions to some difficult ecological problems. Measuring the association among quantitative, semiquantitative or qualitative variables is based on *parametric* and *nonparametric statistical methods* and on *information theory* (Chapters 4, 5 and 6, respectively).

These approaches all contribute to the analysis of complex ecological data sets (Fig. 1.2). Because such data usually come in the form of highly interrelated variables, the capabilities of elementary statistical methods are generally exceeded. While elementary methods are the subject of a number of excellent texts, the present manual focuses on the more advanced methods, upon which ecologists must rely in order to understand these interrelationships.

Table 1.1 Numerical analysis of ecological data.

Relationships between the natural conditions and the outcome of an observation	Methods for analysing and modelling the data
Deterministic: Only one possible result	Deterministic models
Random: Many possible results, unpredictable individually but with characteristic probabilities of occurrence	Methods described in this book (Figure 1.2)
Strategic: Results depend on the respective strategies of the organisms and of their environment	Game theory
Chaotic: Many possible results with short-term predictability and long-term unpredictability	Chaos theory
Uncertain: Many possible, unpredictable results	

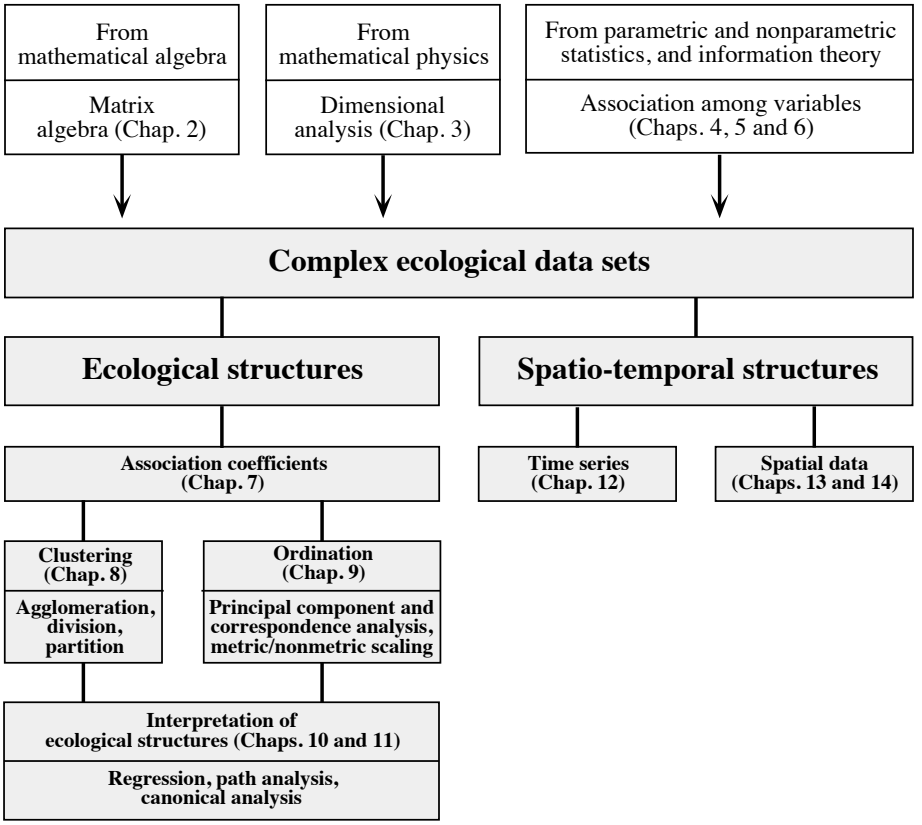


Figure 1.2 Numerical analysis of complex ecological data sets.

In ecological spreadsheets, data are typically organized in rows corresponding to sampling sites or times, and columns representing the variables; these may describe the biological communities (species presence, abundance, or biomass, for instance) or the physical environment. Because many variables are needed to describe communities and environment, ecological data matrices are, for the most part, *multidimensional* (or *multivariate*). Multidimensional data, i.e. data consisting of several variables, structure what is known in geometry as a *hyperspace*, which is a space with many dimensions. One now classical example of ecological hyperspace is the *fundamental niche* of Hutchinson (1957, 1965). According to Hutchinson, the environmental variables that are critical for a species to exist may be thought of as orthogonal axes, one for each factor, of a multidimensional space. On each axis, there are limiting conditions within which the species can exist indefinitely; this concept is

Fundamental niche

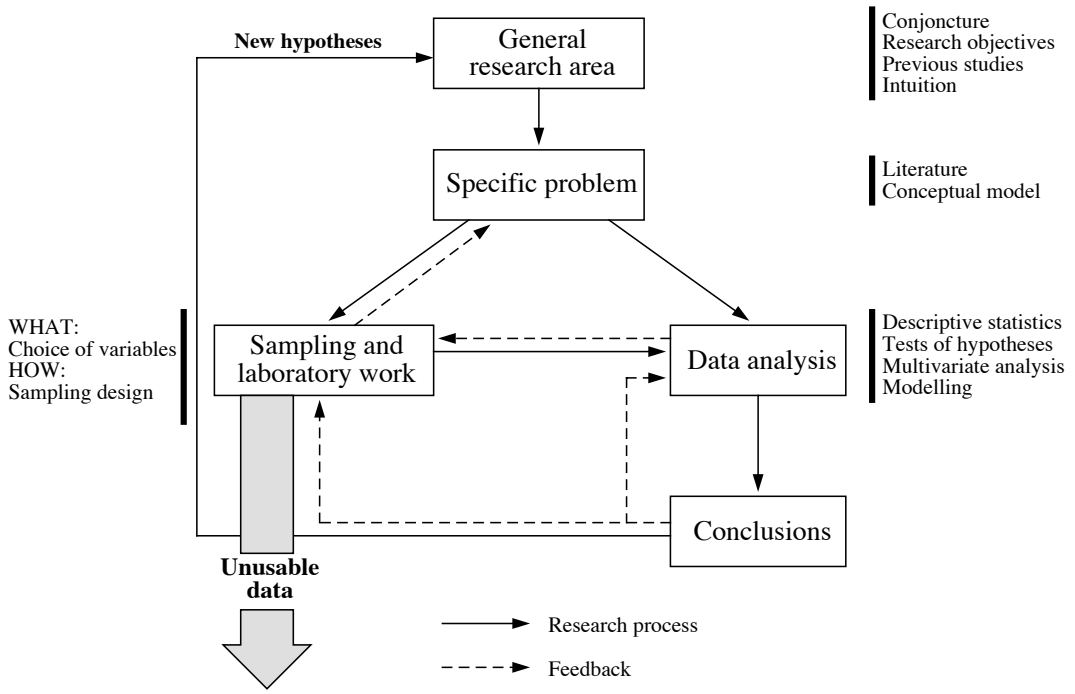


Figure 1.3 Relationships among the various phases of an ecological research.

called upon in Subsection 7.2.2, which discusses unimodal species distributions and their consequences on the choice of resemblance coefficients. In Hutchinson's theory, the set of these limiting conditions defines a hypervolume called the species' fundamental niche. The spatial axes describe the geographical distribution of the species.

The quality of the analysis and subsequent interpretation of complex ecological data sets depends, in particular, on the compatibility between data and numerical methods. It is important to take into account the requirements of the numerical techniques when planning a sampling programme, because it is obviously useless to collect quantitative data that are inappropriate to the intended numerical analyses. Experience shows that, too often, poorly planned collection of costly ecological data, for "survey" purposes, generates large amounts of unusable data (Fig. 1.3).

The search for ecological structures in multidimensional data sets is always based on *association matrices*, of which a number of variants exist, each one leading to slightly or widely different results (Chapter 7); even in so-called association-free methods, like principal component or correspondence analysis, or *K*-means partitioning, there is always an implicit resemblance measure hidden in the method.

Two main avenues are open for analysis: (1) ecological *clustering* using agglomerative, divisive or partitioning algorithms (Chapter 8), and (2) *ordination* in a space with a reduced number of dimensions, using principal component or correspondence analysis, principal coordinate analysis, or nonmetric multidimensional scaling (Chapter 9). The *interpretation of ecological structures,* derived from clustering and/or ordination, may be conducted either directly or indirectly, as will be seen in Chapters 10 and 11, depending on the nature of the problem and on the additional information available.

Ecological data may be sampled along time or space in order to study *temporal or spatial processes* driven by physics or biology (Chapters 12, 13 and 14). These data may be univariate or multivariate. Time or space sampling requires intensive field work. Time sampling can often be automated using equipment that allows the automatic recording of ecological variables. For spatial surveys, the analysis of satellite images, or of information collected by airborne or shipborne equipment, provides important support to field work, and the geographic positions of the observations can be determined using geographic positioning systems. In physical or ecological applications, a *process* is a phenomenon or a set of phenomena organized along time or through space. Mathematically speaking, such ecological data represent one of the possible realizations of a random process, also called a *stochastic process*.

Process

Two major approaches may be used for inference about the population parameters of such processes (Särndal, 1978; Koch & Gillings, 1983; de Gruijter & ter Braak, 1990; de Gruijter *et al*., 2006). In the *design-based approach*, one is interested only in the sampled population and assumes that a fixed value of the variable exists at each location in space, or point in time. A representative subset of the space or time units is selected using an appropriate (randomized) *sampling design* (for 8 different meanings of the expression "representative sampling", see Kruskal & Mosteller, 1988). Design-based (or *randomization-based*; Kempthorne, 1952) inference results from statistical analyses whose only assumption is the random selection of observations; this requires that the target population (i.e. that for which conclusions are sought) be the same as the sampled population. The probabilistic interpretation of this type of inference (e.g. confidence intervals of parameters) refers to repeated selection of observations from the same finite population using the same sampling strategy. The classical (Fisherian) methods for estimating the confidence intervals of parameters like the mean, for variables observed over a given surface or time period, are fully applicable in the design-based framework.

Design-based

Model-based

In the *model-based* (or *superpopulation*) *approach*, the assumption is that the target population is much larger than the sampled population. So, the value associated with each location, or point in time, is not fixed but random, since the geographic surface (or time period) available for sampling (i.e. the statistical population) is but one representation of the superpopulation of such surfaces or time periods — all resulting from the same generating process — about which conclusions are to be drawn. The observed population is related to the superpopulation through a *statistical model*, e.g. a variogram (Section 13.1). Under this model, even if the whole sampled population

Super-population

could be observed, uncertainty would still remain about the model parameters. So, the confidence intervals of parameters estimated over a single surface or time period are obviously too small to account for the among-surface variability, and some kind of correction must be made when estimating these intervals. The type of variability of the superpopulation of surfaces or time periods may be estimated by studying the spatial or temporal correlation of the available data (i.e. over the statistical population). This subject is discussed at some length in Section 1.1. Ecological survey data can often be analysed under either model, depending on the emphasis of the study or the type of conclusions one wishes to derive from them.

In some instances in time series analysis, the sampling design must meet the requirements of the numerical method, because some methods are restricted to data series that meet some specific conditions, such as equal spacing of observations. Inadequate planning of the sampling may render the data series useless for numerical treatment with these particular methods. There are several methods for analysing *ecological series* (Chapter 12). Regression, moving averages, and the variate difference method are designed for identifying and extracting general trends from time series. Correlogram, periodogram, and spectral analysis identify rhythms (characteristic periods) in series. Other methods can detect discontinuities in univariate or multivariate series. Variation in a series may be correlated with variation in other variables measured simultaneously. One may also develop forecasting models using the Box & Jenkins approach.

Similarly, methods are available to meet various objectives when analysing spatial data (Chapters 13 and 14). Structure functions such as variograms and correlograms, as well as point pattern analysis, may be used to confirm the presence of a statistically significant spatial structure and to describe its general features. A variety of interpolation methods are used for mapping univariate data, whereas multivariate data can be mapped using methods derived from ordination or cluster analysis. Models may also be developed that include spatial structures among their explanatory variables; in these models, spatial relationships among the study sites may be represented in a variety of ways.

For ecologists, numerical analysis of data is not a goal in itself. However, a study based on quantitative information must take data processing into account at all phases of the work, from conception to conclusion, including the planning and execution of sampling, the analysis of data proper, and the interpretation of results. Sampling, including laboratory analyses, is generally the most tedious and expensive part of ecological research, and it is therefore important that it be optimized in order to reduce to a minimum the collection of useless information. Assuming appropriate sampling and laboratory procedures, the conclusions to be drawn depend on the results of the numerical analyses. It is, therefore, important to make sure in advance that sampling and numerical techniques are compatible. It follows that numerical processing is at the heart of ecological research; the quality of the results cannot exceed the quality of the numerical analyses conducted on the data (Fig. 1.3).

Of course, the quality of ecological research is not solely a function of the expertise with which quantitative work is conducted. It depends to a large extent on creativity, which calls upon imagination and intuition to formulate hypotheses and theories (Legendre, 2004, 2008a). It is, however, advantageous for the researcher's creative abilities to be grounded into solid empirical work (i.e. work involving field data), because little progress may result from continuously building upon untested hypotheses.

Figure 1.3 shows that a correct interpretation of analyses requires that the sampling phase be planned to answer a specific question or questions. Ecological sampling programmes are designed in such a way as to capture the variation occurring along a number of axe of interest: space, time, or other ecological indicator variables. The purpose is to describe variation occurring along the given axis or axes, and to interpret or model it. Contrary to experimentation, where sampling may be designed in such a way that observations are independent of one another, ecological data are often *spatially or temporally correlated* (Section 1.1).

While experimentation is often construed as the opposite of ecological surveys, there are cases where field experiments are conducted at sampling sites, allowing one to measure rates or other processes ("manipulative experiments" *sensu* Hurlbert, 1984; Subsection 10.2.3). In aquatic ecology, for example, nutrient enrichment bioassays are a widely used approach for testing hypotheses concerning nutrient limitation of phytoplankton. In their review on the effects of enrichment, Hecky & Kilham (1988) identified four types of bioassays, according to the level of organization of the test system: cultured algae; natural algal assemblages isolated in microcosms or sometimes larger enclosures; natural water-column communities enclosed in mesocosms; whole systems. The authors discuss one major question raised by such experiments, which is whether results from lower-level systems are applicable to higher levels, and especially to natural situations. Processes estimated in experiments may be used as independent variables in empirical models accounting for survey results, while "static" survey data may be used as covariates to explain the variability observed among blocks of experimental treatments. Spatial and time-series data analysis have become an important part of the analysis of the results of ecological experiments.

1.1 Spatial structure, spatial dependence, spatial correlation

Students in elementary biostatistics courses are trained, implicitly if not explicitly, in the belief that Nature follows the assumptions of classical statistics, one of them being the independence of observations. However, field ecologists know from experience that organisms are not randomly or uniformly distributed in the natural environment, because processes such as growth, dispersal, reproduction, and mortality, which create the observed distributions of organisms, generate spatial correlation in data, as detailed below. The same applies to the physical variables that structure the environment.

Following hierarchy theory (Simon, 1962; Allen & Starr, 1982; O'Neill *et al*., 1991), we may look at the environment as primarily structured by broad-scale physical processes — orogenic and geomorphological processes on land, currents and winds in fluid environments — which, through energy inputs, create gradients in the physical environment as well as patchy structures separated by discontinuities (interfaces). These broad-scale structures lead to similar responses in biological systems, spatially and temporally. Within these relatively homogeneous zones, finer-scaled contagious biotic processes take place, causing the appearance of more spatial structuring through reproduction and death, predator-prey interactions, food availability, parasitism, and so on. This is not to say that biological processes are necessarily small-scaled and nested within physical processes; indeed, biological processes may be broad-scaled (e.g. bird and fish migrations) and physical processes may be fine-scaled (e.g. turbulence). The theory only purports that stable complex systems are often hierarchical. The concept of scale, as well as the expressions *broad scale* and *fine scale*, are discussed in Section 13.0.

In ecosystems, spatial heterogeneity is therefore functional, meaning that ecosystem functioning depends on it (Levin, 2000). It is not the result of some random, noise-generating process. So, it is important to study this type of variability for its own sake. One of the consequences is that ecosystems without spatial structuring would be unlikely to function. Let us imagine the consequences of a non-spatially-structured ecosystem: broad-scale homogeneity would cut down on diversity of habitats; feeders would not be close to their food; mates would be located at random throughout the landscape; soil conditions in the immediate surrounding of a plant would not be more suitable for its seedlings than any other location; newborn animals would be spread around instead of remaining in favourable environments; and so on. Unrealistic as this view may seem, it is a basic assumption of many of the theories and models describing the functioning of populations and communities. The view of a spatially structured ecosystem requires a new paradigm for ecologists: spatial [and temporal] structuring is a fundamental component of ecosystems (Levin, 1992; Legendre, 1993). Hence ecological theories and models, including statistical models, must be revised to include realistic assumptions about the spatial and temporal structuring of communities.

Spatial dependence, which is also called spatial correlation, is used here as the general case; temporal correlation, also called serial correlation in time series analysis, behaves essentially like its spatial counterpart but along a single sampling dimension. The difference between the spatial and temporal cases is that causality is unidirectional in time series, i.e. it proceeds from $(t - 1)$ to t and not the opposite. Temporal processes, which generate temporally correlated data, are studied in Chapter 12, whereas spatial processes are the subject of Chapters 13 and 14. The following discussion is partly inspired from the papers of Legendre & Fortin (1989), Legendre (1993), and Dray *et al*. (2012).

Spatial structures in variables may be generated by different processes. These processes produce relationships between values observed at neighbouring points in space, hence the lack of independence of values of the observed variable (Box 1.1, first

Independence Box 1.1

This word has several meanings. Five of them will be used in this book. Another important meaning in statistics concerns *independent random variables*, which refer to properties of the distributions and density functions of a group of variables; for a formal definition, see Morrison (1990, p. 7).

Independent observations. — Observations drawn from the statistical population in such a way that no observed value has any influence on any other. In the time-honoured example of tossing a coin, observing a head does not influence the probability of a head (or tail) coming out at the next toss. Spatially correlated data violate this condition because their errors are correlated across observations.

Independent descriptors. — Descriptors (variables) that are not related to one another are said to be independent. *Related* is taken here in some general sense applicable to quantitative, semiquantitative as well as qualitative data (Table 1.2).

Linear independence. — Two descriptors are said to be *linearly dependent* if one can be expressed as a linear transformation of the other, e.g. $x_1 = 3x_2$ or $x_1 = 2 - 5x_2$ (Subsection 1.5.1). Descriptors within a set are said to be *linearly dependent* if at least one of them is a linear combination of the other descriptors in the set (Section 2.7). Orthogonality (Section 2.5) is not the same as linear independence. Two vectors may be linearly independent and not orthogonal, but two orthogonal vectors are always linearly independent.

Independent variable(s) of a model. — In a regression model, the variable to be modelled is called the *dependent variable*. The variables used to model it, usually found on the right-hand side of the equation, are called the *independent variables* of the model. In empirical models, one may talk about *response* (or *target*) and *explanatory* variables for, respectively, the dependent and independent variables, whereas, in a causal framework, the terms *criterion* and *predictor* variables may be used. Some forms of canonical analysis (Chapter 11) allow the modelling of a whole matrix of dependent (target or criterion) variables in a single regression-like analysis.

Independent samples are opposed to *related* or *paired samples*. In related samples, each observation in a sample is paired with one in the other sample(s), hence the name *paired comparisons* for the tests of significance carried out on such data. Authors also talk of *independent* versus *matched* pairs of data. Before-after comparisons of the same elements also form related samples (matched pairs).

Spatial correlation

definition of independence). In many instances, observations that are closer together tend to display values that are more similar than observations that are further apart, resulting in *positive spatial dependence* also called *positive spatial correlation*. Repulsion phenomena (e.g. spatial distributions of territorial organisms that prevent other organisms from occupying neighbouring territories) may produce the opposite effect, with values of closer pairs of points being less similar than the values of pairs of observations that are further apart (*negative spatial correlation* at short distances). Closeness may be measured in a distance metric such as metres, or may be represented by counts of graph edges traversed between observations on connection networks (Subsection 13.3.1). A spatial structure may be present in data without it being caused by true autocorrelation, which is defined below. Two models for spatial structure are presented in Subsection 1.1.1; the first one (eq. 1.1 below) does not correspond to autocorrelation *sensu stricto* whereas the second does (eq. 1.2).

Because it indicates lack of independence among the observations, spatial correlation creates problems when attempting to use tests of statistical significance that assume independence of the observations. This point is developed in Subsection 1.1.2. Other types of dependencies (or, lack of independence) may be encountered in biological data. For example, related samples, discussed in more detail in Section 5.2, should not be analysed as if they were independent (Box 1.1, last definition of independence); this would result in a loss of power for the statistical test.

Spatial correlation is a very general property of ecological variables and, indeed, of most natural variables observed over geographic space (spatial correlation) or along time series (temporal correlation). Spatial [or temporal] correlation may be described by mathematical functions such as correlograms and variograms, called structure functions, which are studied in Chapters 12 and 13. The two possible approaches concerning statistical inference for spatially correlated data (i.e. the design- or randomization-based approach, and the model-based or superpopulation approach) were discussed in Section 1.0.

1 — Origin of spatial structures

A spatial structure may appear in a variable **y** because **y** depends upon one or several causal variables **X** that are spatially correlated (Model 1 below) or because the process that has produced the values of **y** is spatial and has generated correlation among the data points (Model 2 below); or some combination of these two processes. In both cases, spatial correlation will be found when analysing the data (Chapters 12 and 13). The spatially-structured causal variables **X** may be explicitly identified in the model, or not; see Table 14.1. The two models, which are also described by Fortin & Dale (2005) and Dray *et al*. (2012), are more precisely defined as follows.

Induced spatial dependence

• Model 1: induced spatial dependence — Spatial dependence may be induced by the functional dependence of the response variables (e.g. species) on explanatory variables (e.g. environmental) **X** that are themselves spatially correlated. We talk about *induced spatial dependence* in that situation where **y** has acquired the spatial structure of **X**.

This phenomenon is a restatement, in the spatial context, of the classical environmental control model (Whittaker, 1956; Bray and Curtis, 1957), which ecologists call upon when they use regression to analyse the variation of a response variable **y** by a table of environmental variables **X**. That model is the foundation of niche theory (Hutchinson, 1957). On the one hand, if all important spatially-structured explanatory variables are included in the analysis, the following model correctly accounts for the spatial structure induced in **y**:

$$y_j = f(\mathbf{X}_j) + \varepsilon_j \qquad \textbf{(1.1)}$$

where y_j is the value of the dependent variable **y** at site j and ε_j is an error term whose value is independent from site to site. On the other hand, if the function is misspecified, for example through the omission of key explanatory variables with spatial patterning such as a broad-scale linear or polynomial trend, or through inadequate functional representation, one may end up incorrectly interpreting the spatial patterning of the residuals as autocorrelation, which is described in the next paragraph.

Autocorrelation

• Model 2: spatial autocorrelation — Spatial dependence may appear in species distributions as the result of "neutral processes" of population and community dynamics (see for instance Hubbell, 2001, and Alonso *et al*., 2006). Neutral processes include ecological drift (variation in species demography due to random reproduction and survival of individuals due to competition, predator-prey interactions, etc.) and random dispersal (migration in animals, propagule dispersion in plants). These processes create *spatial autocorrelation* (*sensu stricto*) in response variables. The value y_j observed at site j on the geographic surface is assumed to be the overall mean of the process (μ_y) in the study area plus a weighted sum of the centred values $(y_i - \mu_y)$ at surrounding sites i, plus an independent error term ε_j:

$$y_j = \mu_y + \Sigma\, w_i\,(y_i - \mu_y) + \varepsilon_j \qquad \textbf{(1.2)}$$

The y_i's are the values of **y** at other sites i located within the zone of spatial influence of the process generating the autocorrelation (Fig. 1.4). The influence of neighbouring sites may be given, for instance, by weights w_i which are function of the distances between sites i and j (eq. 13.20); other functions may be used. The total error term is $[\Sigma\, w_i\,(y_i - \mu_y) + \varepsilon_j]$; it contains the autocorrelated component of variation $[\Sigma\, w_i\,(y_i - \mu_y)]$, which is noted SA_j below. The model assumes spatial stationarity (Subsection 13.1.1). Its equivalent in time series analysis is the autoregressive (AR) response model (eq. 12.29) where each observation in the time series is modelled as a function of preceding observations.

The term autocorrelation is sometimes loosely used to designate any type of spatial dependence; in that case, one would refer to spatial dependence resulting from neutral processes of population and community dynamics as "true autocorrelation", "inherent autocorrelation", or "autogenic autocorrelation" (Fortin & Dale, 2005), or as the "interaction model" (meaning: interaction among the sites) by Cliff & Ord (1981,

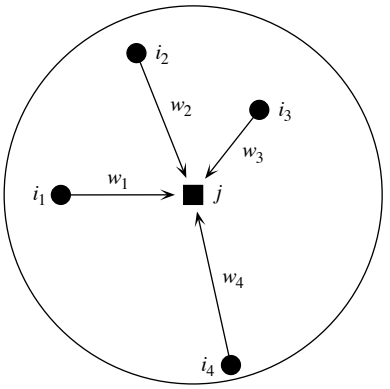


Figure 1.4 The value at site j may be modelled as a weighted sum (with weights w_i) of the influences of other sites i located within the zone of influence of the process generating the autocorrelation (large circle).

p. 141). In statistics, spatial autocorrelation is the spatial dependence found in the error component of a response variable **y** observed trough space after the effect of all important spatially-structured explanatory variables **X** has been accounted for.

The full model describing the value y_j of a response variable **y** at site j is written as follows:

$$y_j = f(\mathbf{X}_j) + u_j \qquad \text{with } u_j = \mathrm{SA}_j + \varepsilon_j$$

where **y** is modelled as a function of the explanatory (e.g. environmental) variables **X,** and u is the spatially autocorrelated residual, which has two components: the spatial autocorrelation (SA_j) in the residual and a random error component (ε_j).

For illustration, Fig. 1.5 describes the two processes that can be at the origin of a spatial structure (i.e. Model 1, induced spatial dependence, and Model 2, spatial autocorrelation) in a simplified system consisting of 4 ponds (large circles) connected by a stream; a light current is flowing from left to right. Five cases of increasing complexity are shown. In each case, circles in the upper row describe the values of an environmental variable **x** whereas the lower row concerns a response variable **y**, for example the abundances of a zooplankton species.

- Case 1 represents the null situation: there are no relationships among the values of **x** nor among those of **y** and no relationship between **x** and **y**. In a simulation program, the values of **y** corresponding to this case could be simulated as $y_j = \varepsilon_j$ where ε_j is a random normal deviate generated independently for each pond j.

- Case 2 is more interesting: it depicts functional dependence of the response variable **y** on the explanatory variable **x**. This is the classical environmental control model mentioned in the description of eq. 1.1 (Model 1). It can be implemented in simulations by equation $y_j = \beta_0 + \beta_x x_j + \varepsilon_j$ where β_0 is a constant and the functional dependence of **y** on **x** is represented by a regression parameter β_x. There is no spatial dependence (spatial correlation) among the values of **x** nor among those of **y** here.

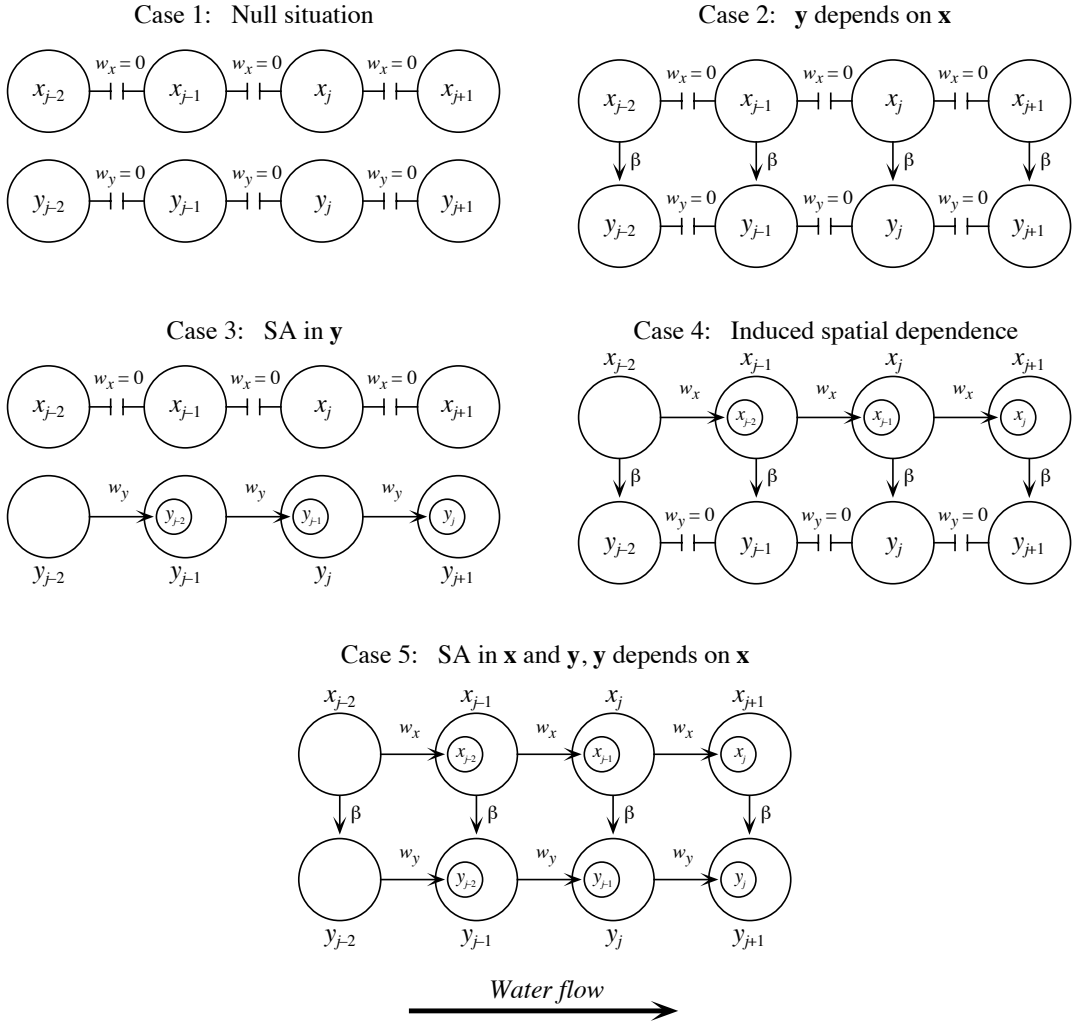


Figure 1.5 Five cases illustrating the origin of spatial structures through different types of relationships between an explanatory variable **x** and a response variable **y** observed across space. Of special interest are case 3 (spatial autocorrelation (SA) in **y**, Model 2) and case 4 (induced spatial dependence, Model 1). Modified from Fortin & Dale (2005, Chapter 5).

• Case 3 describes the process producing *spatial autocorrelation* (SA) in the response variable **y**. The arrows indicate that a random fraction of the zooplankton from pond $(j - 2)$ moves near the outflow stream and is transferred to pond $(j - 1)$ (the small circle inside the second large circle), and so on down the chain of ponds. There is no river-

like strong current moving water across the chain of ponds. As a result, zooplankton abundances in neighbouring ponds are more similar than expected in case 1. This similarity in the values of **y** due to proximity in space is called spatial autocorrelation. In numerical simulations, this process can be simulated by generating a random deviate in the first pond, $y_1 = \varepsilon_1$, and propagating it down the chain of ponds with the equation $y_j = w_y y_{j-1} + \varepsilon_j$. Equation 1.2 (Model 2) describes a similar process for sites on a 2-dimensional map with bidirectional exchanges between sites. In case 3, there is no autocorrelation in explanatory variable **x** and no functional dependence of **y** on **x**.

• Case 4 describes *induced spatial dependence*. A spatial structure is observed in **y** because that variable reflects the autocorrelated spatial structure of **x** through functional dependence of **y** on **x**. Two equations are necessary to represent this process in numerical simulations: the first describes the autocorrelation in **x** along the chain of ponds: $x_j = w_x x_{j-1} + \zeta_j$, and the second describes the spatial dependence of **y** on **x**: $y_j = \beta_0 + \beta_x x_j + \varepsilon_j$. A more general form for surfaces is eq. 1.1 (Model 1).

• Case 5 is the most complex as it combines the processes of cases 3 and 4. This is a situation often encountered in nature. There is spatial autocorrelation (SA) in **x** and in **y**, plus functional dependence of **y** on **x**. The equations describing this case in a simulation program would be: $x_j = w_x x_{j-1} + \zeta_j$ for the spatial autocorrelation (SA) in **x** and $y_j = \beta_0 + \beta_x x_j + w_y y_{j-1} + \varepsilon_j$ for the spatial dependence and autocorrelation in **y** (combination of Models 1 and 2). Methods described in Chapter 14 will show how to disentangle the two processes, using the fact that they often correspond to different spatial scales. More complex cases could be explored, e.g. the simultaneous autoregressive (AR) model and the conditional AR model (Cliff & Ord, 1981, Sections 6.2 and 6.3; Griffith, 1988, Chapter 4).

Figure 1.6 shows an example of simulated data corresponding to case 5. In the upper half of the figure, an environmental variable **x** is constructed on a map (400-point grid) as the sum of: a deterministic structure (here a unimodal distribution, upper-left map), plus spatial autocorrelation (SA) in **x**, plus random error at each point (ζ_j term in the first equation of case 5). The response variable **y** is constructed in the lower half of the figure. The effect of **x** on **y** is obtained by transporting the **x** surface (upper-right map), weighted by a regression coefficient $\beta_x = 0.3$ causing a change in the range of values in this example, to the lower-left corner where it becomes the first element in the construction of **y**. To that map, we add spatial autocorrelation (SA) in **y** and random error at each point (ε_j term in the second equation of case 5). The sum of these three surfaces produces the response variable **y** in the lower-right map. In this example, the **x** and **y** variables are sampled using a cross-shaped sampling design, represented in grey on the surface, containing 39 sampling units; any other sampling design appropriate to the study could have been used.

When there is a significant spatial structure in the data (Chapters 13 and 14), a hypothesis of induced spatial dependence (Model 1) can be examined by multiple regression (Subsection 10.3.3) or canonical analysis (Sections 11.1 and 11.2). Variation partitioning (Sections 10.3.5 and 11.1.11) and multiscale ordination (MSO,

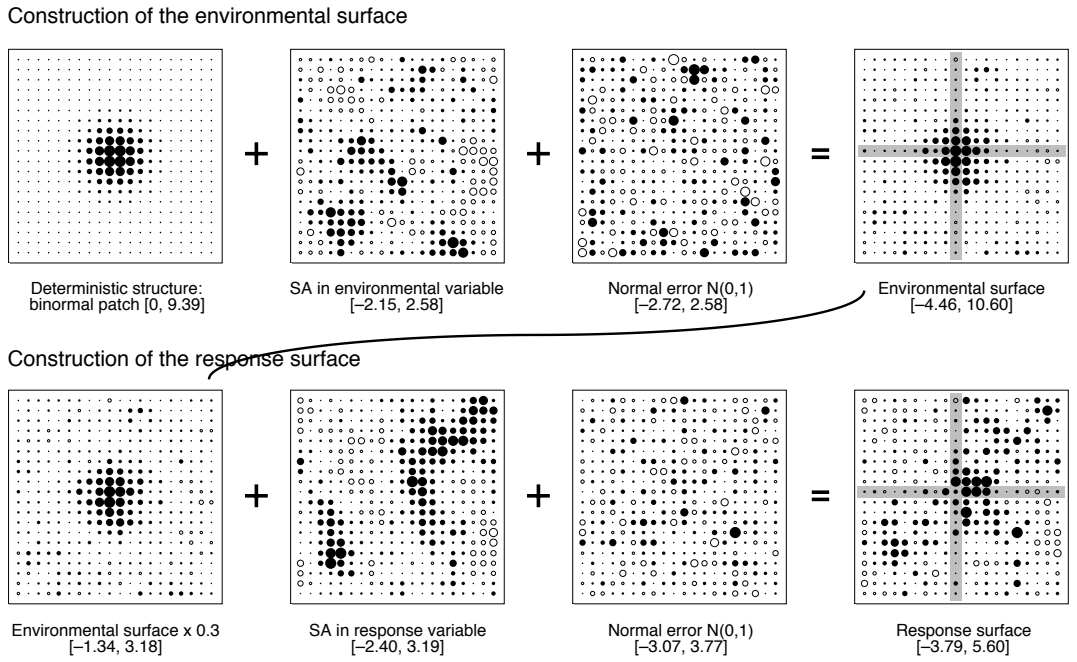


Figure 1.6 Construction of an explanatory (environmental) surface **x** and a response surface **y** in a simulation study. Each square is a bubble map of the study area. Large empty bubbles represent large negative values, and large filled bubbles, large positive values. The range of values in each map is shown in brackets underneath. The sampling design, shown in grey, is a cross with 39 sampled points in this example. Modified from Legendre *et al*. (2002, Fig. 1).

Section 14.4) can be used to determine whether or not the entire spatial structure detectable in the response data can be explained by the environmental variables (case 4) or if there remains an unexplained portion of spatial variation that would support a hypothesis of spatial autocorrelation in **y** (case 5).

Trend

A broad-scale spatial structure larger than the extent of the study area is called a *trend*. When there is a trend in the data, methods of spatial analysis detect spatial correlation due to the trend irrespective of the presence, or not, of finer-scaled sources of spatial correlation. In order to study the finer-scaled spatial structures, the trend must be removed from the data by an operation called *detrending*. One can then proceed with the analysis of the multi-scale spatial structure, for instance by spatial eigenfunction analysis (Sections 14.1 to 14.3). Linear detrending is done by regressing the response data on the geographic coordinates of the study sites (Section 13.2.1). Likewise, detrending must be done on time series before periodic or spectral analysis (Section 12.2).

Detrending

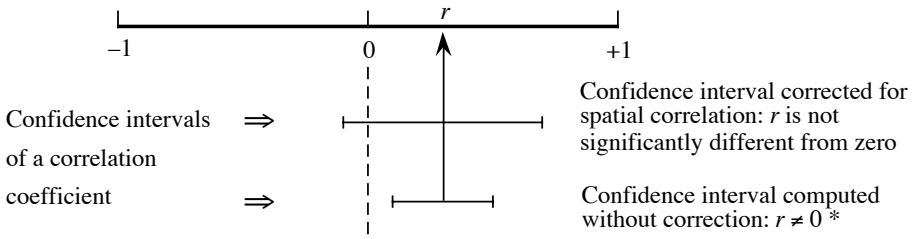


Figure 1.7 Effect of positive spatial correlation on tests of correlation coefficients; * means that the coefficient is (incorrectly) declared significantly different from zero in this example.

It is difficult to determine whether a given observed variable has been generated under Model 1 (eq. 1.1) or Model 2 (eq. 1.2). That question is further discussed in Subsection 13.1.2 in the case of gradients ("false gradients" and "true gradients") and in Chapter 14.

2 — Tests of significance in the presence of spatial correlation

Spatial correlation in a variable brings with it a statistical problem in the model-based approach (Section 1.0): it impairs the ability to perform standard statistical tests of hypotheses (Section 1.2). Let us consider an example of spatially autocorrelated data. The observed values of an ecological variable of interest — the abundances of a species for example — are most often influenced, at any given site, by the spatial distribution of the variable at surrounding sites, because of contagious biotic processes such as growth, dispersion, reproduction, and mortality. Make a first observation at site A and a second one at site B located near A. Since the ecological process is understood to some extent, one can assume that the data are spatially correlated. Using this assumption, one can anticipate to some degree the value of the variable at site B before the observation is made. Because the value at any one site is influenced by, and may be at least partly forecasted from the values observed at neighbouring sites, these values are not stochastically independent of one another.

The influence of spatial correlation on statistical tests may be illustrated using the correlation coefficient (Pearson r, Section 4.2). The problem lies in the fact that, when the two variables under study are positively spatially correlated, the confidence interval, estimated by the classical procedure around a Pearson correlation coefficient (whose calculation assumes independent and identically distributed error terms for all observations), is narrower than it is when calculated correctly, i.e. taking spatial correlation into account. The consequence is that one would declare too often that Pearson r coefficients are significantly different from zero (Fig. 1.7).

An important point is that in correlation or regression analysis, spatial correlation has a deleterious effect on tests of significance only when it is present in both variables. Simulation studies have shown that when spatial correlation was present in only one of the two variables, the test had a correct rate of type I error (Bivand, 1980; Legendre *et al.,* 2002). These simulations have also shown that deterministic spatial structures present in both variables have the same effect as spatial autocorrelation. For example, with a deterministic structure in one of the variables and spatial autocorrelation in the other, tests of significance had inflated rates of type I error.

All the usual statistical tests, nonparametric and parametric, have the same behaviour: in the presence of positive spatial correlation, the computed test statistics are too often declared significant. Negative spatial correlation may produce the opposite effect, for instance in analysis of variance (ANOVA).

The effects of spatial correlation on statistical tests may also be examined from the point of view of the *degrees of freedom*. As explained in Box 1.2, in classical statistical testing, one degree of freedom is counted for each independent observation, from which the number of estimated parameters is subtracted. The problem with spatially correlated data is their lack of independence or, in other words, the fact that new observations do not each bring with them one full degree of freedom, because the values of the variable at some sites give the observer some prior knowledge of the values the variable will take at other sites. The consequence is that new observations cannot be counted for one full degree of freedom. Since the size of the fraction they bring with them is difficult to determine, it is not easy to know what the proper reference distribution for the test should be. All that is known for certain is that positive spatial correlation at short distance distorts statistical tests (references in the next paragraph), and that this distortion is on the "liberal" side. This means that, when positive spatial correlation is present in the small distance classes, the usual statistical tests lead too often to the decision that Pearson or Spearman correlations, regression coefficients, or differences among groups are significant, when in fact they may not be.

This problem has been well documented in correlation analysis (Bivand, 1980; Cliff & Ord, 1981, §7.3.1; Clifford *et al.*, 1989; Haining, 2003, Section 8.2.1; Dutilleul, 1993a; Legendre *et al.,* 2002), linear regression (Cliff & Ord, 1981, §7.3.2; Chalmond, 1986; Griffith, 1988, Chapter 4; Haining, 1990, pp. 330-347), analysis of variance (Crowder & Hand, 1990; Legendre *et al.*, 1990, Legendre *et al.,* 2004), and tests of normality (Dutilleul & Legendre, 1992). The problem of estimating the confidence interval of the mean when the sample data are spatially correlated has been studied by Cliff & Ord (1975, 1981, §7.2) and Legendre & Dutilleul (1991).

When the presence of spatial correlation has been demonstrated, one may wish to remove the spatial dependency among observations; it would then be valid to compute the usual statistical tests. This might be done, in theory, by removing observations until spatial independence is attained; this solution is not recommended because it entails a net loss of information that was often costly to obtain. Another solution is detrending (Subsection 1.1.1) if the spatial structure is a broad-scale trend in the data; if spatial

Degrees of freedom — Box 1.2

Statistical tests of significance often call upon the concept of degrees of freedom. A formal definition is the following: "The degrees of freedom of a model for expected values of random variables is the excess of the number of variables [observations] over the number of parameters in the model" (Kotz & Johnson, 1982).

In practical terms, the number of degrees of freedom associated with a statistic is equal to the number of its independent components, i.e. the total number of components used in the calculation minus the number of parameters one had to estimate from the data before computing the statistic. For example, the number of degrees of freedom associated with a variance is the number of observations minus one (noted $\nu = n - 1$): n components $(x_i - \bar{x})$ are used in the calculation, but one degree of freedom is lost because the mean of the statistical population $(\bar{x})$ is estimated from the sample data; this is a prerequisite before estimating the variance.

There is a different t-distribution for each number of degrees of freedom. The same is true for the F and χ^2 families of distributions, for example. So, the number of degrees of freedom determines which statistical distribution, in these families (t, F, or χ^2), should be used as the reference for a given test of significance. Degrees of freedom are discussed again in Chapter 6 with respect to the analysis of contingency tables.

correlation is part of the process under study, however, this would amount to throwing out the baby with the water of the bath. It is better to analyse the spatially correlated data as such (Chapters 13 and 14), acknowledging the fact that spatial correlation in a variable may result from various causal mechanisms (physical or biological, see Subsection 1.1.1), acting simultaneously and additively.

The alternative for testing statistical significance is to modify the statistical method in order to take spatial correlation into account, as described in the following paragraphs. When such a correction is available, this approach is to be preferred if one assumes that spatial correlation is an intrinsic part of the ecological process to be analysed or modelled.

Corrected tests rely on modified estimates of the variance of the statistic, and on corrected estimates of the effective sample size and of the number of degrees of freedom. Simulation studies have been used to demonstrate the validity of the modified tests. In these studies, a large number of spatially correlated data sets are generated under the null hypothesis (e.g. for testing the difference between two means, pairs of observations are drawn at random from *the same* simulated, spatially

correlated statistical distribution, which corresponds to the null hypothesis of no difference between population means) and tested using the modified procedure; this experiment is repeated a large number of times to demonstrate that the modified testing procedure leads to the nominal rate of rejection of H_0, e.g. 0.05.

Cliff & Ord (1973) proposed a method for correcting the standard error of parameter estimates for the simple linear regression in the presence of spatial correlation. This method was extended to linear correlation, multiple regression, and *t*-test by Cliff & Ord (1981, Chapter 7: approximate solution) and to the one-way analysis of variance by Griffith (1978, 1987). Bartlett (1978) perfected a previously proposed method of correction for the effect of spatial correlation due to an autoregressive process in randomized field experiments, adjusting plot values by covariance on neighbouring plots before the analysis of variance; see also the discussion by Wilkinson *et al*. (1983) and the papers of Cullis & Gleeson (1991) and Grondona & Cressie (1991). Cook & Pocock (1983) suggested another method for correcting multiple regression parameter estimates by maximum likelihood, in the presence of spatial correlation. Using a different approach, Legendre *et al*. (1990) proposed a permutational method for the analysis of variance of spatially correlated data, in the case where the classification criterion is a division of a territory into nonoverlapping regions and one wants to test for differences among the means of these regions. Numerical simulations showed that, using this permutation method, ANOVA was insensitive to spatial correlation and effectively provided a test with a correct rate of type I error. They illustrated the method with an ecological application.

Clifford *et al*. (1989) tested the significance of the correlation coefficient between two spatial processes by estimating a modified number of degrees of freedom, using an approximation of the variance of the correlation coefficient computed from the data. Empirical results showed that their method worked fine for positive spatial correlation in large samples. Dutilleul (1993a) generalized the procedure and proposed an exact method to compute the variance of the sample covariance; the new method is valid for any sample size. In a simulation study, Legendre *et al*. (2002) showed that Dutilleul's modified *t*-test for the correlation coefficient effectively corrects for any kind of spatial correlation in the data: deterministic structures or spatial autocorrelation.

A general method to control for spatial correlation in tests of significance involving univariate or multivariate data was proposed by Peres-Neto & Legendre (2010). It involves partialling out the effect of spatial structures in partial regression (for univariate response data **y**) or partial canonical analysis (for multivariate response data **Y**). Spatial structures are represented in these analyses by spatial eigenfunctions. This method is described in Subsection 14.5.3.

Other major contributions to this topic are found in the literature on time series analysis, especially in the context of regression modelling. Important references are Cochrane & Orcutt (1949), Box & Jenkins (1976), Beach & MacKinnon (1978), Harvey & Phillips (1979), Chipman (1979), and Harvey (1981).

When methods specifically designed to handle spatial correlation are not available, it is sometimes possible to rely on permutation tests, where the significance is determined by random reassignment of the observations (Section 1.2). For some analytical situations, special permutational schemes have been developed that leave spatial correlation invariant; examples are found in Besag & Clifford (1989), Legendre *et al*. (1990) and ter Braak (1990, Section 8). The difficulty encountered in these complex problems is to design a permutation procedure that preserves the spatial or temporal correlation of the data.

The methods of clustering and ordination described in Chapters 8 and 9 to study ecological structures do not rely on tests of statistical significance. So, they are not affected by the presence of spatial correlation. The impact of spatial correlation on numerical methods will be stressed wherever appropriate.

3 — Classical sampling and spatial structure

Random or systematic sampling designs have been advocated as a way of controlling the dependence among observations (Cochran, 1977; Green, 1979; Scherrer, 1982). This was then believed to be a necessary and sufficient safeguard against violations of the independence of errors, which is a basic assumption of classical statistical tests. It is adequate, of course, when one is trying to estimate the parameters of a well-localized statistical population, for example the total number of trees in a forest plot. In such a case, a random or systematic sample is suitable to obtain unbiased estimates of the parameters since, *a priori*, each point has the same probability of being included in the sample. Of course, the variance and, consequently, also the standard error of the mean increase if the distribution is patchy, but their estimates remain unbiased.

Even with random or systematic allocation of observations through space, observations may retain some degree of spatial dependence if the average distance between first neighbours is shorter than the zone of spatial influence of the underlying ecological phenomenon. In the case of broad-scale spatial gradients, no point is far enough to lie outside this zone of spatial influence. Correlograms and variograms (Chapter 13), combined with maps, are used to assess the magnitude and shape of spatial correlation present in data sets.

Classical books such as Cochran (1977) adequately describe the rules that should govern sampling designs. Such books, however, only emphasize design-based inference (Section 1.0) and do not discuss the influence of spatial correlation on sampling designs. At the present time, most of the literature on this subject is from the field of geostatistics, where important references are: David (1977, Ch. 13), McBratney & Webster (1981), McBratney *et al*. (1981), Webster & Burgess (1984), Borgman & Quimby (1988), and François-Bongarçon (1991). In ecology, see Legendre *et al*. (2002).

Ecologists interested in designing field experiments should read the paper of Dutilleul (1993b), who discusses how to accommodate an experiment to spatially

heterogeneous conditions. Legendre *et al*. (2004) have also shown how one can effectively control for the effect of spatial correlation by the design of the experiment, and which experimental designs lead to tests of significance that have greater power.

Heterogeneity

The concept of spatial heterogeneity is discussed at some length in the multi-author book edited by Kolasa & Pickett (1991), in the review paper of Dutilleul & Legendre (1993), in the book of Dutilleul (2011), and in Section 13.0.

1.2 Statistical testing by permutation

The role of a statistical test is to decide whether some *parameter* of the reference population may take a value assumed by hypothesis, given the fact that the corresponding statistic, whose value is estimated from a sample of objects, may have a somewhat different value. A *statistic* is any quantity that may be calculated from the data and is of interest for the analysis (examples below); in tests of significance, a statistic is called *test statistic* or *test criterion*. The assumed value of the statistic, in the reference population, is given by the statistical null hypothesis (written H_0), which translates the biological null hypothesis into numerical terms; it often negates the existence of the phenomenon that the scientist is hoping to evidence. The reasoning behind statistical testing directly derives from the scientific method; it allows the confrontation of experimental or observational findings to intellectual constructs that are called hypotheses, with the explicit purpose of determining whether or not the data support the null hypothesis (see below) at some predetermined confidence level.

Statistic

Testing is the central step of inferential statistics. It allows one to generalize the conclusions of statistical estimation to the reference population from which the observations have been drawn and that they are supposed to represent. Within that context, the problem of multiple testing is too often ignored (Box 1.3). Another legitimate section of statistical analysis, called descriptive statistics, does not rely on testing. The methods of clustering and ordination described in Chapters 8 and 9, for example, are descriptive multidimensional statistical methods. The interpretation methods described in Chapters 10 and 11 may be used in either descriptive or inferential mode.

1 — Classical tests of significance

Consider, for example, a correlation coefficient (which is the statistic of interest in correlation analysis) computed between two variables (Section 4.2). When inference to the statistical population is sought, the null hypothesis is often that the value of the correlation parameter (ρ, rho) is zero in the statistical population; the null hypothesis may also be that ρ has some value other than zero, value provided by the ecological hypothesis. To judge of the validity of the null hypothesis, the only information available is an *estimate* of the correlation coefficient, *r*, obtained from a sample of objects drawn from the statistical population. (Whether the observations adequately

Null hypothesis

Multiple testing Box 1.3

When several tests of significance are carried out simultaneously, the probability of a type I error becomes larger than the nominal value α. Consider for example a correlation matrix among 5 variables: 10 tests are carried out simultaneously. For randomly generated data, there is a probability $p = 0.401$ (computed from the binomial distribution) of rejecting the null hypothesis at least once over 10 tests at the nominal $\alpha = 0.05$ level; this is called the *familywise* or *experimentwise* error rate. So, when conducting multiple tests, one should perform a global test of significance to determine whether there is any significant value at all in the set.

A general approach is the Bonferroni (1935) correction for k independent tests: replace the significance level, say $\alpha = 0.05$, by an adjusted level $\alpha' = \alpha/k$, and compare the probabilities p_i to α'. This is equivalent to adjusting individual p-values p_i to $p'_i = kp_i$ and comparing p'_i to the unadjusted significance level α. In the Sidák (1967) correction, α is replaced by an adjusted level $\alpha' = 1 - (1 - \alpha)^{1/k}$; or one can compare individual corrected values $p'_i = 1 - (1 - p_i)^k$ to the original α significance level. Although the Bonferroni and Sidák methods are appropriate to test the null hypothesis for the whole set of simultaneous hypotheses (i.e. reject H_0 for the family of k hypotheses if the smallest unadjusted p-value in the set is less than or equal to α'), these two methods are overly conservative and often lead to rejecting too few individual hypotheses in the set k, resulting in tests with low power.

Several alternatives have been proposed in the literature; see Wright (1992) for a review. For non-independent tests, Holm's procedure (1979) is nearly as simple to carry out as the Bonferroni adjustment and it is much more powerful, leading to rejecting the null hypothesis more often. It is computed as follows. (1) Order the p-values from left to right so that $p_1 \leq p_2 \leq \ldots \leq p_i \ldots \leq p_k$. (2) Compute adjusted probability values $p'_i = (k - i + 1)p_i$; adjusted probabilities may be larger than 1. (3) Proceeding from left to right, if an adjusted p-value in the ordered series is smaller than the one occurring at its left, make the smallest equal to the largest one. (4) Compare each adjusted p'_i to the unadjusted α significance level and make the statistical decision. The procedure could be formulated in terms of successive corrections to the α significance level, instead of adjustments to individual probabilities.

An even more powerful solution is that of Hochberg (1988), which has the desired overall ("experimentwise") error rate α only for independent tests, i.e. tests that do not share part of their data (Wright, 1992). This procedure is identical to Holm's except for step 3: proceeding this time from right to left, if an adjusted p-value in the series is smaller than the one at its left, make the largest equal to the smallest value. Because the adjusted p-values form a nondecreasing series, both procedures present the properties (1) that a hypothesis in the ordered series cannot be rejected unless all previous hypotheses in the series have also been rejected and (2) that equal p-values receive equal adjusted p-values. Hochberg's method has the further characteristic that no adjusted p-value can be larger than the largest unadjusted p-value or exceed 1. More complex and powerful procedures are described by Wright (1992).

Fisher's *combined probability test* allows one to combine probabilities p_i from k tests computed on independent data sets (meta-analysis). The value $-2\sum \log_e(p_i)$ is distributed as χ^2 with $2k$ degrees of freedom if H_0 is true in all k tests (Fisher, 1954; Sokal & Rohlf, 1995).

represent the statistical population is another question, for which the readers are referred to the literature on sampling design.) We know, of course, that a sample is quite unlikely to produce a parameter estimate that is exactly equal to the value of the parameter in the statistical population. A statistical test tries to answer the following question: given a hypothesis stating, for example, that $\rho = 0$ in the statistical population and the fact that the estimated correlation is, say, $r = 0.2$, is it justified to conclude that the difference between 0.2 and 0.0 is due to sampling variation?

Pivotal statistic

The choice of the statistic to be tested depends on the problem at hand. For instance, in order to find whether two samples may have been drawn from the same statistical population or from populations with equal means, one would choose a statistic measuring the difference between the two sample means ($\bar{x}_1 - \bar{x}_2$) or, preferably, a *pivotal* form like the usual *t*-statistic used in such tests; a pivotal statistic has a distribution under the null hypothesis that remains the same for any value of the measured effect (here, $\bar{x}_1 - \bar{x}_2$) because the difference of means statistic is divided by its standard error. In the same way, the slope of a regression line is described by the slope parameter of the linear regression equation, which is assumed, under the null hypothesis, to be either zero or some other value suggested by ecological theory. The test statistic describes the difference between the observed and hypothesized values of the slope; the pivotal form of this difference is a *t* or *F*-statistic.

Alternative hypothesis

Another aspect of a statistical test is the alternative hypothesis (H_1), which is also imposed by the ecological problem at hand. H_1 is the opposite of H_0, but there may be several statements that represent some opposite of H_0. In correlation analysis for instance, if one is satisfied to determine that the correlation coefficient in the reference population (ρ) is significantly different from zero in either the positive or the negative direction, meaning that *some* linear relationship exists between two variables, then a *two-tailed* alternative hypothesis is stated about the value of the parameter in the statistical population: $\rho \neq 0$. On the contrary, if the ecological phenomenon underlying the hypothesis imposes that a relationship, if present, should have a given sign, one formulates a *one-tailed* hypothesis. For instance, studies on the effects of acid rain are motivated by the general paradigm that acid rain, which lowers the pH, has a negative effect on terrestrial and aquatic ecosystems. In a study of the correlation between pH and diversity, one would formulate the following hypothesis H_1: pH and diversity are positively correlated (i.e. low pH is associated with low diversity; H_1: $\rho > 0$). Other situations would call for a different alternative hypothesis, symbolized by H_1: $\rho < 0$.

The expressions *one-tailed* and *two-tailed* refer to the fact that, in a two-tailed test, one would look in both tails of the reference statistical distribution for values as extreme as, or more extreme than the observed value of the statistic (i.e. the one computed from the actual data). In a correlation study for instance, where the reference distribution (*t*) for the test statistic is symmetric about zero, the probability of the data under the null hypothesis in a two-tailed test is given by the proportion of values in the *t*-distribution that are, *in absolute value*, as large as, or larger than the *absolute value* of the observed *t*-statistic. In a one-tailed test, one would look only in the tail corresponding to the sign given by the alternative hypothesis. For instance, for a test in

the right-hand tail (H_1: $\rho > 0$), look for the proportion of values in the *t*-distribution that are as large as or larger than the *signed value* of the observed *t*-statistic.

In standard statistical tests, the *test statistic* computed from the data is referred to one of the usual statistical distributions printed in books or computed by some appropriate computer software; the best-known are the *z*, *t*, *F* and χ^2 distributions. This, however, can only be done if certain assumptions are met by the data, depending on the test. The most commonly encountered are the assumptions of normality of the variable(s) in the reference population, normality of the regression residuals, homoscedasticity (Box 1.4), and independence of the observations (Box 1.1). Refer to Siegel (1956, Chapter 2), Siegel & Castellan (1988, Chapter 2), or Snedecor & Cochran (1967, Chapter 1), for concise yet clear classical exposés of the concepts related to statistical testing.

2 — Permutation tests

The method of *permutation*, also called *randomization*, is a very general approach to testing statistical hypotheses. Following Manly (1997), permutation and randomization are considered synonymous in the present book, although *permutation* may also be considered to be the technique by which the principle of *randomization* is applied to data during permutation tests. Other points of view are found in the literature. For instance, Edgington (1995) considers that a randomization test is a permutation test based on randomization, by opposition to restricted permutations in a loop for time series or by toroidal shift for grid data on a map. A different although related meaning of *randomization* refers to the random assignment of replicates to treatments in experimental designs.

Randomization

Permutation testing can be traced back to at least Fisher (1935, Chapter 3). Instead of comparing the actual value of a test statistic to a standard statistical distribution, the reference distribution is generated from the data themselves, as described below; other randomization methods are mentioned at the end of the present section. Permutation provides an efficient approach to testing when the data do not conform to the distributional assumptions of the statistical method one wants to use (e.g. normality). Permutation testing is applicable to very small samples, like nonparametric tests. It *does not,* however, solve problems of independence of the observations, including those caused by spatial correlation. Nor does the method solve distributional problems that are linked to the hypothesis subjected to a test*. Permutation remains the method of choice to test novel or other statistics whose distributions are poorly known.

* For instance, when studying the differences among sample means (two groups: *t*-test; several groups: *F*-test of ANOVA), the classical Behrens-Fisher problem (Robinson, 1982) reminds us that two null hypotheses are tested simultaneously by these methods, i.e. equality of the means and equality of the variances. Testing the *t* or *F*-statistics by permutations does not change the dual aspect of the null hypothesis; in particular, it does not allow one to unambiguously test the equality of the means without checking first the equality of the variances using another, more specific test (two groups: *F* ratio; several groups: Bartlett's test of equality of variances).

Furthermore, results of permutation tests are valid even with observations that are not a random sample of some statistical population; this point is further discussed in Subsection 1.2.4. Edgington (1995) and Manly (1997) have written excellent introductory books about the method. A short account is given by Sokal & Rohlf (1995) who use the expression "randomization test". Permutation tests are used in several chapters of the present book.

The speed of modern computers would allow users to perform any statistical test using the permutation method. The chief advantage is that one does not have to worry about the distributional assumptions of classical testing procedures; the disadvantage is the extra computer time required to actually perform a large number of permutations, each one being followed by recomputation of the test statistic. Permutation tests are fairly easy to program and are increasingly available in computer packages. As an example, let us consider the situation where the significance of a correlation coefficient between two variables, $\mathbf{x}_1$ and $\mathbf{x}_2$, is to be tested.

Hypotheses

- H_0: The correlation between the variables in the reference population is zero ($\rho = 0$).
- For a two-tailed test, H_1: $\rho \neq 0$.
- Or for a one-tailed test, either H_1: $\rho > 0$, or H_1: $\rho < 0$, depending on the ecological hypothesis.

Test statistic

- Compute the Pearson correlation coefficient r. Calculate the pivotal statistic $t = \sqrt{n-2}\,[r/\sqrt{1-r^2}]$ (eq. 4.13; n is the number of observations) and use it as the observed value of the test statistic in the remainder of the test.

In this specific case, the permutation test results would be the same using either r or t as the test statistic, because t is a monotonic function of r for any constant value of n; r and t are "equivalent statistics for permutation tests", *sensu* Edgington (1995). This is not always the case. For example, when testing a partial regression coefficient in multiple regression, the test should not be based on the distribution of permuted partial regression coefficients because they are not monotonic to the corresponding partial t-statistics. The partial t should be preferred because it is pivotal and, hence, it is expected to produce correct type I error.

Considering a pair of equivalent test statistics, one could choose the statistic which is the simplest to compute if calculation time would otherwise be longer in an appreciable way. This is not the case in the present example: calculating t involves a single extra line in the computer program compared to r. So the test is conducted using the usual t-statistic.

Distribution of the test statistic

The argument invoked to construct a null distribution for the statistic is that, if the null hypothesis is true, all possible pairings of the two variables are equally likely to occur. The pairing found in the observed data is just one of the possible, equally likely pairings, so that the value of the test statistic for the unpermuted data should be typical, i.e. located in the central part of the permutation distribution.

- It is always the null hypothesis that is subjected to testing. Under H_0, the rows of $\mathbf{x}_1$ are exchangeable with one another if the rows of $\mathbf{x}_2$ are fixed, or conversely, and the observed pairing of $\mathbf{x}_1$ and $\mathbf{x}_2$ values is due to chance alone; accordingly, any value of $\mathbf{x}_1$ could have been paired with any value of $\mathbf{x}_2$.

- A realization of H_0 is obtained by permuting at random the values of $\mathbf{x}_1$ while holding the values of $\mathbf{x}_2$ fixed, or the opposite (which would produce, likewise, a random pairing of values). Recompute the value of the correlation coefficient and the associated *t*-statistic for the randomly paired vectors $\mathbf{x}_1$ and $\mathbf{x}_2$, obtaining a value t^*.

- Repeat this operation a large number of times (say, 999 or 9999 times). The different permutations produce a set of values t^* obtained under H_0.

- Add to these the observed value of the *t*-statistic, computed for the unpermuted vectors. Since H_0 is being tested, this value is considered to be one of those that could be obtained under H_0 and, consequently, it should be added to the distribution of *t* values (Hope, 1968; Edgington, 1995; Manly, 1997). Together, the unpermuted and permuted values form an estimate of the sampling distribution of *t* under H_0, which will be used as the reference distribution in the next step.

Statistical decision

- As in any other statistical test, the decision is made by comparing the observed value of the test statistic (*t*) to the reference distribution obtained under H_0. If the observed value of *t* is typical of the values obtained under the null hypothesis (which states that there is no relationship between $\mathbf{x}_1$ and $\mathbf{x}_2$), H_0 cannot be rejected; if it is unusual, being too extreme to be considered a likely result under H_0, H_0 is rejected and the alternative hypothesis is considered to be a more likely explanation of the data.

- Compute the associated p-value, which is the proportion of values in the reference distribution that are as extreme as, or more extreme than the observed value of the test statistic. The p-value is either computed from the reference distribution obtained by permutations, or found in a table of the appropriate statistical distribution. The p-value is a statement about the probability of obtaining a result as extreme as, or more extreme than the one actually obtained from the sample data, assuming that H_0 is true for the reference population. Researchers often write in short that it is *the probability of the data under the null hypothesis*. Fisher (1954) saw the p-value as a measure of the strength of evidence against the null hypothesis; the smaller the p-value, the stronger the evidence against H_0.

Significance level

• Compare the p-value to a predetermined significance level α. Following the Neyman-Pearson (or *frequentist*) approach (Neyman & Pearson, 1966), one rejects H_0 if $p \leq \alpha$, and does not reject it if $p > \alpha$. Or one can use the Fisher approach: Fisher left the interpretation of the p-value and the ensuing statistical decision to the researcher.

3 — Numerical example

Let us consider the following case of two variables observed over 10 objects:

$\mathbf{x}_1$	–2.31	1.06	0.76	1.38	–0.26	1.29	–1.31	0.41	–0.67	–0.58
$\mathbf{x}_2$	–1.08	1.03	0.90	0.24	–0.24	0.76	–0.57	–0.05	–1.28	1.04

These values were drawn at random from a positively correlated bivariate normal distribution, as shown in Fig. 1.8a. Consequently, they would be suitable for parametric testing. So, it is interesting to compare the results of a permutation test to the usual parametric *t*-test of the correlation coefficient. The statistics and associated probabilities for this pair of variables, for $\nu = (n - 2) = 8$ degrees of freedom, are:

$r = 0.70156$, $t = 2.78456$, $n = 10$:
prob (one-tailed) = 0.0119, prob (two-tailed) = 0.0238.

There are $10! = 3.6288 \times 10^6$ possible permutations of the 10 values of variable $\mathbf{x}_1$ (or $\mathbf{x}_2$). Here, 999 of these permutations were generated using a random permutation algorithm; they represent a random sample of the 3.6288×10^6 possible permutations. The computed values for the test statistic (t) between permuted $\mathbf{x}_1$ and fixed $\mathbf{x}_2$ have the distribution shown in Fig. 1.8b; the observed value, $t = 2.78456$, has been added to this distribution. The permutation results are summarized in the following table, where $|t|$ is the (absolute) observed value of the *t*-statistic ($|t| = 2.78456$) and t^* is a value obtained after permutation. The absolute value of the observed t is used in the following table to make it a general example since there are cases where t is negative.

	$t^* < -\lvert t \rvert$	$t^* = -\lvert t \rvert$	$-\lvert t \rvert < t^* < \lvert t \rvert$	$t^* = \lvert t \rvert$	$t^* > \lvert t \rvert$
Statistic t	8	0	974	1[†]	17

[†] This count corresponds to the observed t value that was added to the reference distribution.

For a one-tailed test (in the right-hand tail in this case, since H_1: $\rho > 0$), one counts how many values in the permutational distribution of the statistic are equal to, or larger than, the observed value ($t^* \geq t$; there are $1 + 17 = 18$ such values in this case). This is the only one-tailed hypothesis worth considering, because the objects are known in this case to have been drawn from a positively correlated distribution. A one-tailed test in the left-hand tail (H_1: $\rho < 0$) would be based on how many values in the permutational distribution are equal to, or smaller than, the observed value ($t^* \leq t$, which are $8 + 0 + 974 + 1 = 983$ in the example). For a two-tailed test, one counts all values that are as extreme as, or more extreme than the observed value *in both tails of the distribution* ($|t^*| \geq |t|$, which are $8 + 0 + 1 + 17 = 26$ in the example).

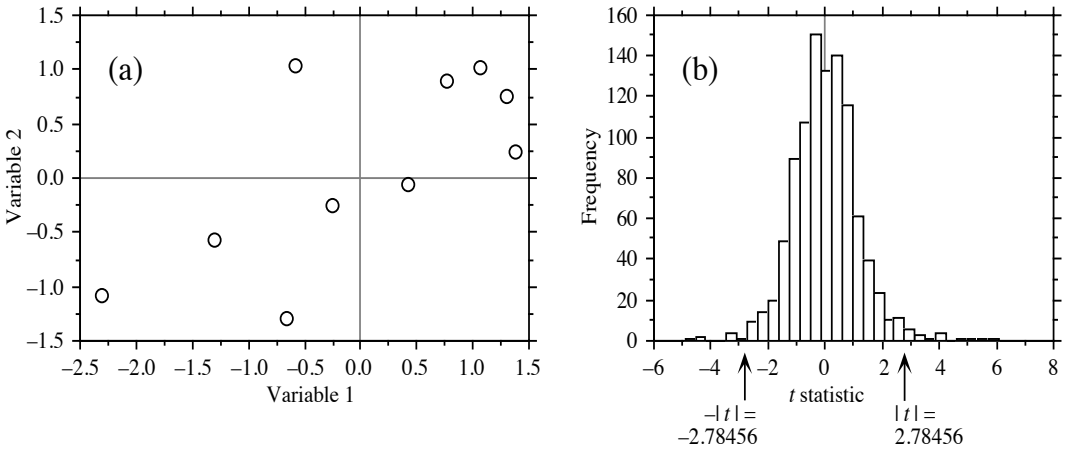


Figure 1.8 (a) Scatter diagram of the 10 points of the numerical example with respect to variables $\mathbf{x}_1$ and $\mathbf{x}_2$. (b) Frequency histogram of the (1 + 999) permutation results (t-statistics for correlation coefficients). The observed value of t, $|t| = 2.78456$, is shown, as well as $-|t| = -2.78456$.

Probabilities associated with these distributions are computed as follows, for a one-tailed and a two-tailed test (results using the r statistic would be the same):

One-tailed test [H_0: $\rho = 0$; H_1: $\rho > 0$]:
prob ($t^* \geq 2.78456$) = (1 + 17)/1000 = 0.018

Two-tailed test [H_0: $\rho = 0$; H_1: $\rho \neq 0$]:
prob($|t^*| \geq 2.78456$) = (8 + 0 + 1 + 17)/1000 = 0.026

Note how similar the permutation results are to the results obtained from the classical test, which referred to a table of the Student t-distribution. The observed difference is partly due to the small number of pairs of points ($n = 10$) sampled at random from the bivariate normal distribution, with the consequence that the data set does not quite conform to the hypothesis of normality. It is also due, to a certain extent, to the use of only 999 permutations, sampled at random among the 10! possible permutations.

4 — Remarks on permutation tests

In permutation tests, the reference distribution against which the statistic is tested is obtained by randomly permuting the data under study, without reference to any statistical population. The test is valid as long as the reference distribution has been generated by a procedure related to a null hypothesis that makes sense for the problem at hand, irrespective of whether or not the data set is representative of a larger statistical population. This is the reason why the data do not have to be a random

sample from some larger statistical population. The only information the permutation test provides is whether the pattern observed in the data is likely, or not, to have arisen by chance. For this reason, one may think that permutation tests are not as "good" or "interesting" as classical tests of significance because they might not allow one to infer conclusions that apply to a statistical population.

A more pragmatic view is that the conclusions of permutation tests may be generalized to a reference population if the data set is a random sample of that population. Otherwise, they allow one to draw conclusions only about the particular data set, measuring to what extent the value of the statistic is "usual" or "unusual" with respect to the null hypothesis implemented in the permutation procedure. Edgington (1995) and Manly (1997) further argue that data sets are very often not drawn at random from statistical populations, but simply consist of observations that happen to be available for the study. The generalization of results, in classical as well as permutation tests, depends on the degree to which the data were actually drawn at random, or are equivalent to a sample drawn at random, from a reference population.

Complete permutation test

For small data sets, one can compute all possible permutations in a systematic way and obtain the complete permutation distribution of the statistic; an *exact* or *complete permutation test* is obtained. For large data sets, only a sample of all possible permutations may be computed because there are too many. When designing a *sampled permutation test*, it is important to make sure that one is using a *uniform random generation algorithm*, capable of producing all possible permutations with equal probabilities (Furnas, 1984). Computer programs use procedures that produce random permutations of the data; these in turn call the 'Random' function of computer languages. Such a procedure is described in Section 5.8 of Manly's book (1997). Random permutation functions are available in subroutine libraries and in R.

Sampled permutation test

The case of the correlation coefficient has shown how the null hypothesis guided the choice of an appropriate permutation procedure, capable of generating realizations of this null hypothesis. A permutation test for the difference between the means of two groups would involve random permutations of the objects between the two groups instead of random permutations of one variable with respect to the other. The way of permuting the data depends on the null hypothesis to be tested.

Some tests may be reformulated in terms of some other tests. For example, the *t*-test of equality of means is equivalent to a test of the correlation between the vector of observed values and a vector assigning the observations to group 1 or 2. The same value of *t* and probability (classical or permutational) are obtained using both methods.

Simple statistical tests such as those of correlation coefficients or differences between group means may be carried out by permuting the original data, as in the example above. Problems involving complex relationships among variables may require permuting the residuals of some *model* instead of the raw data; *model-based permutation* is discussed in Subsection 11.1.8. The effect of a nominal covariable may be controlled for by *restricted permutations*, limited to the objects within the groups

Restricted permutations

defined by the covariable. This method is discussed in detail by Manly (1997). Applications are found in Brown & Maritz (1982; restrictions within replicated values in a multiple regression) and in Sokal *et al*. (1987; Mantel test), for instance.

In sampled permutation tests, adding the observed value of the statistic to the distribution has the effect that it becomes impossible for the test to produce no value "as extreme as, or more extreme than the observed value", as the standard expression goes. This way of computing the probability is biased, but it has the merit of being statistically valid (Edgington, 1995, Section 3.5). The precision of the probability estimate is the inverse of the number of permutations performed; for instance, after (999 + 1) permutations, the precision of the probability statement is 0.001.

How many permutations?

The number of permutations one should perform is always a trade-off between precision and computer time. The more permutations the better, since probability estimates are subject to error due to sampling the population of possible permutations (except in the rare cases of complete permutation tests), but it may be tiresome to wait for the permutation results when studying large data sets. Jackson & Somers (1989) recommend to compute 10000 to 100000 permutations in order to ensure the stability of the probability estimates in Mantel tests (Subsection 10.5.1). The following recommendation can be made. In exploratory analyses, 500 to 1000 permutations may be sufficient as a first contact with the problem. If the computed probability is close to the preselected significance level, run more permutations. In any case, use more permutations (e.g. 10000) for final results submitted for publication.

Interestingly, tables of critical values in nonparametric statistical tests for small samples are based on permutations. The authors of these tables computed how many cases can be found, in the complete permutation distribution, that are as extreme as, or more extreme than the computed value of the statistic. Hence, probability statements obtained from small-sample nonparametric tests are exact probabilities (Siegel, 1956).

Monte Carlo

Named after the city that hosts the famous casino in the principality of Monaco, Monte Carlo methods use random numbers to study either real data sets or the behaviour of statistical methods through simulations. Permutation tests are Monte Carlo methods because they use random numbers to randomly permute data. Other such methods are based on computer-intensive resampling. Among these are the jackknife (Tukey, 1958; Sokal & Rohlf, 1995) and the bootstrap (Efron, 1979; Efron & Tibshirani, 1993; Manly, 1997). In the latter methods, the values used in each iteration to compute a statistic are a subsample of the original data. In the jackknife, each subsample leaves out one of the original observations and sampling is done *without replacement*. In the bootstrap, each subsample is obtained by resampling the original sample *with replacement*; the justification is that resampling the original sample approximates a resampling of the original population.

Jackknife
Bootstrap

As an exercise, readers are invited to figure out how to perform a permutation test for the difference between the means of two groups of objects on which a single variable has been measured, using the *t*-statistic, and create a permutational *t*-test R

function*. Other types of permutation tests are discussed in Sections 5.4, 7.3, 8.9, 10.2, 10.3, 10.5, 10.6, 11.1, 11.4, 11.5, 12.6, 13.1 and 13.3.

1.3 Computer programs and packages

Processing complex ecological data sets almost always requires the use of computers, as much for the amount of data to be processed as for the fact that the operations to be performed are complex and often repetitious.

Powerful statistical packages such as SAS®, SPSS®, Statistica® and others are commercially available for general statistical analysis. Many other programs are either commercially or freely available on the Web pages of researchers or research institutions; some of these programs will be mentioned in *Software* sections in the following chapters.

This book will pay special attention to statistical functions available in the R language, which was developed in 1990 by Ross Ihaca and Robert Gentleman at the University of Auckland. R is a dialect of the S language. The S freeware was created in 1976 by John Chambers and colleagues at *AT&T Bell Laboratories*. R became freeware in 1995 and an international project in 1997. Its source code is freely available under the GNU General Public License. For most users, R is a powerful environment to carry out statistical analyses. R is also a programming language that allows scientists to easily write new functions. For computationally-intensive tasks, R functions can call compiled code written in C, C++ and Fortran.

The main features of the R language are described on the Web page http://en.wikipedia.org/wiki/R_(programming_language). Other computer languages such as S-PLUS® (a commercial implementation of S) and MATLAB® offer features comparable to R; however, they are not free.

The use of R has grown tremendously among researchers during the past 15 years and it has become a *de facto* standard for software development and computing in most fields of science. The fact that it is free and multi-platform explains in part its success: functions can be used in the same way on all major personal computer operating systems (presently Microsoft Windows, Mac OS X, and Linux). R is also available for a wide variety of Unix platforms. The other part of the explanation holds in the fact that the *R Development Core Team* has encouraged contributions from the community of users and methods developers, who have joined in the movement wholeheartedly. As a result, thousands of R packages are now available on the *Comprehensive R Archive Network* (CRAN) main site (http://cran.r-project.org/) and on mirror sites.

* Readers can compare their solution to the R function ***t.perm()*** available on the Web page http://numericalecology.com/rcode.

Thousands more packages and individual functions are distributed by researchers on their Web pages or are attached to scientific papers describing new numerical methods. All functions found in R packages come with documentation files, called by the ***help()*** function or by a question mark, and they are all presented in the same format.

There are many reference books published about the R language and its application to various fields. A good starting point to learn about R is *The R book* of Crawley (2007). The Venables & Ripley (2002) textbook is the acknowledge reference for many functions found in the R and S languages. In several of the following chapters, we will refer to the book *Numerical ecology with R* by Borcard *et al*. (2011), which was written as a companion to the 1998 and the present editions of *Numerical ecology*. The Borcard *et al*. (2011) book is of particular interest to readers who wish to implement the methods described in this book using available R software.

Here is an example of how R packages and functions will be referred to in this book: package VEGAN, function ***rda()***. The parentheses after function names contain data file names and other parameters necessary to run functions.

Ecologists should bear in mind that easy computation has two pitfalls: the fact that computations are done and results are produced does not ensure (1) that the data satisfy the conditions required by the method, or (2) that the results produced by the computer are interpreted correctly in ecological terms. This book provides colleagues with the theoretical and practical information they need to avoid these pitfalls.

1.4 Ecological descriptors

Descriptor
Variable

Any ecological study, classical or numerical, is based on *descriptors*. In the present text, the terms *descriptor* and *variable* will be used interchangeably. These refer to the attributes, or characters (also called items in the social sciences, and profiles or features in the field of pattern recognition) used to describe or compare the *objects of the study*.

Object

The *objects* that ecologists compare are the sites, quadrats, observations, sampling units, individual organisms, or subjects; these are defined *a priori* by the sampling design, before making the observations (Section 2.1). Observation units are often called "samples" by ecologists. The term *sample* is only used in its statistical sense in this book; it refers to *a set of observations* resulting from a sampling action or campaign. Objects may be called individuals or OTUs (*Operational taxonomic units*) in numerical taxonomy, OGUs (*Operational geographic units*) in biogeography, cases, patterns or items in the field of pattern recognition, etc.

The descriptors, used to describe or qualify the objects, are the physical, chemical, ecological, or biological characteristics of these objects that are of interest for the study. In particular, biological species are *descriptors* of sites for ecologists; in (numerical) taxonomy on the contrary, the species are the *objects* of the study, and the sites where the species are observed or collected may be used by the taxonomist as

descriptors of the species. It all depends on the variable, defined *a priori,* that specifies the objects of a study. In ecology, sites are compared using the species they contain, there being no possibility of choosing the species, whereas taxonomists compare populations or other taxonomic entities obtained from a number of different sites.

Descriptor

A *descriptor is a law of correspondence established by the researcher* to describe and compare, on the same basis, all the objects of the study. This definition applies to all types of descriptors discussed below (Table 1.2). The fundamental property of a descriptor, as understood in the present book, is that it distributes the objects among non-overlapping states. Each descriptor must, therefore, operate like *a law that associates with each object in the group under study one and only one element of a set of distinguishable states that belong to the descriptor.*

Descriptor state

The *states* that constitute a descriptor *must necessarily be mutually exclusive*. In other words, two different states of the same descriptor must not be applicable to the same object. Descriptors, on the contrary, do not have to be independent of one another (see Box 1.1: independent descriptors). In Chapter 6, it will be seen that the information contained in one descriptor may partially or totally overlap with the information in another descriptor.

1 — Mathematical types of descriptors

The states that form a descriptor — i.e. the qualities observed or determined on the objects — may be of a qualitative or quantitative nature, so that descriptors may be classified into several types. In ecology, a descriptor may be biological (presence, abundance, or biomass of different species), physical, chemical, geological, geographical, temporal, climatic, etc. Table 1.2 presents a classification of descriptors according to their mathematical types. That classification is independent of the particular discipline to which the descriptors belong. The mathematical type of a descriptor determines the type of numerical processing that can be applied to it. For example, parametric correlations (Pearson's r) may be calculated between quantitative descriptors, while nonparametric correlations (such as Kendall's τ) may be used on ordered but not necessarily quantitative descriptors, as long as their relationship is monotonic. To measure the dependence among descriptors that are not in monotonic relationship, or among qualitative descriptors, requires the use of other methods based on contingency tables (Chapter 6). Subsection 1.5.7 will show how descriptors of different mathematical types can be made compatible, in order to use them together in ecological studies.

Quantitative descriptors, which are the most usual type in ecology, are found at the bottom of Table 1.2. They include all descriptors of abundance and other quantities that can be plotted on a continuous axis of real numbers. They are called quantitative, or *metric* (Falconer, 1960), because they measure changes in a phenomenon in such a way that the difference between 1 and 2, for example, is quantitatively the same as the difference between, say, 6 and 7. Such descriptors may be further subdivided into *relative-scale* quantitative variables, where value 'zero' means the absence of the characteristic of interest, and *interval-scale* variables where the 'zero' is chosen arbitrarily. For the latter type, the fact that the 'zero' reference is chosen

Relative scale
Interval scale

Table 1.2 The different mathematical types of descriptors.

Descriptor types	Ecological examples
Binary (two states, presence-absence)	Species present or absent
Multi-state (many states)	
Nonordered (qualitative, nominal, attributes)	Geological group
Ordered	
Semiquantitative (rank-ordered, ordinal)	Importance or abundance scores
Quantitative (metric, measurement)	
Discontinuous (meristic, discrete)	Equidistant abundance classes
Continuous (metric)	Temperature, length

arbitrarily prevents comparisons of the type "this temperature (°C) is twice as high as that one". Species abundance data, or temperatures measured in Kelvin, are examples of the first type, while temperature measured in degrees Celsius, dates, or geographic directions (of wind, currents, etc.) in degrees, are examples of the second.

Continuous quantitative descriptors are usually processed as they are. If they are divided into a small number of *equidistant* classes of abundance (further discussed below), the discontinuous descriptors that are obtained may usually be processed as if they were continuous, because the distortion due to grouping is negligible for the majority of distribution types (Sneath & Sokal, 1973). Before the advent of computers, it was usual practice, in order to facilitate calculations, to divide continuous descriptors into a small number of classes. This transformation is still necessary when, due to low precision of the measurements, only a small number of classes can be distinguished in practice, or when comparisons are sought between quantitative and semiquantitative descriptors.

Meristic variables (the result of enumeration, or counting) theoretically should be considered as discontinuous quantitative. In ecology, however, these descriptors are most often counts of the number of specimens belonging to the various species, whose range of variation is so large that they behave, for all practical purposes, as continuous variables. When they are transformed (Sections 1.5 and 7.7), as is often the case, they become real numbers instead of integers.

In order to speed up field observations or counts in the laboratory, it is often interesting for ecologists to record observations in the form of *semiquantitative* descriptors. Usually, it is possible to estimate environmental characteristics very rapidly by ascribing them a score using a small number of ordered classes: score 1 < score 2 < score 3, etc. Ecologists may often proceed

in this way without losing pertinent information, whereas precise counts would have necessitated more considerable efforts than required by the ecological phenomenon under study. For example, in studying the influence of the unevenness of the landscape on the fauna of a given area, it may be enough to describe the relief using ordered classes such as flat, undulated, rough, hilly and mountainous. In the same way, counting large numbers of organisms may be done using abundance scores instead of precise numbers of individuals. Frontier (1973), for example, established such a scoring scale to describe the variability of zooplankton. Another score scale, also developed by Frontier (1969) for counting zooplankton, was used to estimate biomass (Dévaux & Millerioux, 1976b) and diversity of phytoplankton (Dévaux & Millerioux, 1977) as well as to evaluate schools of cetaceans at sea (Frontier & Viale, 1977). Frontier & Ibanez (1974) as well as Dévaux & Millerioux (1976a) have shown that this rapid technique is as informative as classical enumeration for principal component analysis (Section 9.1). It must be noted that nonparametric statistical tests of significance, which are used on such semiquantitative descriptors, have a discriminatory power almost equal to that of their parametric equivalent. Naturally occurring semiquantitative descriptors, which give *ranks* to the objects under study, as well as quantitative descriptors divided into non-equidistant classes (which is done either to facilitate data collection or to evidence holes in frequency distributions), are included among the semiquantitative descriptors. Method 6.4 in Subsection 1.5.6 shows how to normalize semiquantitative descriptors if they have to be used in methods of data analysis that perform better in the presence of normality. Normalized semiquantitative descriptors should only be interpreted in terms of the ordinal value that they really represent. In addition, methods designed for quantitative data analysis may often be adapted to ranked data. This is the case, for example, with principal component analysis (Lebart *et al*., 1979; Subsection 9.1.7) and linear regression (Iman & Conover, 1979).

Qualitative descriptors often present a problem to ecologists, who are tempted to discard them, or reduce them to a series of binary variables (Subsection 1.5.7). Let us forget the cases where descriptors of this kind have been camouflaged as ordered variables by scientists who did not quite know what to do with them …Various methods based on contingency tables (Chapter 6) may be used to compare such descriptors with one another, or to ordered descriptors divided into classes. Special resemblance coefficients (Chapter 7) allow these descriptors to be used as a basis for clustering (Chapter 8) or ordination (Chapter 9). The first paragraph of Chapter 6 gives examples of qualitative descriptors. An important class is formed by classifications of objects, which may in turn become descriptors of these objects for subsequent analyses, since the definition of a classification (Section 8.1) corresponds to the definition of a descriptor given above.

Binary or *presence-absence* descriptors may be noted + or –, or 1 or 0. In ecology, the most frequently used type of binary descriptors is the presence or absence of a species, when reliable quantitative information is not available. It is only for historical reasons that they are considered as a special class: programming the first computers was greatly facilitated by such descriptors and, as a result, several methods have been developed for processing them. Sneath & Sokal (1973) present various methods to recode variables into binary form; see also Subsection 1.5.7. Binary descriptors encountered in ecology may be processed either as qualitative, semiquantitative or quantitative variables. Even though the mean and variance parameters of binary descriptors are difficult to interpret, such descriptors may be used with methods originally designed for quantitative variables — in a principal component or correspondence analysis, for instance, or as independent variables in regression or canonical analysis models.

When collecting ecological data, the level of precision with which descriptors are recorded should be selected with consideration of the problem at hand. Quantitative descriptors may often be recorded either in their original form or in semiquantitative or qualitative form. The degree of precision should be chosen with respect to the following factors: (1) What is the optimal degree of precision of the descriptor for analysing this particular ecological phenomenon? (2) What type of mathematical treatment will be used? This choice may determine the mathematical types of the descriptors. (3) What additional cost in effort, time or money is required to raise the level of precision? Would it not be more informative to obtain a larger number of less precise data?

2 — Intensive, extensive, additive, and non-additive descriptors

There are other useful ways of looking at variables. Margalef (1974) classified ecological variables as either *intensive* or *extensive*. These notions are derived from

Intensive

thermodynamics (Glansdorff & Prigogine, 1971). A variable is said to be *intensive* if its value is defined independently of the size of the sampling unit in which it is measured. For example, water temperature is defined independently of the size of the bucket of water in which a thermometer is placed: we do not say "12°C per litre" but simply "12°C". This does not mean that the *measured value* of temperature may not vary from place to place in the bucket; it may indeed, unless water is well-mixed and therefore homogeneous. Concentration of organisms (number per unit surface or volume), productivity, and other rate variables (e.g. birth, death) are also intensive because, in a homogeneous system, the same value is obtained whether the original

Extensive

measurements are made over 1 m^2 or over 100 m^2. In contrast, an *extensive* variable is one whose value, in a homogeneous system, changes proportionally (in linear relationship) to the size of the sampling unit (which may consist in a line, a surface, or a volume). It is formally defined as an integral over the sampling unit. Number of individuals and biomass in a quadrat or volume, at a given point in time, are examples of extensive variables.

Extensive variables have the property that the values they take in two sampling units can be added to provide a meaningful estimate of the value in the combined unit: they are additive (next paragraph). Other variables do not have this property; either they do not vary at all (e.g. temperature in a homogeneous bucket of water, which is an intensive variable), or they vary in a nonlinear way with the size of the sampling unit. For example, species richness in a sampling unit (surface or volume) cannot be computed as the sum of the numbers of species found in two sub-units; that sum would usually be larger than the number of species actually found in the combined unit because some species are common to the two sub-units. Species diversity (Chapter 5) also has this property. The relationship of such variables to scale is complex and depends on the distribution patterns of the species and the size of the sampling units (grain size of the measurements; Section 13.0).

Additive

Another, more statistical point of view concerns additivity. This notion is well-known in geostatistics (Olea, 1991, p. 2; Journel & Huijbregths, 1978). A variable is

said to be *additive* if its values can be added while retaining the same meaning as the original variable. A good example is the number of individuals in a quadrat. Concentrations, which are intensive variables, are additive if they are referred to the same linear, surface or volume unit measure (e.g. individuals m^{-2}; kg m^{-3}) (Journel & Huijbregths, 1978, p. 199); values may be added to compute a mean for example.

Extensive variables (e.g. number of individuals) are, by definition, additive; a sum or a mean has the same meaning as the original data although, if the sampling units differ in size, the values must be weighted by the sizes of the respective sampling units for their mean to be meaningful. For intensive additive variables (e.g. temperature or concentration), only the (weighted) mean has the same meaning as the original values. Variables may be additive over either time or space (Walliser, 1977); numbers of individuals in quadrats, for example, are additive over space, but not over time if the time lag between observations is shorter than the generation time of the organisms (the same individuals would be counted several times).

Non-additive

Examples of *non-additive variables* are pH values, logarithms and ratios of random variables, indices of various kinds, and directions of vectors (wind direction, aspect of a slope, etc.). Values of non-additive variables must be transformed in some way before (and if) they can be meaningfully combined. Logarithms of counts of organisms, for instance, have to be back-transformed using antilogarithms before values can be added. For ratios, the numerator and denominator must be added separately, and the ratio recomputed from these sums. Other non-additive variables, such as species richness and diversity, cannot be numerically combined; these indices for combined sampling units must be recomputed from the combined raw data.

These notions are of prime importance when analysing spatial data (Chapters 13 and 14). To appreciate their practical usefulness, let us consider a study in which the following variables have been measured at a site in a lake or in the ocean, at different times: incident solar energy at water surface (W m^{-2}), temperature (°C), pH, O_2 concentration (g m^{-3}), phytoplankton production (g C m^{-3} s^{-1}), and concentration of zooplankton (individuals m^{-3}). All these variables are intensive; they all have complex physical units, except temperature (simple unit) and pH (no unit). Assuming that some form of random sampling has been conducted with constant-sized observation units, how could estimates be obtained for the whole study area? This question may be viewed from two different angles, i.e. one may be looking for a mean or for an integral value over the study area. For additive variables (i.e. all except pH), values can be computed that represent the mean over the study area. However, integrating over the study area to obtain values for total incident solar energy, zooplankton, etc. is not that simple, because it requires the variables to be extensive. No extensive variable can be derived from temperature or pH. In the case of variables with complex physical units, new variables may be derived with units that are appropriate for integration:

- Consider O_2 concentration. Its physical dimensions (Section 3.1) are $[ML^{-3}]$, with units g m^{-3}. This indicates that the "mass" part (dimension [M], with unit g), which is extensive, may be integrated over a volume, for example that of the surface mixed

layer over the whole study area. Also, values from different depths in the mixed layer may be vertically integrated, to provide areal concentrations (dimensions $[ML^{-2}]$, with units g m^{-2}). The same applies to the concentration of zooplankton.

- Flux variables can be turned into variables that are additive over both space and time. Phytoplankton production (dimensions $[ML^{-3}T^{-1}]$, with units g C m^{-3} s^{-1}) is a flux variable since it is expressed per unit space and time. The extensive "mass" part may be integrated over a volume or/and over time, e.g. the euphotic zone over the whole study area or/and for the duration of the study. Values from different depths in the euphotic zone may be vertically integrated, thus providing areal concentrations (dimensions $[ML^{-2}T^{-1}]$, with units g C m^{-2} s^{-1}), which can then be integrated over time.

- Incident solar energy (W m^{-2}) represents a more complex case. The "power" part (W) can be integrated over space (m^2) only. However, because W = J s^{-1} (Table 3.2), it is possible to integrate the "energy" part (J) over both space and time. Since incident solar energy is either in W m^{-2} or J m^{-2} s^{-1}, the "power" part may be integrated over space or, alternatively, the "energy" part may be integrated over both surface (m^2) and time (s). For example, one can compute solar energy over a given area during 24 h.

1.5 Coding

Coding is a technique by which original data are transformed into other values, to be used in the numerical analysis. All types of descriptors may be coded, but nonordered descriptors must necessarily be coded before they may be analysed numerically. The functions or laws of correspondence used for coding qualitative descriptors are generally discontinuous; positive integers are usually associated with the various states.

Consider the case where one needs to compute the dependence between a variable with a high degree of precision and a less precisely recorded descriptor. Two approaches are available. In the first approach, the precision of the more precise descriptor is lowered, for example by dividing continuous descriptors into classes. Computers can easily perform such transformations. Dependence is then computed using a mathematical method adapted to the descriptor with the *lowest* level of precision. In the second approach, the descriptor with the lower precision level will be given a numerical scale adjusted to the more precise one. This operation is called *quantification* (Cailliez & Pagès, 1976; Gifi, 1990); one method of quantification through canonical correspondence analysis is described in Subsection 11.2.2. Other transformations of variables, that adjust a descriptor to another, have been developed in the regression framework discussed in Section 10.3.

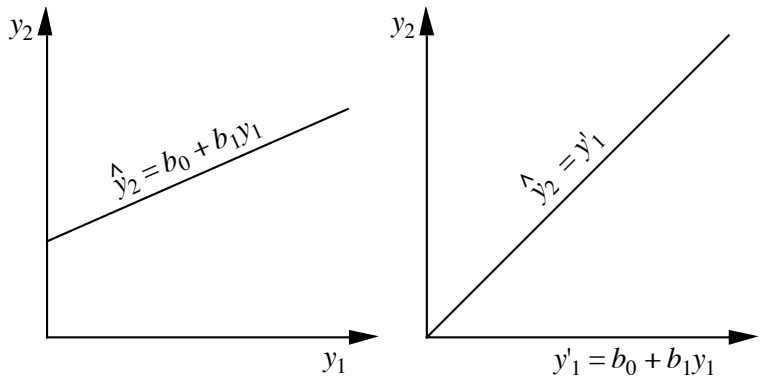


Figure 1.9 The regression parameters (b_0 and b_1) found by regressing y_2 on y_1 (left panel) may be used (right panel) to transform y_1 into y'_1 such that y'_1 is now on the same scale as y_2.

1 — Linear transformation

In a study where there are quantitative descriptors of different types (metres, litres, mg L^{-1}, …), it may be useful to put them all on the same scale in order to simplify the mathematical forms of relationships. It may be difficult to find an ecological interpretation for a relationship that includes a high level of artificial mathematical complexity, where scale effects are intermingled with functional relationships. Such changes of scale may be linear (of the first order), or of some higher order.

A linear change of scale of variable y is described by the transformation $y' = b_0 + b_1 y$ where y' is the value after transformation. Two different transformations are actually included in this equation. The first one, *translation*, consists in adding or subtracting a constant (b_0 in the equation) to all data. Graphically, this consists in sliding the scale beneath the data distribution. Translation is often used to bring to zero the mean, the modal class, the weak point of a bimodal distribution, or another point of interest in the distribution. The second transformation, *expansion*, is a change of scale obtained by multiplying or dividing all observed values by a constant (b_1 in the equation). Graphically, this operation is equivalent to contracting or expanding the scale beneath the distribution of a descriptor.

Two variables that are linearly related can always be put on the same scale by a combination of an expansion followed by a translation, or the opposite, the values of parameters b_0 and b_1 being found by linear regression (model I or model II: Chapter 10). For example (Fig. 1.9), if a linear regression analysis shows the equation relating y_2 to y_1 to be $\hat{y}_2 = b_0 + b_1 y_1$ (where $\hat{y}_2$ represents the values estimated by the regression equation for variable y_2), then transforming y_1 into $y'_1 = b_0 + b_1 y_1$ successfully puts variable y_1 on the same scale as variable y_2, since $\hat{y}_2 = y'_1$. If one

wishes to transform y_2 instead of y_1, the regression equation should be computed the other way around.

2 — Nonlinear transformations

The methods of multidimensional analysis described in this book are often based on covariances or linear correlations. Using them requires that the relationships among variables be made linear by an appropriate transformation. When two variables are not linearly related, their relationship may be described by a second- or higher-degree equation, or by other functional forms, depending on the situation. If the nonlinear form of the equation is derived from ecological theory, as it is often the case in population dynamics models, interpretation of the relationship poses no problem. If, however, a nonlinear transformation is chosen empirically, for reasons of mathematical elegance and without grounding in ecological theory, it may be difficult to find an ecological meaning to it.

The relationship between two variables may be determined with the help of a scatter diagram of the objects in the plane formed by the variables. The principles of analytical geometry may then be used to recognize the type of relationship (Fig. 1.10), which in turn determines the most appropriate type of transformation. A relationship frequently found in ecology is the exponential function, in which a variable y_2 increases in geometric progression with respect to y_1, according to one of the following equations:

$$y_2 = b^{(y_1)} \text{ or } y_2 = b_0 b_1^{(y_1)} \text{ or } y_2 = b_0 b_1^{(y_1 + b_2)} \text{ or else } y_2 = b_0 b_1^{(b_2 y_1)} \qquad \textbf{(1.3)}$$

Logarithmic transformation

depending on the number of constants b that shift or amplify the function. Such relationships can easily be linearized by using the logarithm of variable y_2 (called y'_2 below) instead of y_2 itself. The above relationships then become:

$$y'_2 = \log(y_2) = b' y_1, \text{ or } y'_2 = b'_0 + b'_1 y_1,$$

$$\text{or } y'_2 = b'_0 + b'_1 (y_1 + b_2), \text{ or } y'_2 = b'_0 + b'_1 b_2 y_1 \qquad \textbf{(1.4)}$$

where the b's are the logarithms of constants b in eq. 1.3.

If two variables display a logarithmic relationship of the form

$$y_2 = \log_b (y_1) \qquad \textbf{(1.5)}$$

where b is the base of the logarithm, their relationship can be made linear by applying a $\log^{-1}$ transformation to y_2:

$$y'_2 = b^{(y_2)} = y_1 \qquad \textbf{(1.6)}$$

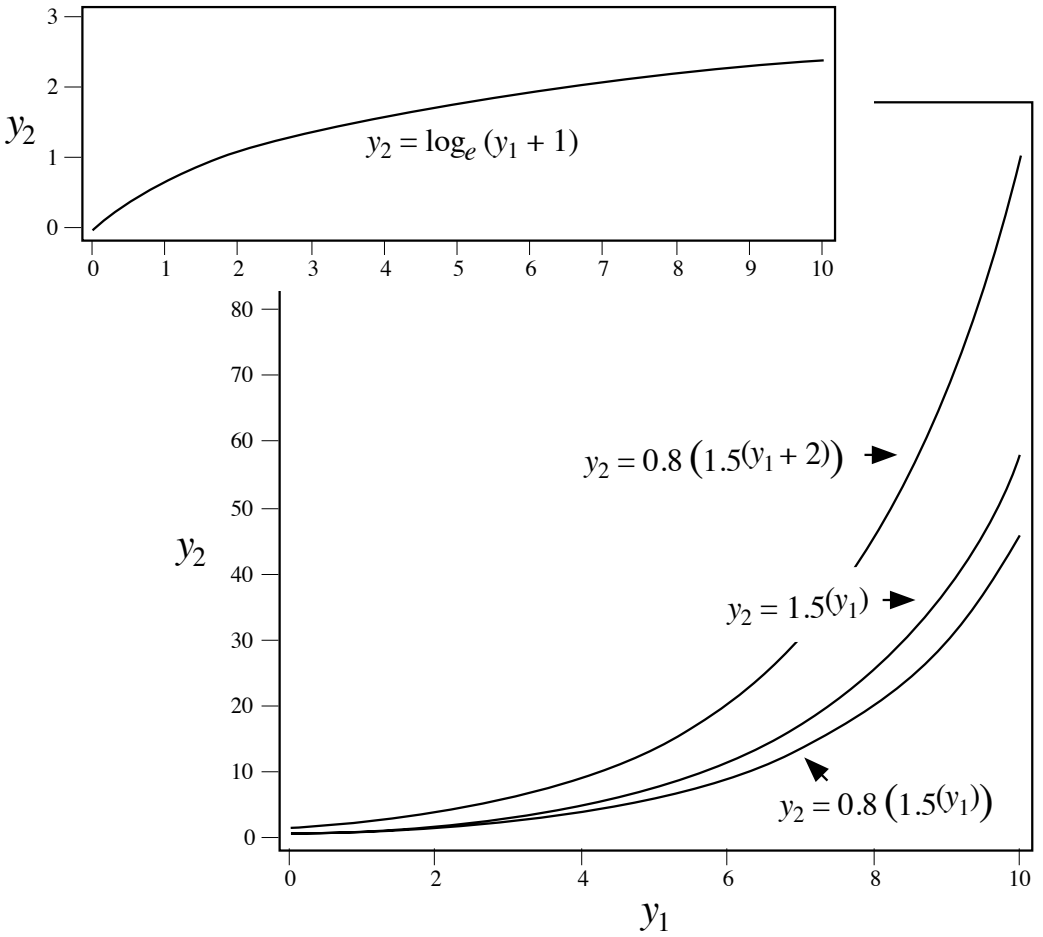


Figure 1.10 The relationship between variables may often be recognized by plotting them one against the other. In the upper panel, y_2 varies as the natural logarithm of y_1. In the lower panel, y_2 is an exponential function of y_1. These curves (and corresponding equations) may take different forms, depending on the modifying constants b, b_0, b_1 and b_2 (eq. 1.3).

When a nonlinear form can be assumed from knowledge of the ecological process involved, the corresponding equation can be used as the basis for a linearizing transformation. For instance, the nonlinear equation

$$N_t = N_0 e^{rt} \tag{1.7}$$

describes the exponential growth of a population, as observed in population explosions. In this equation, the independent variable is time (t); N_0 and N_t are the

population sizes at times 0 and t, respectively; r is the Malthus parameter describing the intrinsic rate of increase of the population. This nonlinear equation indicates that N_t should be transformed into its natural logarithm to make the relationship linear. After this transformation, $\log_e(N_t)$ is linearly related to t: $\log_e(N_t) = \log_e(N_0) + rt$.

3 — Combining descriptors

Another transformation that is often used consists in combining different descriptors by addition, subtraction, multiplication or division. In limnology, for example, the ratio (surface O_2/ bottom O_2) is often used as a descriptor. So is the Pearsall ionic ratio, all ions being in the same physical units:

$$y = \frac{\mathrm{Na} + \mathrm{K}}{\mathrm{Mg} + \mathrm{Ca}} \tag{1.8}$$

Beware, however, of the spurious correlations that may appear when comparing a ratio variable $\mathbf{y}/\mathbf{z}$ to $\mathbf{z}$, or two ratio variables $\mathbf{y}_1/\mathbf{z}$ and $\mathbf{y}_2/\mathbf{z}$ (Pearson 1897). Jackson & Somers (1991a) illustrate the problem using simulated data and recommend that such correlations be tested using permutation tests (Section 1.2) involving permutation of the parent variables, followed by construction of the ratios from the permuted variables and computation of the correlation coefficient under permutation.

Permutation test

One may want to take into account a factor of magnitude or size. For example, when observation units are of different sizes, the number of specimens of each species may be divided by the area or the volume of the unit (depending on whether the units come from an area or a volume), or by some other measure of the sampling effort. One must exert care when interpreting the results, however, since large observation units are more representative of populations and have smaller variances than small ones.

4 — Ranging and standardization

Quantitative variables, used in ecology as environmental descriptors, are often expressed in incompatible units such as metres, mg L^{-1}, pH units, etc. In order to compare such descriptors, or before using them together in a classification or ordination procedure, they must be brought to some common scale. Among the methods available, some only eliminate size differences while others reduce both the size and variability to a common scale.

Translation, a method previously discussed, allows one to *centre* the data, eliminating size differences due to the position of the zero on the various scales. Centring is done by subtracting the mean of the observations $(\bar{y})$ from each value y_i:

$$y'_i = y_i - \bar{y} \tag{1.9}$$

For relative-scale variables (Subsection 1.4.1), dividing each y_i by the largest observed value is a way, based on expansion, to bring all values in the range [0, 1] (Cain & Harrison, 1958):

$$y'_i = y_i / y_{max} \quad \textbf{(1.10)}$$

For interval-scale variables, whose range may include negative values, the absolute value of the largest positive or negative value is used as divisor. The transformed values are in the interval [–1, +1].

Ranging

Other methods allow the simultaneous adjustment of the magnitude and the variability of the descriptors. The method of *ranging*, proposed by Sneath & Sokal (1973), reduces the values of a variable to the interval [0, 1] by first subtracting the minimum observed for each variable and then dividing by the range:

$$y'_i = \frac{y_i - y_{min}}{y_{max} - y_{min}} \quad \textbf{(1.11)}$$

For relative-scale variables (Subsection 1.4.1) for which y_{min} is always zero, ranging can be achieved as well with eq. 1.10.

Standardization

The most widely used method for making descriptors compatible is to *standardize* the data (transformation into so-called "z-scores"). This method will be fully discussed in Section 4.2, which deals with correlation. Principal components (Section 9.2) are frequently computed using standardized data. Standardization is achieved by subtracting the mean (translation) and dividing by the standard deviation (s_y) of the variable (expansion):

$$z_i = \frac{y_i - \bar{y}}{s_y} \quad \textbf{(1.12)}$$

The position of each object on the transformed variable z_i is expressed in standard deviation units; as a consequence, it refers to the group of objects from which s_y has been estimated. The new variable z_i is called a *standardized variable*. Such a variable has three interesting properties: its mean is zero ($\bar{z} = 0$); its variance and hence its standard deviation are 1 ($s_z^2 = s_z = 1$); it is also a *dimensionless variable* (Chapter 3) since the physical dimensions (metres, mg L^{-1}, etc.) in the numerator and denominator cancel out. Transformations 1.8, 1.10 and 1.11 also produce dimensionless variables.

Beware of the "default options" of computer programs that may implicitly or explicitly suggest to standardize all variables before data analysis. Milligan & Cooper (1988) report simulation results showing that, for clustering purposes, if a transformation is needed, the ranging transformation (eqs. 1.10 and 1.11) gives results that are in general better to those obtained using standardization (eq. 1.12).

5 — Implicit transformation in association coefficients

When descriptors with different scales are used together to compare objects, the choice of the association coefficient (Section 7.6) may partly determine the type of transformation that must be applied to the descriptors. Some coefficients give equal weights to all variables independently of their scales while others take into account the magnitude of variation of each one. Since the amount of information (in the sense of information theory; Chapter 6) in a quantitative descriptor increases as a function of its variance, equalizing the variances before the association coefficient is computed is a way to ensure that all descriptors have the same weight. It is for ecologists to decide the kind of contribution they expect from each descriptor; again, beware of the "default options" of computer programs.

Some association coefficients require that the data be expressed as integers. Depending on the capabilities of the computer program and the degree of discrimination required, ecologists may decide to use the closest integer value, or to multiply first all values by 10 or 100, or else to apply some other simple transformation to make the data compatible with the coefficient to be computed.

6 — Normalization

Another type of transformation, called *normalizing transformation*, is performed on descriptors to make the frequency distributions of their data values look like the normal curve of errors — or, at least, as unskewed as possible. Indeed, several of the methods used in multivariate data analysis have been developed under the assumption that the variables are normally distributed. Although most of these methods do not actually require full normality (i.e. no skewness nor kurtosis), they may perform better if the distributions of values are, at least, not skewed. Skewed distributions, as in Fig. 1.11, are such that the variance of the distribution is controlled mostly by the few points in the extreme right tail; so, variance-partitioning methods such as principal component analysis (Chapter 9) or spectral analysis (Chapter 12) would bring out components expressing the variation of these few data points first instead of the variation of the bulk of data values. Normalizing transformations also have the property of reducing the *heteroscedasticity* of descriptors (Box 1.4). The data analysis phase of research should always start by looking at the distributions of values for the different variables, i.e. computing basic distribution statistics (including skewness and kurtosis, eqs. 4.41 and 4.42), drawing histograms of frequency distributions, and testing for normality (described in Section 4.6). A normalizing transformation may have to be found for each variable separately; in other cases, one is looking for the best transformation that would normalize several variables.

• 6.1 — Ecologists often encounter distributions where a species is abundant in a few observation units (quadrats, etc.), fairly abundant in more, present in even more, and absent in many; this is in agreement with the concept of ecological niche briefly explained in Section 1.0, if the sampling programme covers a large enough area or environmental gradient. Distributions of this type are clearly not normal, being

Homoscedasticity Box 1.4

Homoscedasticity, also called *homogeneity* or *equality of the variances*, technically means that the variances of the error terms are equal for all observations. Its antonym is **heteroscedasticity** or *heterogeneity of the variances*. Homoscedasticity may actually refer to different properties of the data.

- *For a single variable*, homoscedasticity of the distribution means that, when the statistical population is sampled repeatedly, the expected value of the variance remains the same, whatever the value of the mean of the data sample. Data drawn from a normal distribution possess this property whereas data drawn from a Poisson distribution, for instance, do not, since the variance is equal to the mean in this type of distribution.

- *In regression analysis*, homoscedasticity means that, for all values of the independent variable, the variances of the corresponding values of the response variable (called error variances or variances of the residuals) are the same.

- *In t-test, analysis of variance and discriminant analysis*, homoscedasticity means that variances are equal in all groups, for each variable.

strongly skewed to the right (long tail in the higher values). Needless to say, environmental variables may also have non-normal distributions. For instance, the scales on which chemical variables are measured are conventions of chemistry which have no relation whatsoever with the processes generating these values in nature. So, any normalizing transformation is as good as the scale on which these data were originally measured.

Skewed data are often transformed by taking logarithms (below) or square roots. *Square root* is the least drastic transformation and is used to normalize data that have a Poisson distribution, where the variance is equal to the mean, whereas the *logarithmic transformation* is applicable to data that depart more widely from a normal distribution (Fig. 1.11). Several intermediate transformations have been proposed between these two extremes (Fig. 1.12): cubic root, $\log^2$, $\log^p$, etc. The *hyperbolic transformation* is useful for one particular type of data, which share the two extreme types at the same time (when the standard deviation is proportional to the mean, with many observations of a very small size which follow a Poisson distribution: Quenouille, 1950; Barnes,

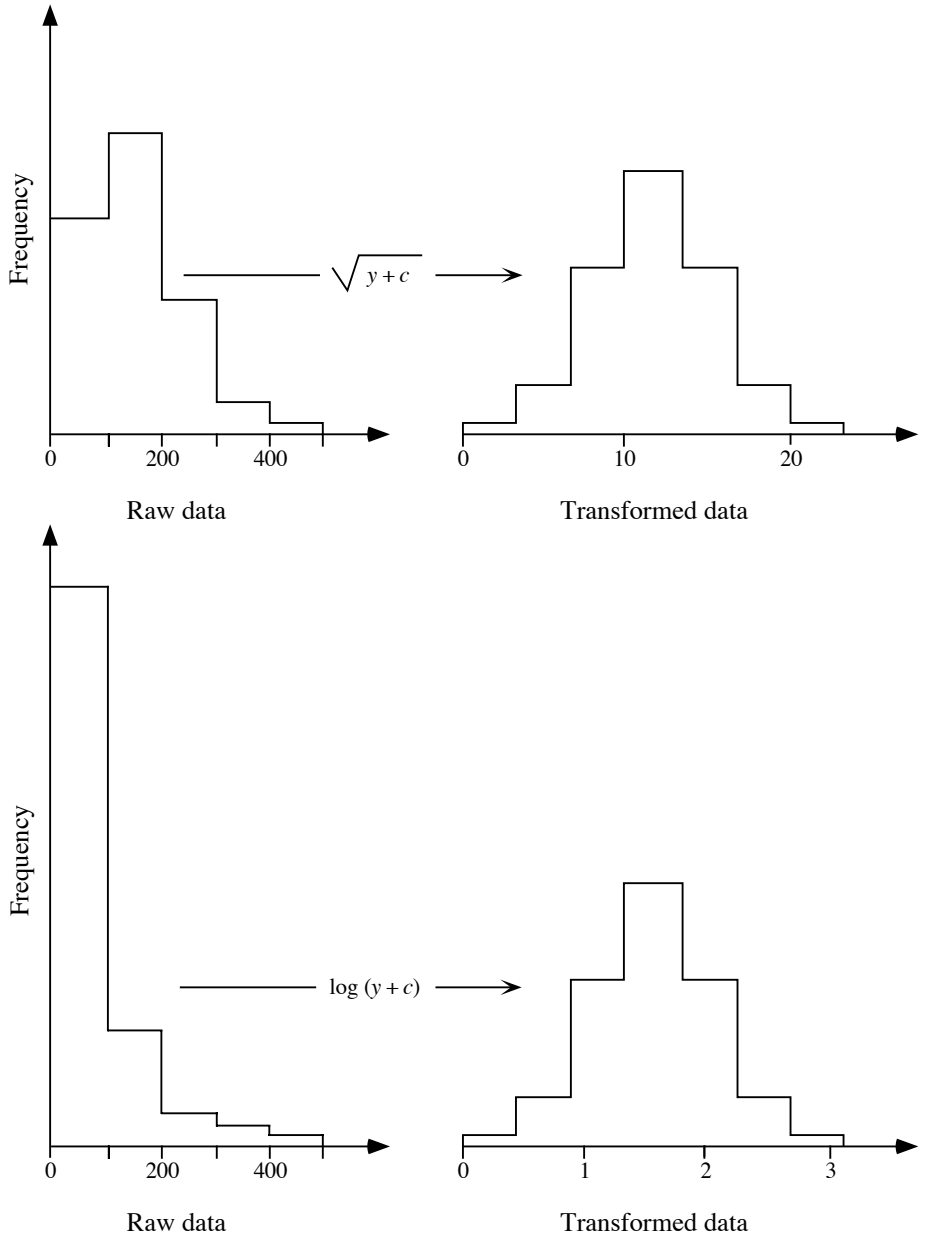


Figure 1.11 Numerical examples. Upper panel: Data that follow a Poisson distribution (left) can be normalized by the square root transformation (right). For a given species, these frequencies may represent the number of quadrats (ordinate) occupied by the number of specimens shown along the abscissa. Lower panel: Data distribution (left) that can be normalized by a logarithmic transformation (right).

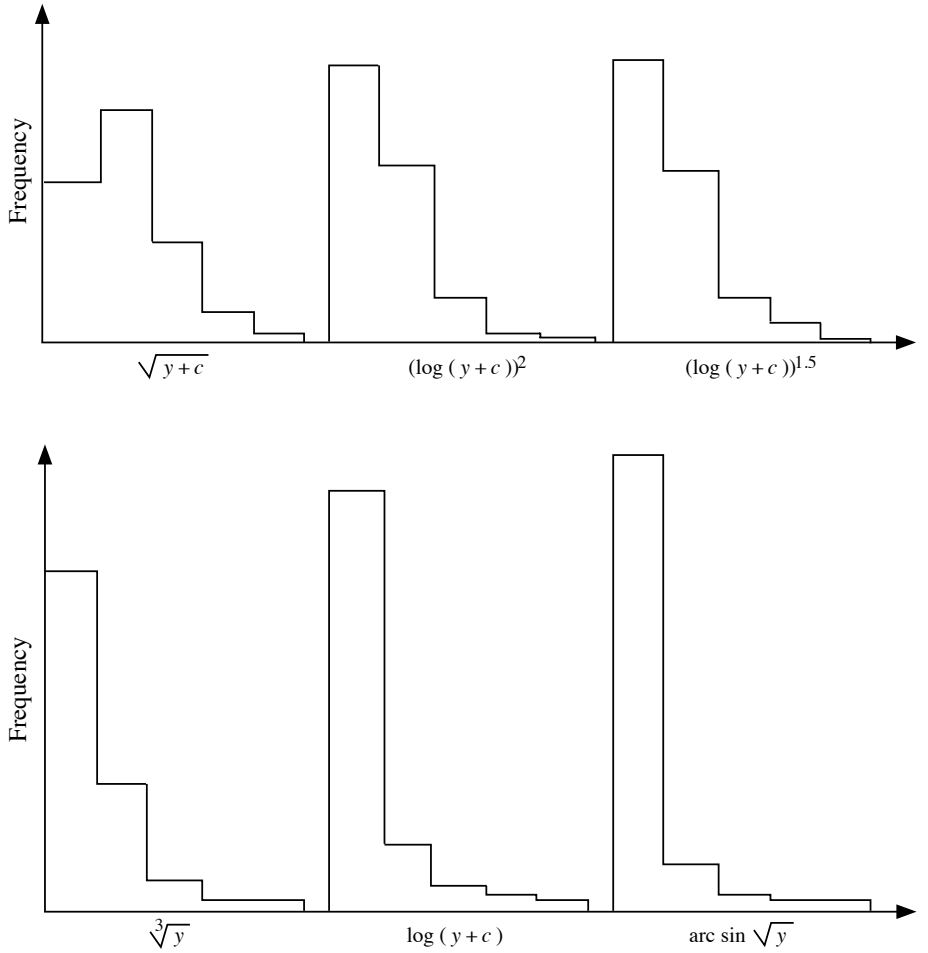


Figure 1.12 Numerical examples. Each histogram is labelled by the normalizing transformation to be used in that case. The bottom rightmost histogram refers to a simplified version of the hyperbolic transformation.

1952). The *angular* or *arcsine transformation* is appropriate for percentages and proportions (Sokal & Rohlf, 1981, 1995):

$$y_i' = \arcsin\sqrt{y_i} \qquad \textbf{(1.13)}$$

In case of doubt, one may try several of these transformations and perform a test of normality (Section 4.6), or compute the skewness of the transformed data, retaining the transformation that produces the most desirable results. Alternatively, the Box-Cox method (point 6.2, below) may be used to find the best normalizing transformation.

A logarithmic transformation is computed as follows:

Logarithmic transformation

$$y'_i = \log(b_0 + b_1 y_i) \qquad \textbf{(1.14)}$$

The base of logarithm chosen has no influence on the normalising power, since transformation from one base (c) to another (d) is a linear change of scale (expansion, see Subsection 1.5.1: $\log_d y_i = \log_c y_i / \log_c d$). When the data to be transformed are all strictly positive (all $y_i > 0$), it is not necessary to carry out a translation ($b_0 = 0$ in eq. 1.14). When the data contain fractional values between 0 and 1, one may multiply all values by some appropriate constant in order to avoid negative transformed values: $y'_i = \log(b_1 y_i)$. When the data to be transformed contain negative or null values, a translation must be applied first, $y'_i = \log(b_0 + y_i)$, since the logarithmic function is defined over the set of positive real numbers only. One should choose for translation a constant b_0 that is of the same order of magnitude as the significant digits of the variable to be transformed; for example, $b_0 = 0.01$ for data between 0.00 and 0.09 (the same purpose would have been achieved by selecting $b_0 = 1$ and $b_1 = 100$ for these data). For species abundance data, this rule produces the classical transformation $y'_i = \log(y_i + 1)$.

Box-Cox method

• 6.2 — When there is no *a priori* reason for selecting one or the other of the above transformations, the Box-Cox method allows one to empirically estimate the most appropriate exponent of the following general transformation function:

$$y'_i = (y_i^{\gamma} - 1)/\gamma \qquad (\text{for } \gamma \neq 0) \qquad \textbf{(1.15)}$$

and

$$y'_i = \log_e(y_i) \qquad (\text{for } \gamma = 0)$$

As before, y'_i is the transformed value of observation y_i. In this transformation, the value γ is used that maximizes the following log likelihood function:

$$L = -(\nu/2)\log_e(s^2_{y'}) + (\gamma - 1)(\nu/n)\sum_i \log_e(y_i) \qquad \textbf{(1.16)}$$

since it is this value that yields the best transformation to normality (Box & Cox, 1964; Sokal & Rohlf, 1995). The value L that maximizes the likelihood function is found by iterative search. In this equation, $s^2_{y'}$ is the variance of the *transformed* values y'_i. When analysing several groups of observations at the same time (below), $s^2_{y'}$ is estimated instead by the within-group, or residual variance computed in a one-way ANOVA. The group size is n and ν is the number of degrees of freedom ($\nu = n - 1$ if the computation is made for a single group). All y_i values must be strictly positive numbers since logarithms are taken in the likelihood function L (eq. 1.16); all values may easily be made strictly positive by translation, as discussed in Subsection 1.5.1. It is interesting to note that, if $\gamma = 1$, the function is a simple linear transformation; if $\gamma = 1/2$, the function becomes the square root transformation; when $\gamma = 0$, the transformation is logarithmic; $\gamma = -1$ yields the reciprocal transformation.

Readers are invited to take a value (say 150) and transform it, using eq. 1.15, with a variety of values of γ gradually tending toward 0 (say 1, 0.1, 0.01, 0.001, etc.). Comparing the results to the logarithmic transformation will make it clear that the natural logarithm is indeed the limit of eq. 1.15 when γ tends towards 0.

Another log likelihood function L' is proposed by Sokal & Rohlf (1995) to achieve homogeneity of the variances for several groups of observations of a given variable, together with the normality of their distributions. This generalized Box-Cox transformation may also be applied to the identification of the best normalizing transformation for several species, for a given set of sampling sites.

• 6.3 — When the data distribution includes several groups, or when the same transformation is to be applied to several quantitative and dimensionally homogeneous descriptors (Chapter 3; for instance, a species abundance data table), Taylor's (1961) power law provides the basis for another general transformation that stabilizes the variances and thus makes the data *more likely* to conform to the assumptions of parametric analysis, including normality (Southwood, 1966; see also Downing, 1979 on this subject). This law relates the means and variances of the k groups through the equation

Taylor's power law

$$s^2_{y_k} = a\,(\bar{y}_k)^b \qquad \textbf{(1.17)}$$

from which constants a and b can be computed by nonlinear regression (Subsection 10.3.6). When the latter is not available, an approximation of b may be calculated by linear regression using the logarithmic form

$$\log s^2_{y_k} = \log a + b \log \bar{y}_k \qquad \textbf{(1.18)}$$

Having found the value of b, the variance stabilizing transformations

$$y'_i = y_i^{\left(1-\frac{b}{2}\right)} \qquad (\text{for } b \neq 2) \qquad \textbf{(1.19)}$$

or

$$y'_i = \log_e(y_i) \qquad (\text{for } b = 2)$$

are applied to the data.

• 6.4 — The following method represents an *omnibus normalizing procedure*, which is able to normalize most kinds of data. The procedure is easy to carry out in R or using a standard statistical packages. The package must have a pseudo-random number generator for random normal deviates, i.e. values drawn at random from a normal distribution.

Omnibus procedure

(1) Write the quantitative or semiquantitative descriptor to be normalized into a vector or a column of a spreadsheet. Sort the vector in order of increasing values.

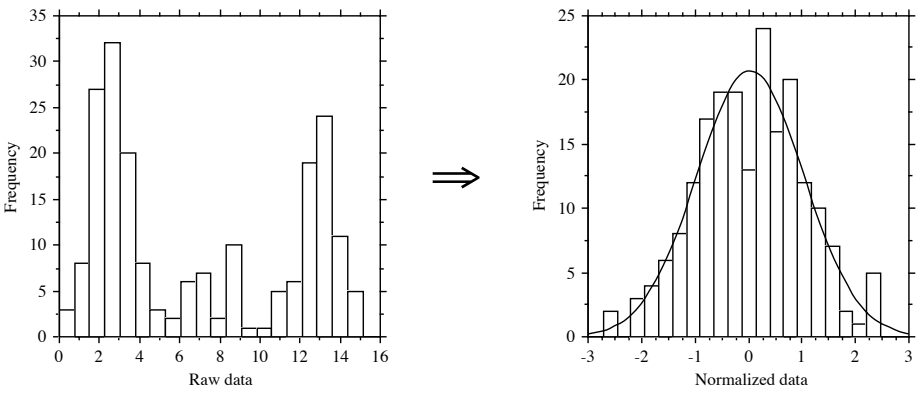


Figure 1.13 The *omnibus* procedure is used here to normalize a set of 200 data values with tri-modal distribution (left). A normal curve is fitted to the normalized data (right). The normalized data could be rescaled to the approximate range of the original data through the linear transformation $y_{rescaled} = 8 + (y_{normalized} \times 16/5.5)$ where 16 is the approximate range of the raw data and 5.5 that of the normalized data; the constant 8 makes all rescaled values positive.

(2) Create a new descriptor with the same number of values, using a pseudo-random normal deviate generator (***rnorm()*** is the function to use in R). Sort this new vector in order of increasing values. (3) Bind the two vectors, or copy the sorted normal deviate values besides the first sorted vector in the spreadsheet. Sort the bound vectors or the spreadsheet back into the original order if necessary. (4) Use the normal deviates as a monotonic proxy for the original descriptor. Figure 1.13 shows an example of this transformation. (5) It may be useful in some cases to rescale the normalized data to the approximate range of the original data through a linear transformation.

This procedure may be modified to handle *ex aequo* (tied) values (Section 5.3). Tied values may either receive the same normal deviate value, or they may be sorted in some random order and given neighbouring normal deviate values; one should select a solution that makes sense considering the data at hand.

Data transformed in this way may be used in methods of data analysis that perform better in the presence of normally distributed data. Several such methods will be studied in Chapters 9 and 11. The main disadvantage is that a back-transformation is difficult. If the study requires that values of the transformed descriptor be forecasted by a model, the database itself will have to be used to find the original descriptor values that are the closest to the forecasted normal deviate. An interpolation may have to be made between observed data values.

7 — Dummy variable coding

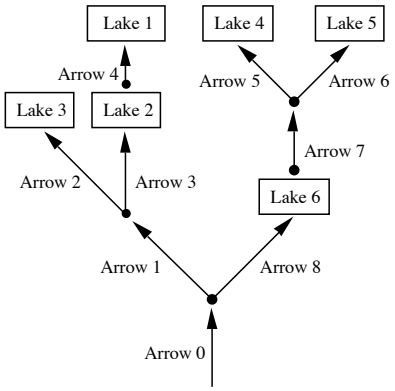
Multistate qualitative descriptors may be binary-coded as *dummy variables*. This coding is interesting because it allows the use of qualitative descriptors in procedures such as multiple regression, discriminant analysis or canonical analysis, which have been developed for quantitative variables and in which binary variables may also be used. A multistate qualitative descriptor with s states can be decomposed into $(s - 1)$ dummy variables V_j, as shown by the following example of a four-state descriptor:

States	Dummy variables			
	V_1	V_2	V_3	V_4
1	1	0	0	0
2	0	1	0	0
3	0	0	1	0
4	0	0	0	1

In this example, three dummy variables, e.g. V_1 to V_3, are sufficient to code for the four states of the nominal descriptor, excluding V_4. Had dummy variable V_4 been included (shaded column above), its information would have been totally *linearly dependent* (Box 1.1 and Section 2.7) on the first three variables, meaning that it would have been entirely predictable from the sum of the other three variables and the intercept represented by a column vector of 1: $V_4 = \mathbf{1}_{\text{intercept}} - (V_1 + V_2 + V_3)$. This shows that the first three dummy variables are sufficient to determine the four states of the multistate qualitative descriptor. Actually, any one of the four dummy variables may be eliminated to return to the condition of linear independence among the remaining ones. Using the coding table above, the objects are coded by three dummy variables instead of a single 4-state descriptor. An object with state 1, for instance, would be recoded as [1 0 0], an object with state 2 as [0 1 0], and so on.

There are other methods to code for a qualitative variable or a factor of an experiment. Helmert contrasts are now briefly described. Consider an experimental factor with s levels. The first Helmert variable contrasts the first and second levels; the second variable contrasts the third level to the first two; the third variable contrasts level 4 to the first three; and so on. The coding rule for Helmert contrasts is illustrated by the following examples:

2 groups: 1 variable	**3 groups: 2 variables**	**4 groups: 3 variables**	**5 groups: 4 variables**	**etc.**
$\begin{bmatrix} -1 \\ +1 \end{bmatrix}$	$\begin{bmatrix} -1 & -1 \\ +1 & -1 \\ 0 & +2 \end{bmatrix}$	$\begin{bmatrix} -1 & -1 & -1 \\ +1 & -1 & -1 \\ 0 & +2 & -1 \\ 0 & 0 & +3 \end{bmatrix}$	$\begin{bmatrix} -1 & -1 & -1 & -1 \\ +1 & -1 & -1 & -1 \\ 0 & +2 & -1 & -1 \\ 0 & 0 & +3 & -1 \\ 0 & 0 & 0 & +4 \end{bmatrix}$	



	Arrows							
	1	2	3	4	5	6	7	8
Lake 1	1	0	1	1	0	0	0	0
Lake 2	1	0	1	0	0	0	0	0
Lake 3	1	1	0	0	0	0	0	0
Lake 4	0	0	0	0	1	0	1	1
Lake 5	0	0	0	0	0	1	1	1
Lake 6	0	0	0	0	0	0	0	1

Figure 1.14 Lakes interconnected by a river network (left) can be binary-coded as shown in the table to the right. Numbers are assigned in an arbitrary order to the directional edges (arrows) of the network. It is not useful to code the root of the network (arrow 0) in the matrix since all lakes would be coded '1' for that arrow. This example is revisited in Subsection 14.3.1.

Contrasts can be constructed based on some quantitative variable of interest associated with the objects, instead of the levels of a qualitative variable. Polynomial contrasts are based on an orthogonal polynomial of the quantitative variable of interest. The reference variable may be the position of the observations along a transect or a time series, or along an ecological gradient of altitude, pH, humidity, and so on. The contrasts are the successive monomials of the polynomial of the variable of interest, centred and made orthogonal to the lower-degree monomials; the monomials are then usually standardized to have a sum-of-squares of 1. Polynomial contrasts are used as explanatory variables in analyses in the same way as Helmert contrasts.

Other forms of coding have been developed for special types of variables. In phylogenetic analysis, the states of multistate characters are sometimes related by a hypothesized transformation series, going from the single hypothesized ancestral state to all the advanced states; such a series can be represented by a directed network where the states are connected by arrows representing evolutionary progression. A transformation series may be coded into binary variables using a method proposed by Kluge & Farris (1969).

River network

This same method may be applied to code the spatial relationships among localities in a geographic network. An example in freshwater ecology is a group of lakes connected by a river network (Fig. 1.14). In this example, a pseudo-map containing rivers and lakes is drawn to represent the network. A number is assigned to each river segment (the river segments are the edges of the connected graph) while nodes represent the furcation points. In Fig. 1.14, the coding is based on the river segments; it could just as well be based on the nodes if one felt that the nodes were the important

carriers of geographic information (as in Magnan *et al*., 1994). If the phenomenon to be modelled is, for example, fish dispersal from downstream, the arrows can be drawn going upstream, as in Fig. 1.14. In the lake-by-arrow matrix, a value '1' is assigned to each arrow found downstream from a lake, representing the fact that the corresponding river segment may allow fish to travel from the root to that lake. All other arrows are coded '0' for that lake. The resulting matrix is a complete numerical coding of the hydrographic network information: knowing the coding procedure, one can entirely reconstruct the network topology from the matrix entries.

The coding method may be tailored to the ecological problem at hand. For a dispersion phenomenon going downstream, arrows could point the other way around; in that case, a lake would be coded '1' in the table for arrows arriving in that lake from upstream. The pattern of interconnections does not even need to be a tree-like structure; it may form a more general type of directed network, but no cycle is allowed. Coding the information allows the use of this type of geographical information in different types of numerical models, like multiple regression (Chapter 10) or canonical analysis (Chapter 11). In many of these methods, zeros and ones are interchangeable. This coding method for directional spatial processes will be further developed in Section 14.3 where it will serve as the basis for the *Asymmetric Eigenvector Maps* (AEM) method of spatial analysis.

1.6 Missing data

Ecological data matrices are often plagued by missing data. The latter do not necessarily result from negligence on the part of the field team; most often, they are caused by the breakdown of measuring equipment during field surveys, weather events that prevented sampling sites from being visited on a given date, lost or incorrectly preserved specimens, improper sampling procedures, and so on.

Three families of solutions are available to cope with this problem for the analysis of field survey data, if one can make the assumption that the missing values occur at random in the data set. Most of the approaches mentioned below are discussed by Little & Rubin (1987), who also proposed methods for estimating missing values in controlled experiments (when the missing values are only found in the outcome variable; their Chapter 2) as well as valid model-based likelihood estimation of missing values for situations where the distribution of missing values does not meet the randomness assumption stated above.

Missing values may be represented in data matrices by numbers that do not correspond to possible data values. Codes such as –1 or –9 are often used when the real data in the table are all positive numbers, as it is the case with species abundance data; otherwise, –99 or –999, or other such unambiguous codes, may be used. In spreadsheets, missing values are often represented by bullets or 'NA' symbols.

1 — Delete rows or columns

Delete any row or column of the data matrix (Section 2.1) containing missing values. If a few rows contain most of the missing values, proceed by *rowwise* (also called *listwise*) *deletion*; conversely, if most missing values are found in a few variables only, proceed by *columnwise deletion*. This is the simplest, yet the most costly method, as it throws away the valuable information present in the remainder of these rows or columns.

2 — Accommodate algorithms to missing data

Accommodate the numerical method in such a way that the missing values are skipped during calculations. For instance, when computing resemblance coefficients among rows (Q-mode) or columns (R-mode) of the data matrix (Chapter 7), a simple method is *pairwise deletion* of missing values. This means, for example, that when computing a correlation coefficient between variables $\mathbf{y}_1$ and $\mathbf{y}_2$, if the value of the tenth object is missing for $\mathbf{y}_2$, object $\mathbf{x}_{10}$ is skipped in the computation of this correlation value. When it comes to comparing $\mathbf{y}_1$ and $\mathbf{y}_3$, if $\mathbf{x}_{10}$ has no missing data for these variables, it is then kept in the calculation for this pair of variables. However, one must be aware that covariance and correlation matrices computed in this way may be indefinite (i.e. they may have negative eigenvalues; Table 2.2).

3 — Estimate missing values

Estimate the missing values, a method called *imputation* by Little & Rubin (1987). This is the best strategy when missing values are located all over the data matrix — contrary to the situation where the missing values are found in a few rows or columns only, in which case deletion of these rows or columns may be the strategy of choice. The assumption one has to make when estimating missing values is that the missing data are not grossly atypical compared to those present in the data set. Methods for estimating missing data are interesting in cases where the numerical algorithm required for analysing the data cannot accommodate missing values. Ecologists should never imagine, however, that the estimated values are ecologically meaningful; as a consequence, they should refrain from attempting to interpret these numbers in ecological terms. Ecologists should also keep in mind that the estimation procedure has not created the missing degrees of freedom that would have accompanied observations carried out in nature or in the laboratory.

Three groups of methods are available for replacing quantitative missing values.

• 3.1 — The easiest way, which is often used in computer programs, is to replace missing values by the mean of the variable, estimated from the values present in the data table. When doing so, one assumes that nothing is known about the data, outside of the weak assumption mentioned above that the missing value comes from the same population as the non-missing data. Although this solution produces covariance and correlation matrices that are positive semidefinite (Section 2.10), the variances and

covariances are systematically underestimated. One way around this problem is to select missing value estimates at random from some distribution with appropriate mean and variance. This is not recommended, however, when the relative positions of the objects are of interest (principal component analysis; Section 9.1). A variant of the same method is to use the median instead of the mean; it is more robust in the sense that it does not assume the distribution of values to be unskewed. It is also applicable to semiquantitative descriptors. For qualitative descriptors, use the most frequent state instead of the mean or median.

• 3.2 — Estimate the missing values by regression. Multiple linear regression (Section 10.3), with rowwise deletion of missing values, may be used when there are only a few missing values to estimate. The dependent (response) variable of the regression is the descriptor with missing value(s) while the independent (explanatory) variables are the other descriptors in the data table. After the regression equation has been computed from the objects without missing data, it can be used to estimate the missing value(s). Using this procedure, one assumes the descriptor with missing values to be linearly related to the other descriptors in the data table (unless some form of nonparametric or nonlinear multiple regression is being used) and the data to be approximately multivariate normal. This method also leads to underestimating the variances and covariances, but less so than in 3.1. An alternative approach is to use a regression program allowing for pairwise deletion of missing values in the estimation of the regression coefficients, although, in that case, a maximum likelihood estimation of the covariance matrix would be preferable (Little & Rubin, 1987, p. 152 *et seq*.).

If such a method cannot be used for estimating the covariance matrix and if the missing values are scattered throughout the data table, an approximate solution may be obtained as follows. Compute a series of simple linear regressions with pairwise deletion of missing values, and estimate the missing value from each of these simple regression equations in turn. The mean of these estimates is taken as the working estimated value. The assumptions are basically the same as in the multiple regression case (above). Other methods of imputation are available in specialized R packages; see Section 1.7.

To estimate missing values in qualitative (nominal) descriptors, use logistic regression (Section 10.3) instead of linear regression.

• 3.3 — Interpolate missing values in spatially correlated data. Positive spatial correlation (Section 1.1) means that near points in time or space are similar. This property allows the interpolation of missing or otherwise unknown values from the values of near points in the series. With spatial data, interpolation is the first step of any mapping procedure, and it may be done in a variety of ways (Subsection 13.2.2), including the kriging method developed by geostatisticians. The simplest such method is to assign to a missing data the value of its nearest neighbour. In time series, interpolation of missing values may be performed using the same methods; see also Shumway & Stoffer, 1982, and Mendelssohn & Cury, 1987, for a maximum likelihood method for estimating missing data in a time series using a state-space model.

Myers (1982, 1983, 1984) proposed a method, called co-kriging, that combines the power of principal component analysis (Section 9.1) with that of kriging. It allows the estimation of unknown values of a data series using both the values of the same variable at neighbouring sites and the known values of other variables, correlated with the first one, observed at the same or neighbouring points in space; the spatial inter-relationships of these variables are measured by a cross-variogram. This method is interesting for the estimation of missing data in broad-scale ecological surveys and to compute values at unobserved sites on a geographic surface.

1.7 Software

The methods presented in this introductory chapter are implemented in the R language.

1. Corrections for multiple testing (Box 1.1) can be done using the ***p.adjust()*** function of the STATS package.

2. Several R functions use permutation tests. They will be identified in later chapters where permutation-based statistical methods are presented. For R functions that do not rely on compiled code for intensive calculations, permutations are produced by the ***sample()*** function of the STATS package. That function can also carry out resampling with replacement (bootstrapping).

3. All standard statistical distributions, and many others, are available in the STATS package. To find out about them, type in the R console: ***help.search(****"distribution", package="stats"****)***. Additional statistical distributions are available in other R packages.

4. Ranging and standardization (Subsection 1.5.4), as well as other transformations, are available in the ***decostand()*** function of VEGAN. Variable standardization is also available through the ***scale()*** function of STATS. The Box-Cox transformation (Subsection 1.5.6) can be done using the ***boxcox.fit()*** function of the GEOR package.

5. Helmert contrasts are available in the ***contr.helmert()*** function of the STATS package; polynomial contrasts can be computed using the ***contr.poly()*** function of the same package. Contrast matrices corresponding to actual data files are generated using the ***model.matrix()*** function of the STATS package; this function calls ***contr.helmert()*** or ***contr.poly()*** for calculation of the contrasts.

6. Imputation of missing values using a principal component analysis model is available in function ***imputePCA()*** of MISSMDA. Function ***mice()*** of package MICE carries out multivariate imputation by chained equations.