1

Introduction

1.1 The nature of paleontological data

Paleontology is a diverse field, with many different types of data and a corresponding variety of analytical methods. For example, the types of data used in ecological, morphological, and phylogenetic analyses are often quite different in both form and quality. Data relevant to the investigation of morphological variation in Ordovician brachiopods are quite different from those gathered from Neogene mammal-dominated assemblages for paleoecological analyses. Nevertheless, there is a surprising commonality of techniques that can be implemented in the investigation of such apparently contrasting data sets.

1.1.1 Univariate measurements

Perhaps the simplest type of data is straightforward, continuous measurements such as length or width. Such measurements are typically made on a number of specimens in one or more samples, commonly from collections from different localities or species. Typically, the investigator is interested in characterizing, and subsequently analyzing, the sets of measurements, and comparing two or more samples to see how they differ. Measurements in units of arbitrary origin (such as temperature on the Celsius or Fahrenheit scales) are known as *interval* data, while measurements in units of a fixed origin (such as distance and mass) are called *ratio* data. Methods for interval and ratio data are presented in chapter 4.

1.1.2 Bivariate measurements

Also relatively easy to handle are bivariate interval or ratio measurements on a number of specimens. Each specimen is then characterized by a *pair* of values, such as both length and width of a given fossil. In addition to comparing samples, we will then typically wish to know if and how the two variables are interrelated. Do they, for example, fit on a straight line, indicating a static (isometric) mode of growth or do they exhibit more complex (anisometric) growth patterns? Bivariate data are discussed in chapters 4 and 6.

Paleontological Data Analysis, Second Edition. Øyvind Hammer and David A.T. Harper.

1.1.3 Multivariate morphometric measurements

The next step in increasing complexity is multivariate morphometric data, involving multiple variables such as length, width, thickness, and, say, distance between ribs (Fig. 1.1). We may wish to investigate the structure of such data and whether two or more samples are different. Multivariate data sets can be difficult to visualize, and special methods have been devised to emphasize any inherent structure. Special types of multivariate morphometric data include digitized outlines and the coordinates of landmarks. Multivariate methods useful for morphometrics are discussed in chapters 5 and 6.

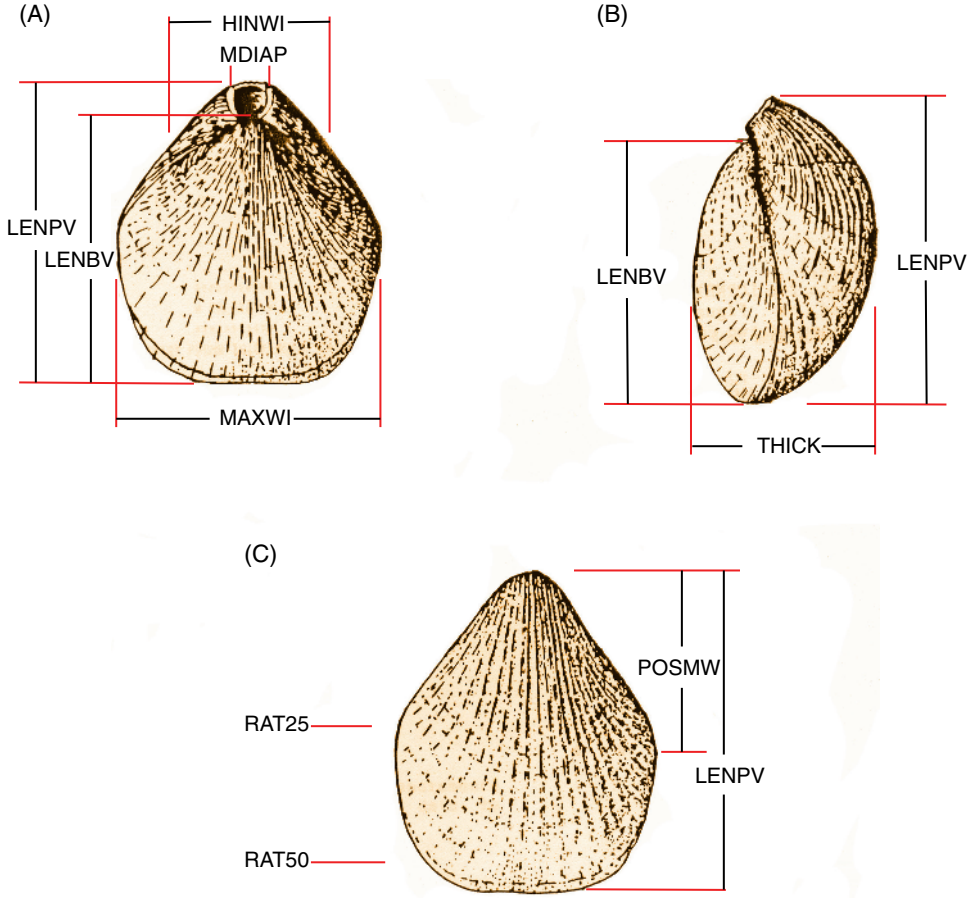

Figure 1.1 Typical set of measurements made on conjoined valves of the living brachiopod *Terebratulina* from the west coast of Scotland. A: Measurements of the length of ventral (LENPV) and dorsal (LENBV) valves, together with hinge (HINWI) and maximum (MAXWI) widths. B: Thickness (THICK). C: Position of maximum width (POSMW) and numbers of ribs per mm at, say, 2.5 mm (RAT25) and 5.0 mm (RAT50) from the posterior margin can form the basis for univariate, bivariate, and multivariate analysis of the morphological features of this brachiopod genus.

1.1.4 Character matrices for phylogenetic analysis

A special type of morphological (or molecular) data is the character matrix for phylogenetic analysis (cladistics). In such a matrix, taxa are conventionally entered in rows and characters in columns. The state of each character for each taxon is typically coded with an integer. Character matrices and their analyses are treated in chapter 14.

1.1.5 Paleoecology and paleobiogeography – taxa in samples

In paleoecology and paleobiogeography, the most common data type consists of taxonomic counts at different localities or stratigraphic levels. Such data are typically given in an *abundance matrix*, either with taxa in rows and samples in columns or vice versa. Each cell contains a specimen count (or abundance) for a particular taxon in a particular sample (Table 1.1).

In some cases, we do not have specimen counts available; instead, we only know whether a taxon is present or absent in the sample. Such information is specified in a *presence-absence table*, where absences are typically coded with zeros and presences with ones. This type of binary data may be more relevant to biogeographic or stratigraphic analyses, where each taxon is weighted equally.

Typical questions include whether the samples are significantly different from each other, whether the biodiversity is different, whether there are well-separated groups of localities, or any indication of an environmental gradient. Methods for analyzing taxa-in-samples data are discussed in chapters 5, 9, and 10 (Fig. 1.2).

1.1.6 Time series

A time series is a sequence of values through time. Paleontological time series include diversity curves, geochemical data from fossils through time, and thicknesses of bands from a series of growth increments. For such a time series, we may wish to investigate whether there is any trend or periodicity. Some appropriate methods are given in chapter 12.

1.1.7 Biostratigraphic data

Biostratigraphy is the correlation and zonation of sedimentary strata based on fossils. Biostratigraphic data are mainly of two different types. The first is a single table with localities in rows and events (such as the first or last appearance of a species) in columns.

Table 1.1 Example of an abundance matrix.

	Spooky Creek	Scary Ridge	Creepy Canyon
M. horridus	0	43	15
A. cornuta	7	12	94
P. giganteus	23	0	32

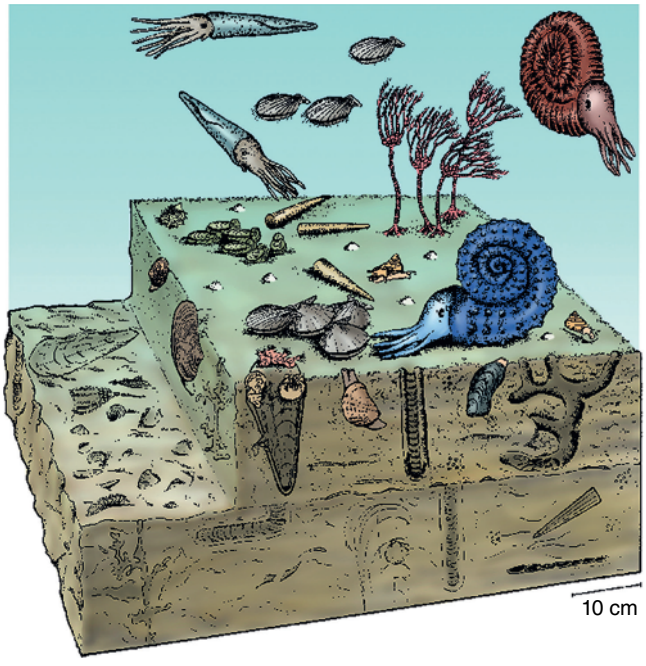

Figure 1.2 A muddy-sand community from the Early Jurassic. Community reconstruction is based on the presence of and partly the relative abundance of taxa. After Benton and Harper (2020), modified from McKerrow (1978).

Table 1.2 Example of an event table.

	***M. horridus* FAD**	***A. cornuta* FAD**	***M. horridus* FAD**
Spooky Creek	3.2	4.0	7.6
Scary Ridge	1.4	5.3	4.2
Creepy Canyon	13.8	13.8	15.9

FAD, first appearance datum.

Each cell of the table contains the stratigraphic level, in meters, feet, or ranked order, of a particular event at a particular locality (Table 1.2).

Such event tables form the input to biostratigraphic methods such as ranking-scaling and constrained optimization.

The second main type of biostratigraphic data consists of a single presence-absence matrix for each locality. Within each table, the samples are sorted according to stratigraphic level. Such data are used in the method of unitary associations.

Quantitative biostratigraphy is the subject of chapter 13 (Fig. 1.3).

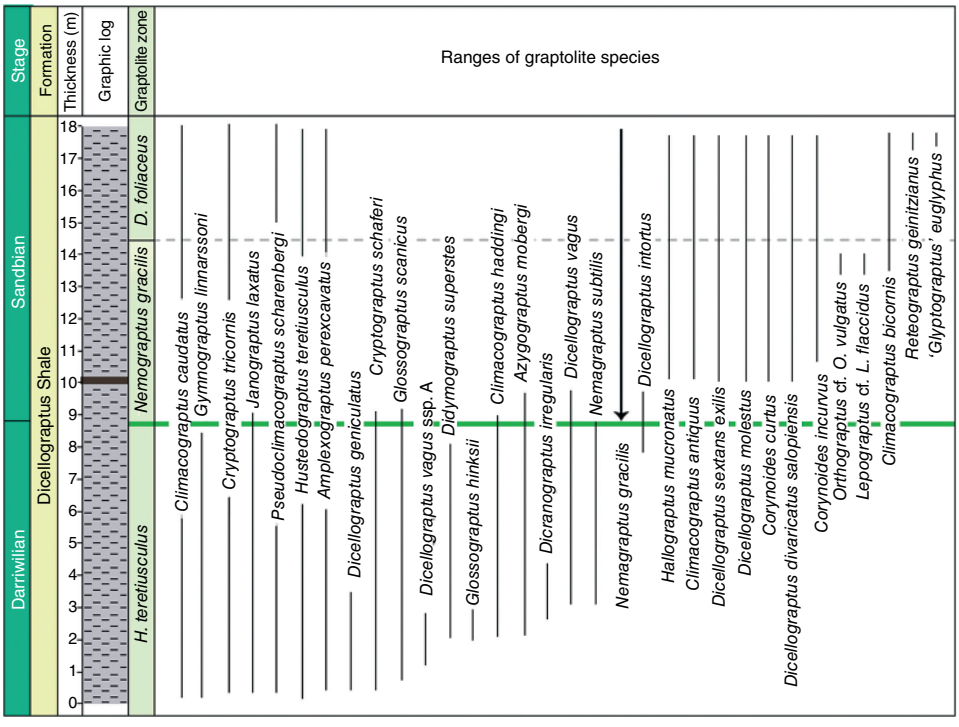

Figure 1.3 Range chart of fossils, mainly graptolites, through Darriwilian and Sandbian (Middle and Upper Ordovician) strata in southern Sweden. The boundary between the two stages (green horizontal line) is fixed to coincide with the first appearance of the graptolite *Nemagraptus gracilis*. Harper et al. (2022)/Geological Society of London.

1.2 Advantages and pitfalls of paleontological data analysis

During the last few decades, paleontology has comfortably joined the other sciences in its emphasis on quantitative methodologies. Statistical and explorative analytical methods, most of them depending on computers for their practical implementation, are now used in all branches of paleontology, from systematics and morphology to paleoecology and biostratigraphy. Over the last 100 years, techniques have evolved with available hardware, from the longhand calculations of the early twentieth century, through the time-consuming mainframe implementation of algorithms in the mid-twentieth century to the microcomputer revolution of the late twentieth century.

In general terms, there are three main components to any paleontological investigation. A detailed description of a taxon or community is followed by an analysis of the data (this may involve a look at ontogenetic or size-independent shape variation in a taxon or the population dynamics and structure of a community) with finally a comparison with other relevant and usually similar paleontological units. Numerical techniques have greatly enhanced the description, analysis, and comparison of fossil taxa and assemblages. Scientific hypotheses can be more clearly framed and statistically tested with numerical data.

The use of rapid computer-based algorithms has rendered even the most complex multivariate techniques accessible to virtually all researchers. Nevertheless, this development has attracted considerable discussion, and critical comments are occasionally made. Such criticisms, if not simply the result of ignorance or misplaced conservatism (which is now rarely the case), are mainly directed at the occasional abuse of quantitative data analysis methods. We will ask the reader to have the following checklist in mind when contemplating the application of a specific method to a specific problem.

1.2.1 Data analysis for the sake of it

First of all, one should always have a concrete problem to solve, or at least some hope that interesting and useful, presently unknown, information may emerge from the analysis (the "data mining" approach is sometimes denounced, but how can we otherwise find new territory to explore?). From time to time, we see technically impressive articles with beautiful illustrations that are almost devoid of new, scientifically important content. Technical case studies showing the application of new methods are of course of great interest, but then they should be clearly labeled as such. Do not use a complicated method of data analysis to show something that would be clearer and more obvious with a simpler approach. Shooting sparrows with cannons, or the smashing of open doors.

1.2.2 The Texas sharpshooter

The infamous Texas sharpshooter fired his gun at random, hit some arbitrary objects, and afterward claimed that he had been aiming at them in particular. In explorative data analysis, we are often "mining" for some yet unknown pattern that may be of interest and can lead to the development of new, testable hypotheses. The problem is that there are so many possible "patterns" out there that one of them is quite likely to turn up just by chance. It may be unlikely that one particular pattern would be found in a random data set (this we can often test statistically), but the probability that *any* pattern will be found is another matter. This subtle problem is always lurking in the background of explorative analysis, but there is really not much we can do about it, apart from keeping it in mind and trying to identify some process that may have been responsible for the pattern we observe.

1.2.3 Explorative method or hypothesis testing?

The distinction between data exploration and hypothesis testing is sometimes forgotten. Many of the techniques in this book are explorative, meaning that they are devices for the visualization of structure in complicated data sets. Such techniques include principal components analysis, cluster analysis, and even parsimony analysis. They are not statistical tests, and their results should in no way be presented as statistical "proof" (even if such a thing should exist, which it does not) of anything.

1.2.4 Incomplete data

Paleontological data are invariably incomplete – that is the nature of the fossil record, already pointed out by Charles Darwin and a number of his nineteenth century contemporaries. This problem is, however, no worse for quantitative than for qualitative analysis.

On the contrary, one of the main points of statistics is to understand the effects of the incompleteness of the data. Clearly, when it comes to this issue, much criticism of quantitative paleontological data analysis is really misplaced. Quantitative analysis can often reveal the incompleteness of the data and its importance – qualitative analysis often simply ignores it.

1.2.5 Statistical assumptions

All statistical methods make assumptions about the nature of the data. These include both "obvious" assumptions that always apply, such as the samples giving an unbiased representation of the parent population, and more specific assumptions such as the shape of the statistical distribution. Sometimes we are a little sloppy about the assumptions, in the hope that the statistical method will be robust to small violations, but then we should at least know what we are doing.

1.2.6 Statistical and biological significance

That a result is statistically significant does not mean that it is biologically or geologically significant. If sample sizes are very large, even minute differences between them may reach statistical significance, without being of any importance in, for example, adaptational, genetic, or environmental terms.

1.2.7 Circularity

Circular reasoning can sometimes be difficult to spot but let us give an obvious example. We have collected a thousand ammonites and want to test for sexual dimorphism. So, we divide the sample into small shells (microconchs) and large shells (macroconchs), and test for difference in size between the two groups. The means are found to differ with a high statistical significance, so the microconchs are significantly smaller than the macroconchs, and dimorphism is proved. Where is the obvious error in this procedure?

1.3 Software

A range of software is available for carrying out both general statistical procedures and more specialized analytical methods. We would like to mention a few, keeping in mind that new, excellent products are released every year.

A number of professional packages are available for general univariate and multivariate statistical analyses. These may be expensive, but have been thoroughly tested, produce high-quality graphic output, and come with extensive documentation. Examples are SAS and Stata. More specialized, free programs are available for special purposes, such as TNT and MrBayes for phylogenetic (cladistic) analysis, or RASC and CONOP for biostratigraphy.

However, such commercial or special-purpose, stand-alone applications have become much less important in the last 20 years, due to the enormous popularity of the statistical programming environment "R." Being basically a programming language, R gives near-total flexibility, but it also has a relatively high learning threshold. Most new statistical methods that are useful in paleontology are now released as packages for R. It also has very good functions for plotting high-quality figures.

The syntax of the R language is somewhat idiosyncratic and does not adhere to many of the standards of more general programming languages. Partly for this reason, the programming language Python is emerging as a strong competitor to R for data analysis. Python is perceived by many, especially trained programmers, as a more intuitive and elegant language than R, and the number of Python packages for data analysis is rapidly increasing.

In an attempt to lower the threshold into quantitative data analysis for paleontologists, we developed the PAST (PAlaeontological STatistics) software project, including a wide range of both general and special methods used within the field (Hammer et al. 2001). We regard this as a solid, extensive, and easy-to-use package, particularly useful for education but also popular in publication-quality research.

Although most of the examples and figures in this book were made with PAST, we will also provide examples using R and Python.

References

Benton, M.J., Harper, D.A.T. 2020. *Introduction to Paleobiology and the Fossil Record*. 2nd edition. Wiley-Blackwell, Chichester, UK.

Hammer, Ø., Harper, D.A.T., Ryan, P.D. 2001. PAST: Paleontological statistics software package for education and data analysis. *Palaeontologia Electronica* 4(1), 9.

Harper, D.A.T., Bown, P.R., Coe, A.L. 2022. Chronostratigraphy: understanding rocks and time. In Coe, A.L. (ed.), *Deciphering Earth's History: The Practice of Stratigraphy*. *Geoscience in Practice*, 227–243. Geological Society, London.

McKerrow, W.S. 1978. *Ecology of Fossils*. Duckworth Company Ltd., London.