Multivariate Analysis of Ecological Data

MICHAEL GREENACRE
Professor of Statistics at the Pompeu Fabra University in Barcelona, Spain

RAUL PRIMICERIO
Associate Professor of Ecology, Evolutionary Biology and Epidemiology
at the University of Tromsø, Norway

Chapter 1 Offprint

Multivariate Data in Environmental Science

First published: December 2013
ISBN: 978-84-92937-50-9

Supporting websites:
www.fbbva.es
www.multivariatestatistics.org

Fundación BBVA

Chapter 1

Multivariate Data in Environmental Science

In this introductory chapter we take a simple *univariate* or *bivariate* view of multivariate data, using a small educational example taken from marine biology. This means we will not venture beyond studying one or two variables at a time, using graphical displays as much as possible. Often we will show many of these representations simultaneously, which facilitates comparison and interpretation. The descriptive graphical methods that we use here – histograms, bar-charts and box-and-whisker plots – are well-known in any basic statistical course, and are invaluable starting points to what we could call a "marginal" understanding of our data before embarking on multivariate analysis. We encourage researchers to make as many graphical displays as possible of their data, to become acquainted with each variable, and to be aware of problems at an early stage, such as incorrect or very unusual values, or unusual distributions.

Contents

Data set "bioenv": Introductory data set from marine biology

As a simple introductory example to motivate and illustrate the concepts and methods explained in this book, consider the data in Exhibit 1.1. These are biological and environmental observations made at 30 sampling points, or *sites*, on the sea-bed. Typically, a number of grabs (for example, five) are made close by at each site and then a fixed volume of each grab is sent to a biological laboratory for analysis. The more grabs one takes at a site, the more species are eventually identified. The biological data consist of species

Exhibit 1.1: *Typical set of multivariate biological and environmental data: the species data are counts, whereas the environmental data are continuous measurements, with each variable on a different scale; the last variable is a categorical variable classifying the sediment of the sample as mainly* C *(= clay/silt),* S *(= sand) or* G *(= gravel/stone)*

Site No.	Species counts					Environmental variables			
	a	*b*	*c*	*d*	*e*	*Depth (x)*	*Pollution (y)*	*Temperature (z)*	*Sediment (s)*
s1	0	2	9	14	2	72	4.8	3.5	S
s2	26	4	13	11	0	75	2.8	2.5	C
s3	0	10	9	8	0	59	5.4	2.7	C
s4	0	0	15	3	0	64	8.2	2.9	S
s5	13	5	3	10	7	61	3.9	3.1	C
s6	31	21	13	16	5	94	2.6	3.5	G
s7	9	6	0	11	2	53	4.6	2.9	S
s8	2	0	0	0	1	61	5.1	3.3	C
s9	17	7	10	14	6	68	3.9	3.4	C
s10	0	5	26	9	0	69	10.0	3.0	S
s11	0	8	8	6	7	57	6.5	3.3	C
s12	14	11	13	15	0	84	3.8	3.1	S
s13	0	0	19	0	6	53	9.4	3.0	S
s14	13	0	0	9	0	83	4.7	2.5	C
s15	4	0	10	12	0	100	6.7	2.8	C
s16	42	20	0	3	6	84	2.8	3.0	G
s17	4	0	0	0	0	96	6.4	3.1	C
s18	21	15	33	20	0	74	4.4	2.8	G
s19	2	5	12	16	3	79	3.1	3.6	S
s20	0	10	14	9	0	73	5.6	3.0	S
s21	8	0	0	4	6	59	4.3	3.4	C
s22	35	10	0	9	17	54	1.9	2.8	S
s23	6	7	1	17	10	95	2.4	2.9	G
s24	18	12	20	7	0	64	4.3	3.0	C
s25	32	26	0	23	0	97	2.0	3.0	G
s26	32	21	0	10	2	78	2.5	3.4	S
s27	24	17	0	25	6	85	2.1	3.0	G
s28	16	3	12	20	2	92	3.4	3.3	G
s29	11	0	7	8	0	51	6.0	3.0	S
s30	24	37	5	18	1	99	1.9	2.9	G

abundances obtained by summing the counts of the species identified in the grabs for each site.

Usually there are dozens or hundreds of species found in an ecological study. Exhibit 1.1 is intentionally a small data set with only five species, labelled *a* to *e*. The number of sites, 30 in this case, is more realistic, because there are usually few sampling locations in marine environmental sampling. As well as the biological data, several environmental variables are typically available that characterize the sites. As examples of these we give four variables, three measurements and one classification – Exhibit 1.1 shows the values of depth *x* (in metres), a pollution index *y*, the temperature *z* in °C and the sediment type (three categories). The *pollution index* is based on data for heavy metal concentrations such as barium, cadmium and lead, as measured in the sea-bed samples – the higher the index, the higher is the overall level of pollution. The last column gives the classification of the sediment in the sample as clay/silt (C), sand (S) or gravel/stone (G). In this chapter we look at well-known univariate and bivariate summaries of these data, before we move on to a multivariate treatment.

Continuous variables

The three variables pollution, depth and temperature are called *continuous variables* because they can – theoretically, at least – have any value on a continuous scale. To take a look at the range of values as well as the shape of the distribution of a continuous variable, we typically plot a *histogram* for each variable – see the first three histograms in Exhibit 1.2. A histogram divides the range of a continuous variable into intervals, counts how many observations are in each interval, and then plots these frequencies. Pollution and temperature are seen to have single-peaked distributions, while the distribution of depth seems more uniform across its range.

Categorical variables

The sediment variable is a *categorical* (or *discrete*) variable because it can take only a few "values", which in this case are sediment types – see the *bar-chart* form of its distribution in Exhibit 1.2. A bar-chart simply counts how many observations correspond to each given category – this is why the bar-chart shows spaces between the categories, whereas in the histogram the frequency bars touch one another. The bar-chart shows a fairly even distribution, with gravel/rock (G) being the least frequent sediment category.

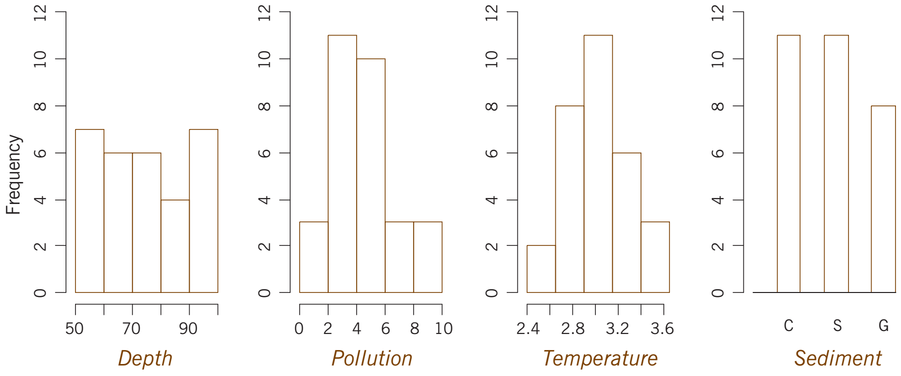


Exhibit 1.2:
Histograms of three environmental variables and bar-chart of the categorical variable

Categorical variables are either *ordinal* or *nominal* depending on whether the categories can be ordered or not. In our case, the categories could be considered ordered in terms of granularity of the sediment, from finest (clay/silt) to coarsest (gravel/rock). An example of a nominal variable, where categories have no inherent ordering, might be "sampling vessel" (if more than one ship was used to do the sampling) or "region of sampling" (if sampling was done in more than one region). Often, continuous variables are categorized into intervals (i.e., discretized), giving an ordinal variable; for example, "depth" could be categorized into several categories of depth, from shallow to deep.

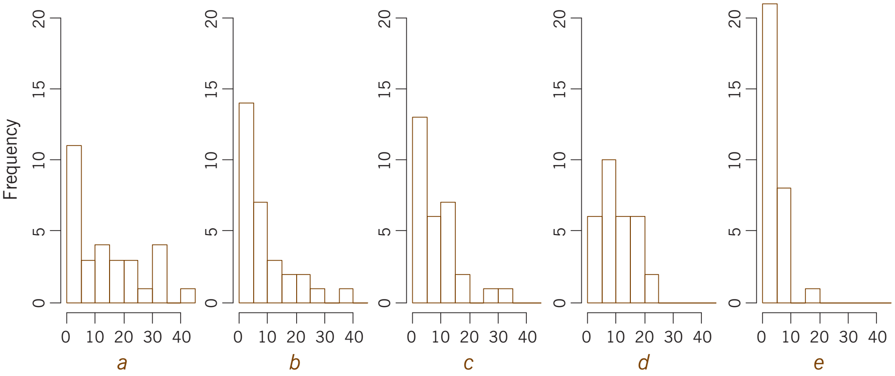
Count variables

The biological data are measured on a different scale from the others – these are *counts*, and have an integer scale: 0, 1, 2, and so on. Counts have a special place in statistics, lying somewhere between a continuous variable and a categorical variable. For the moment, however, we shall treat these data as if they were continuous variables; later we shall discuss various special ways to analyse them. Exhibit 1.3 shows histograms of the five species, with highly skew distributions owing to the many zeros typically found in such species abundance data.

Relationships amongst the environmental variables

The usual way to investigate the relationships between two continuous variables is to make a scatterplot of their relationship. The scatterplots of the three pairs of variables are shown in Exhibit 1.4, as well as the numerical value of their correlation coefficients. The only statistically significant correlation is between depth and pollution, a negative correlation of –0.396 ($p = 0.0305$, using the two-tailed t-test[1]).

Exhibit 1.3: *Histograms of the five species, showing the usual high frequencies of low values that are mostly zeros, especially in species* e



[1] Note that the t-test is not the correct test to use on such nonnormal data. An alternative is the distribution-free permutation test, which gives an estimated p-value of 0.0315, very close to the 0.0305 of the t-test. The permutation test for a correlation is described in Chapter 6, with a full treatment of permutation testing in Chapter 17.

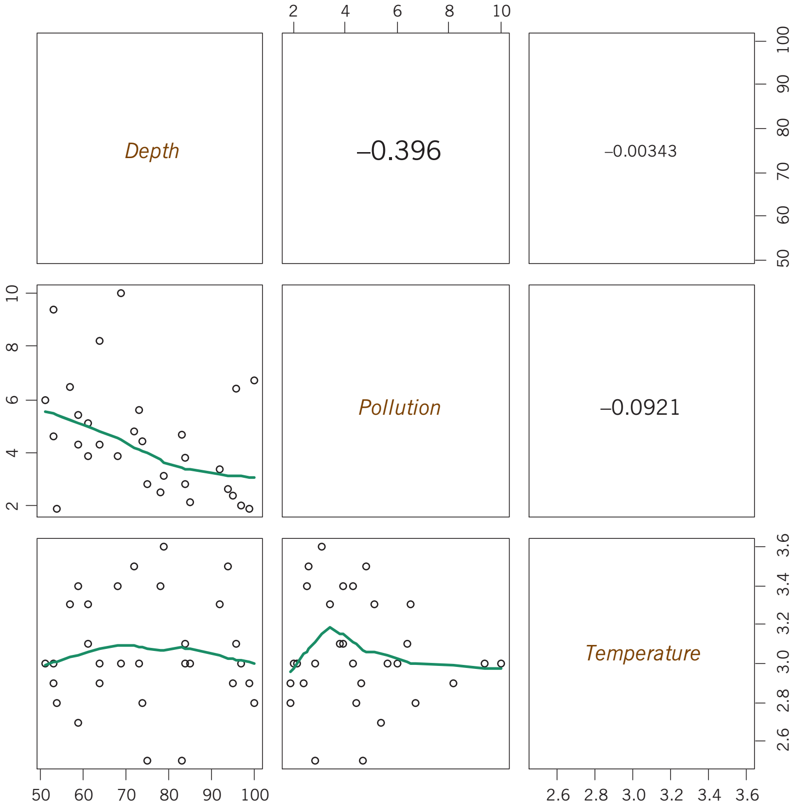
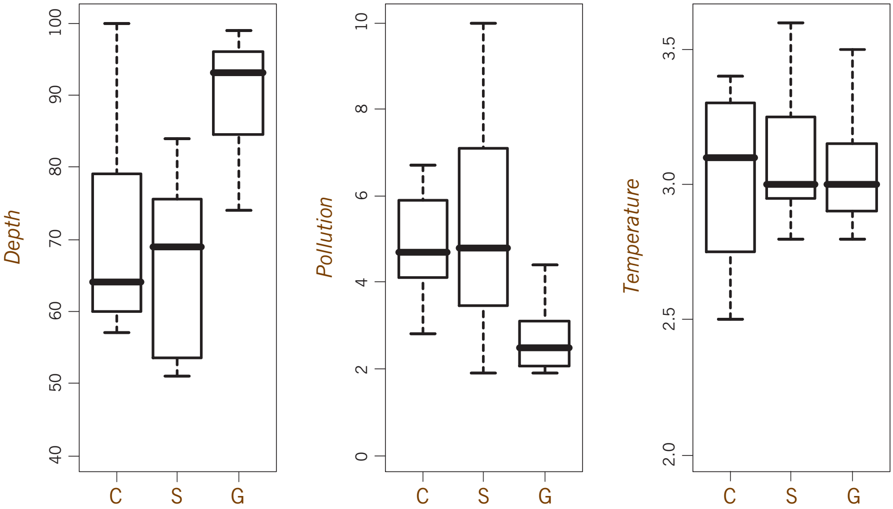


Exhibit 1.4:
Pairwise scatterplots of the three continuous variables in the lower triangle, showing smooth relationships (in brown, a type of moving average) of the vertical variable with respect to the horizontal one; for example, at the intersection of depth an pollution, pollution defines the vertical ("y") axis and depth the horizontal ("x") one. The upper triangle gives the correlation coefficients, with size of numbers proportional to their absolute values

To show the relationship between the continuous environmental variables and the categorical one (sediment), the *box-and-whisker* plots in Exhibit 1.5 compare the distributions of each continuous variable within each category. The boxes are drawn between the lower and upper quartiles of the distribution, hence the central 50% of the data values lie in the box. The median value is indicated by a line inside the box, while the whiskers extend to the minimum and maximum values in each case. These displays show differences between the gravel samples (G) and the other samples for the depth and pollution variables, but no differences amongst the sediment types with respect to temperature. A correlation can be calculated if the categorical variable has only two categories, i.e., if it is *dichotomous*. For example, if clay and sand are coded as 0 and gravel as 1, then the correlations[2] between the three variables depth, pollution and temperature, and

[2] This correlation between a continuous variable and a dichotomous categorical variable is called the *point biserial* correlation. Based on permutation testing (see Chapters 6 and 17), the *p*-values associated with these correlations are estimated as 0.0011, 0.0044 and 0.945 respectively.

Exhibit 1.5: *Box-and-whisker plots showing the distribution of each continuous environmental variable within each of the three categories of sediment* (C = *clay/silt,* S = *sand,* G = *gravel/stone). In each case the central horizontal line is the median of the distribution, the boxes extend to the first and third quartiles, and the dashed lines extend to the minimum and maximum values*



this dichotomous sediment variable, are 0.611, –0.520 and –0.015 respectively, confirming our visual interpretation of Exhibit 1.5.

Relationships amongst the species abundances

Similar to Exhibit 1.4 the pairwise relationships between the species abundances can be shown in a matrix of scatterplots (Exhibit 1.6), giving the correlation coefficient for each pair. Species a, b and d have positive inter-correlations, whereas c tends to be correlated negatively with the others. Species e does not have a consistent pattern of relationship with the others.

Relationships between the species and the continuous environmental variables

Again, using scatterplots, we can make a first inspection of these relationships by looking at each species-environmental variable pair in a scatterplot. The simplest way of modelling the relationship, although perhaps not the most appropriate way (see Chapter 18), is by a linear regression, shown in each mini-plot of Exhibit 1.7. The coefficient of determination R^2 (variance explained by the regression) is given in each case, which for simple linear regression is just the square of the correlation coefficient. The critical point for a 5% significance level, with $n = 30$ observations, is $R^2 = 0.121$ ($|R| = 0.348$); but because there are 15 regressions we should reduce the significance level accordingly. A conservative way of doing this is to divide the significance level by the number of tests, in which case the R^2 for significance is 0.236 ($|R| = 0.486$).[3] This would lead to the conclusion

[3] This is known as the *Bonferroni correction*. If many tests are performed, then the chance of finding a significant result by chance increases; that is, there is higher probability of committing a "type I" error. If the significance level is α and there are M tests, the Bonferroni correction is to divide α by M, then use the α/M significance level for the tests. This is a conservative strategy because the tests are usually not independent, so the correction overcompensates for the problem. But in any case, it is good to be conservative, at least in statistics!

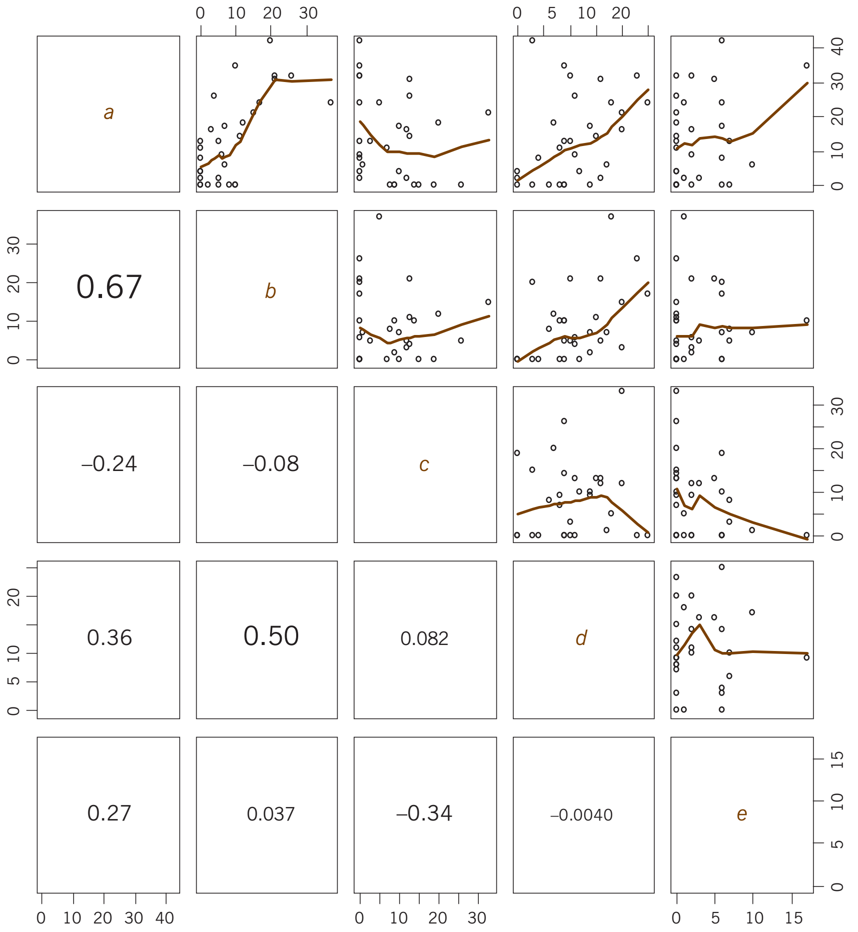


Exhibit 1.6:
Pairwise scatterplots of the five species abundances, showing in each case the smooth relationship of the vertical variable with respect to the horizontal one; the lower triangle gives the correlation coefficients, with size of numbers proportional to their absolute values

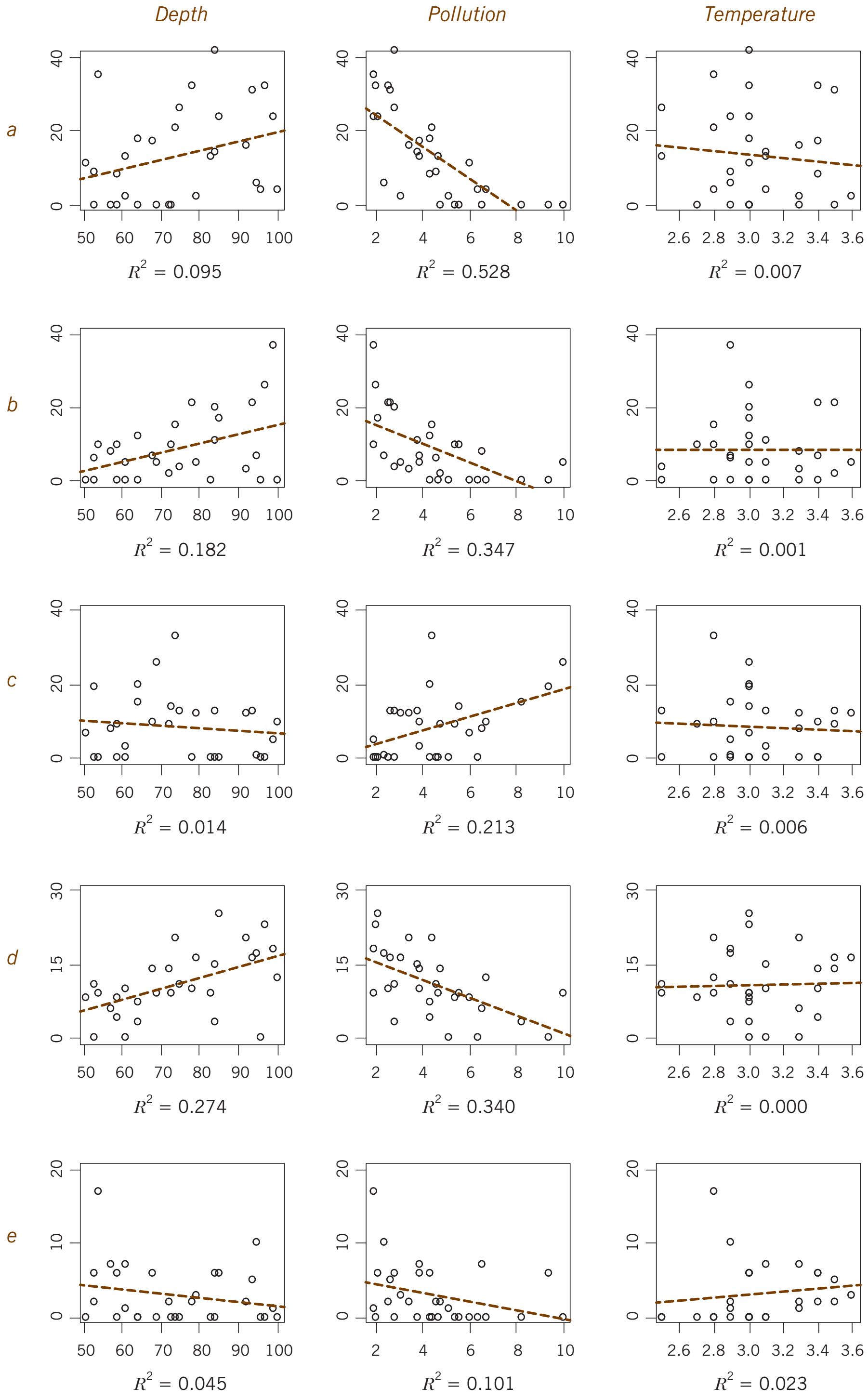
that *a*, *b* and *d* are significantly correlated with pollution, and that *d* is also significantly correlated with depth.

Relationships between the species and the categorical environmental variables

To show how the species abundances vary across the sediment groups, we use the same boxplot style of analysis as Exhibit 1.5, shown in Exhibit 1.8. The statistic that can be used to assess the relationship is the *F*-statistic from the corresponding analysis of variance (ANOVA). For this size data set and testing across three groups, the 5% significance point of the *F* distribution[4] is $F = 3.34$. Thus, species groups *a*, *b* and *d* show apparent significant differences between the sediment types, with the abundances generally showing an increase in gravel/stone.

[4] Note that using the *F*-distribution is not the appropriate way to test differences between count data, but we use it anyway here as an illustration of the *F* test.

Exhibit 1.7:
Pairwise scatterplots of the five groups of species with the three continuous environmental variables, showing the simple least-squares regression lines and coefficients of determination (R^2)



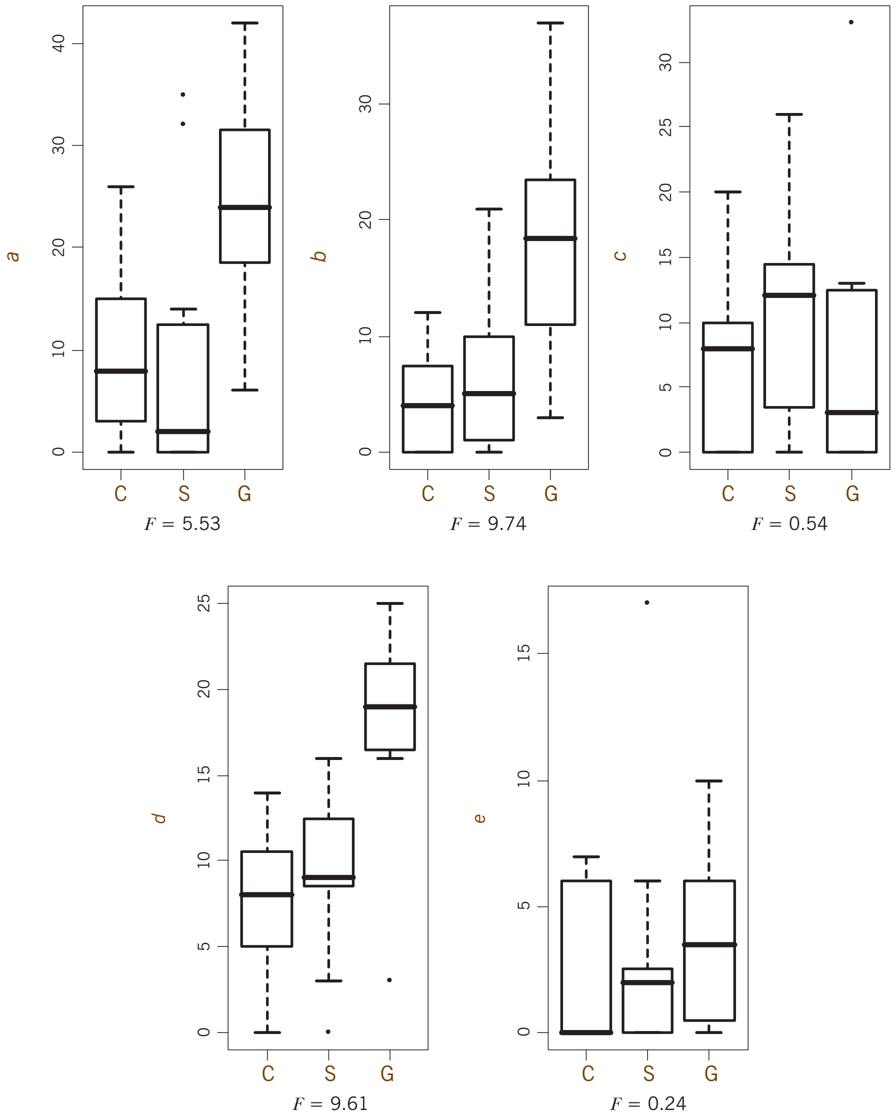


Exhibit 1.8:
Box-and-whisker plots showing the distribution of each count variable across the three sediment types (C = *clay/silt,* S = *sand,* G = *gravel/stone) and the F-statistics of the respective ANOVAs*

SUMMARY:
Multivariate data in environmental science

1. Large numbers of variables are typically collected in environmental research: it is not unusual to have more than 100 species, and more than 10 environmental variables.

2. The scale of the variables is either continuous, categorical or in the form of counts.

3. For the moment we treat counts and continuous data in the same way, whereas categorical data are distinct in that they usually have very few values.

4. The categorical data values do not have any numerical meaning, but they might have an inherent order, in which case they are called *ordinal*. If not, they are called *nominal*.

5. The univariate distributions of count and continuous variables are summarized in histograms, whereas those of categorical variables are summarized in bar-charts.

6. The bivariate distributions of continuous and count variables are summarized in typical "*x*-*y*" scatterplots. Numerically, the relationship between a pair of variables can be summarized by the correlation coefficient.

7. The relationship between a continuous and a categorical variable can be summarized using box-and-whisker plots side by side, one for each category of the categorical variable. The usual correlation coefficient can be calculated between a continuous variable and a dichotomous categorical variable (i.e., with only two categories).

List of Exhibits