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 3 Offprint

Measurement Scales, Transformation and Standardization

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

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

Fundación **BBVA**

Chapter 3

Measurement Scales, Transformation and Standardization

To conclude this introductory part on multivariate data analysis, we present a discussion about scales of measurement and the various possibilities for transforming variables. Questions such as the following can plague environmental biologists: "Should I log-transform my data?", "How do I analyse a data set where there is a mixture of continuous and categorical variables?", "My data are not normally distributed, does this matter? And if it does, help!", "Do I need to standardize my data?" and "My data are percentages that add up to 100: does this make a difference to the analysis?" The answers to some of these questions will only become fully apparent later, but at least in this chapter we will catalogue some of the issues involved and list some of the standard ways of transforming data. Readers can optionally skip this chapter for the moment if they are keen to proceed, and dip into it later as we refer back to these issues when they come up in real applications.

Contents

Data theory

Data are the manifestation of statistical variables, and these variables can be classified into various types according to their scales of measurement. Our own personal hierarchical classification of measurement scales is depicted in Exhibit 3.1. The main division, as we have mentioned already, is between categorical and continuous scales. This is a pragmatic distinction, because in reality all observed data are categorical. As Michael types these words his age is 59.5475888 years

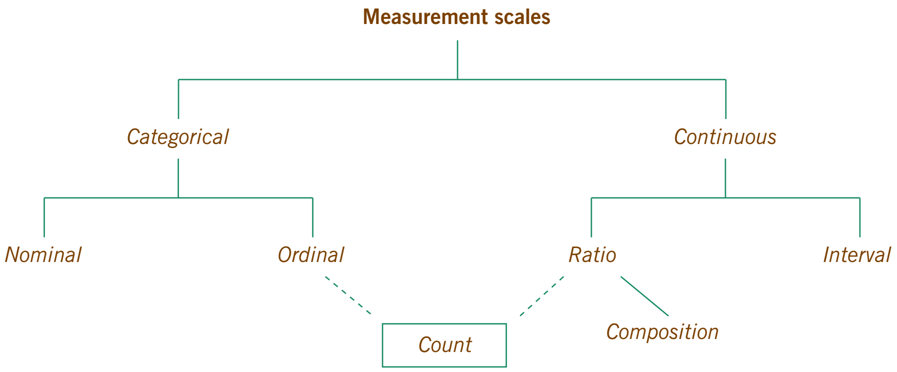

Exhibit 3.1: *A classification of data in terms of their measurement scales. A variable can be categorical (nominal or ordinal) or continuous (ratio or interval). Count data have a special place: they are usually thought of as ratio variables but the discreteness of their values links them to ordinal categorical data. Compositional data are a special case of ratio data that are compositional in a collective sense because of their "unit-sum" constraint*

(to seven decimal places and, theoretically, we could give it to you with even more accuracy), but it has now already advanced to 59.5475902 years in the interim, and is increasing by the second! Of course, in any statistical study, for example an epidemiological study, his age would be recorded simply as 59, having been *discretized*. But we still consider the value 59 to be the manifestation of the continuous variable "age".

Categorical data can be measured on a *nominal* or *ordinal* scale. Nominal categories have no ordering: for example, region of sampling (1. Tarehola, 2. Skognes, 3. Njosken, or 4. Storura), and habitat (1. pelagic, or 2. littoral), hence the numbers recorded in the database have no numerical meaning apart from assigning the samples into groups. Ordinal categories do have an ordering: for example, nature of substrate (1. clay, 2. silt, 3. sand, 4. gravel, or 5. stone – these are ordered by grain size), and month of sampling (1. June, 2. July, 3. August, 4. September), hence the ordering of the numbers (but not their actual values) can be taken into account in the subsequent statistical analysis. Circular ordering of categories (e.g., directions N, NE, E, SE, S, SW, W, NW) is a very special case, as are angular data in the continuous case, where 360° is identical to 0°.

Continuous data can be measured on a *ratio* or *interval* scale. A continuous scale is classified as ratio when two numbers on the scale are compared *multiplicatively*, and an interval scale is when they are compared *additively*. For example, age – in fact, any variable measuring time – is an interval variable. We would not say that Michael's age increased by 0.000002% (the multiplicative increase) in the time it took him to write that sentence above, but we would simply say that it increased by 44 seconds (the additive increase). Ratio variables are almost always nonnegative and have a fixed zero value: for example, biomass, concentration, length, euros and tonnage. Temperature, even though it does have an absolute zero, is an interval variable, unless you like to say that today is 2.6% hotter than yesterday (with respect to absolute zero) – we prefer to say that the temperature today has risen by 7 °C compared to yesterday's 20 °C.

Count data have a special place in the scheme of Exhibit 3.1, as they can be considered both ordinal and ratio. When 23 individuals of *Galothowenia oculata* are counted in a marine benthic sample, is that a continuous variable? We could not have counted a fraction of an individual, so this sounds like an ordinal categorical observation, but with many possible categories. On the other hand, in a survey of family sizes in Europe, we find only a few values – 0, 1, 2, 3 and 4 children and a sprinkling of families with 5 or more. This sounds more ordinal categorical and less continuous than the *Galothowenia oculata* count. The truth is that they can be validly considered one or the other, depending on how many possible values are observable. If there are many possible values, as in the case of species abundance, then we tend to think of it as a ratio variable. Another aspect is whether we model the expected, or average, count, which is theoretically continuous: for example, at a given sampling location we might predict an average *Galothowenia oculata* abundance of 10.57, even though individual counts are, of course, integers.

Finally, we have singled out compositional data as a special case – these are proportions that add up to 1, a property called *closure*, or the *unit-sum constraint*. The compositional label applies to a set of variables, not to a single one, since it is the property of the set that gives it that nature. Compositional data are usually created from a set of counts or a set of ratio variables when their total is not as relevant as the composition formed by the parts. For example, when we count different species sampled at a particular site, it is likely that the total number is not so relevant, but rather the proportion that each species contributes to the overall count. But if the sampling sites were exactly the same size, as in quadrat sampling in botany, then the overall counts would also be valid measures of overall abundance per unit area sampled. By contrast, a geochemist looking at a mineral sample is not concerned about the weight or volume of the particular sample but in the breakdown of that sample into its components. The situation is identical for fatty acid studies in biology where the data are inherently proportions or percentages, with the overall size of the material sampled having no relevance at all.

The myth of the normal distribution

One of the thorniest issues for applied researchers is that of the normal distribution – most would think that their data should be normal or close to normal in order to arrive at valid conclusions subsequently. This belief is mostly misguided, however, and is a myth created in idealized statistics courses that assume that everything is normally distributed and teach very little about nonparametric statistics, categorical data analysis and modern hypothesis testing using computer-based algorithms such as permutation testing and bootstrapping (see Chapter 17). In any case, it is important to distinguish between *exploratory* and *confirmatory* data analysis. In data exploration, which is actually the theme of most of the present book, we are considering methods to summarize and interpret large data sets, to give us an understanding of the information that we have painstakingly collected and to diagnose relationships between the observed variables. The normal distribution is a minor issue here, but outliers and standardization and transformations are major ones, which we deal with soon. In the second case of confirmatory analysis, which we will touch on now and again in passing, data are assumed to be representative of a wider population and we want to make conclusions, called *inferences*, about that population. An example of an inference might be that a particular underlying gradient detected in the sample exists in the population with a high degree of probability, based on statistical hypothesis testing. Here we need to know the probabilistic characteristics of the population, and the assumption of normality is the easiest (and most studied) choice. There are, however, other solutions which do not depend on this assumption at all. But, having said this, the idea of data being approximately normally distributed, or at least symmetrically distributed, does have some advantages in exploratory analysis too.

Most of the methods we use are what we call *least-squares* methods that were developed in the context of well-behaved normally distributed data. By "least squares" we mean that solutions are found by minimizing an error criterion defined as a sum of squared differences between our estimated (or "fitted") solution and the observed data. Even in our simple data set of Chapter 1 (Exhibit 1.1) we have seen that the variables are generally not symmetrically distributed around their means. The count variables in Exhibit 1.3, for example, show very skew distributions, with mostly low values and a few much higher ones. Data analysis with these variables, using standard least-squares procedures to fit the models, will be sensitive to the higher values, where the larger error in fitting the high values is even larger when squared. There are several solutions to this problem: one is to use a different theory – for example, maximum likelihood rather than least squares – or make some transformation of the data to make the distributions more symmetric and closer to "well-behaved" normal. Another possibility, used often in the case of count data, is to introduce weights into the analysis, where rare or unusual values are downweighted and contribute less to the results (for example, see Chapters 13 and 14 on correspondence analysis and log-ratio analysis).

Logarithmic transformation of ratio data

Since most ratio variables are skew with long tails to the right, a very good all-purpose transformation is the logarithmic one. This not only pulls in the long tails but also converts multiplicative relationships to additive ones, since $\log(ab) = \log(a) + \log(b)$ – this is advantageous not only for interpretation but also because most of the methods we use involve addition and subtraction. The logarithmic function is shown in Exhibit 3.2 (the lowest curve) as well as other

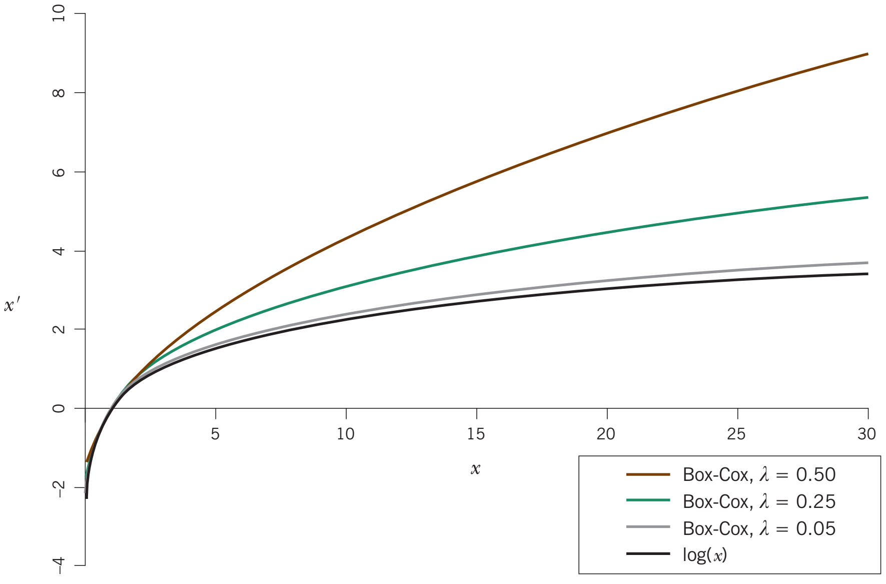

Exhibit 3.2: *The natural logarithmic transformation x' = log(x) and a few Box-Cox power transformations, for powers λ = ½ (square root), ¼ (double square root, or fourth root) and 0.05*

functions that will be described in the next section. Notice how large values of the original variable *a* are pulled down by the log-transformation.

To illustrate the effect the log-transformation has on the interpretation of a variable, consider first the simple linear additive relationship expressed in this equation between the average abundance of a certain marine species and the concentration of the heavy metal barium:

$$abundance = \mathrm{C} - 0.023\ Ba \tag{3.1}$$

where C is some constant. The interpretation is that abundance decreases on average by 0.023 per unit increase of barium (measured in ppm), or 2.3 per 100 units increase in barium. Now consider another equation where abundance has been log-transformed using the natural logarithm (sometimes denoted by “ln”):

$$\log(abundance) = \mathrm{C}' - 0.0017\ Ba \tag{3.2}$$

where C′ is another constant. A unit increase in *Ba* now decreases the logarithm of *abundance* on average by 0.0017. If we exponentiate both sides of equation (3.2), which is the inverse transformation of the natural logarithm, we obtain:

$$abundance = e^{(\mathrm{C}' - 0.0017\ Ba)} = e^{\mathrm{C}'} \cdot e^{(-0.0017\ Ba)} \tag{3.3}$$

That is, a unit increase in barium changes $\exp(-0.0017\,Ba)$ to $\exp(-0.0017\,[Ba+1]) = \exp(-0.0017\,Ba)\cdot\exp(-0.0017)$. So the effect is that abundance is multiplied by $\exp(-0.0017) = 0.9983$, in other words a 0.17% decrease. For a 100 unit increase in barium, abundance is multiplied by $\exp(-0.0017 \times 100) = \exp(-0.17) = 0.8437$, a 15.63% decrease. Notice that this is not a $100 \times 0.17\% = 17\%$ decrease since the multiplicative effect is compounded (just like interest calculations in finance where the "capital" is being diminished). The above example shows how the logarithmic transformation converts an additive effect into a multiplicative one.

Power transformations and Box-Cox

In Exhibit 3.2 three other curves are shown corresponding to power transformations of the variable x, called *Box-Cox transformations* after two of the most influential statisticians of the 20th century, the American George Box and the Englishman Sir David Cox. These are a slight modification of a simple power transformation x^λ and take the following form:

$$x' = \frac{1}{\lambda}(x^\lambda - 1) \qquad (3.4)$$

The advantage of this form is that it tends to the log-transformation as the power parameter tends to 0, as shown in Exhibit 3.2 – as λ decreases the curve approaches the logarithmic curve. The division by λ conveniently keeps the scale of the original variable from collapsing: for example, if you take the 20th roots (that is, $x^{0.05}$) of a set of data, you will quickly see that all the values are close to 1, so the division by 0.05, which multiplies the values by 20, restores them to an almost logarithmic scale.

Box-Cox transformations serve as a flexible way of symmetrizing data and have found extensive application in regression analysis. The inverse transformation is:

$$x = (1 + \lambda x')^{\frac{1}{\lambda}} \qquad (3.5)$$

where x' is the transformed value in (3.4). We shall refer to these transformations in Chapter 14 in our treatment of compositional data.

Dummy variables

In functional methods of regression and classification, there is no problem at all to have some continuous and some categorical predictors. The categorical variables are coded as *dummy variables*, which are variables that take on the values 0 or 1. For example, suppose one of the predictors is sampling region, with four regions. This variable is coded as four dummy variables which have values [1 0 0 0] for region A, [0 1 0 0] for region B, [0 0 1 0] for region C and

[0 0 0 1] for region D. For a technical reason only 3 out of these 4 dummies can be used – the statistical program usually does this all automatically, omitting (for example) the last dummy for region D. Then the results for the three included dummies are interpreted as the differences of those three regions compared to the omitted region. If the categorical variable has only two categories, for example pelagic or littoral habitat, then only one dummy variable is included, omitting the one for littoral, for example, in which case the model estimates the effect of the difference between pelagic and littoral habitats.

For structural methods, however, the situation is more complicated, because we are trying to explore structure amongst all the variables and here the coding does matter. We could resort to dummy variable coding of all the categorical variables but this is not satisfactory because of the inherently different variances in the dummy variables compared to the continuous ones. For example, a danger might exist that the dummy variables have much less variance than the continuous variables, so when we look for structure we only see patterns in the continuous variables while those in the categorical variables are more or less "invisible" to our investigation. We need to balance the contributions of the variables in some way that gives them all a fair chance of competing for our attention. This is a problem of *standardization*, which we treat in detail in a later section.

Fuzzy coding

An alternative approach to the problem of mixed-scale data is to recode the continuous variables also as dummy variables, so that we put them on the same scale as the categorical dummies. This can be achieved by dividing up the continuous scale into intervals, for example three intervals which can be labelled "low", "medium" and "high". Clearly, this loses a lot of information in the continuous variables, so there is a way to avoid data loss called *fuzzy coding*. If we again choose the three-category option, then a continuous variable can be fuzzy coded as shown in Exhibit 3.3.

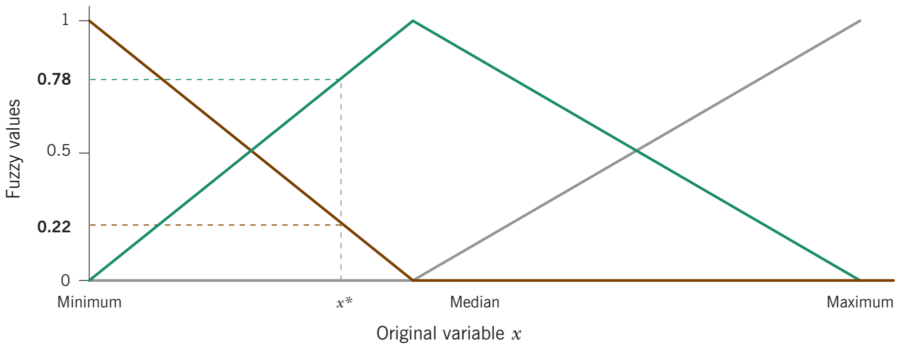

Exhibit 3.3:
Fuzzy coding of a continuous variable x into three categories, using triangular membership functions. The minimum, median and maximum are used as hinge points. An example is given of a value x^* just below the median being fuzzy coded as [0.22 0.78 0]

In this example we have used the simplest procedure, *triangular membership functions,* for fuzzy coding. For three categories we need three *hinge points*, which we have chosen to be the minimum, median and maximum of the continuous variable x. Triangles are drawn as shown and these provide the fuzzy values for the three categories – notice that the third category, drawn in gray, has value zero below the median. The general algorithm for computing the three fuzzy values $[z_1 \;\; z_2 \;\; z_3]$ is as follows:

$$z_1(x) = \begin{cases} \dfrac{m_2 - x}{m_2 - m_1}, & \text{for } x \le m_2 \\ 0 & \text{otherwise} \end{cases}$$

$$z_2(x) = \begin{cases} \dfrac{x - m_1}{m_2 - m_1}, & \text{for } x \le m_2 \\ \dfrac{m_3 - x}{m_3 - m_2}, & \text{for } x > m_2 \end{cases} \tag{3.6}$$

$$z_3(x) = \begin{cases} \dfrac{x - m_2}{m_3 - m_2}, & \text{for } x > m_2 \\ 0 & \text{otherwise} \end{cases}$$

where m_1, m_2 and m_3 denote the three hinge points. For example, in Exhibit 3.3, the hinges were $m_1 = 3.69$, $m_2 = 8.64$ and $m_3 = 19.65$. The value x^* was 7.55 and was fuzzy coded as $z_1(7.55) = (8.64 - 7.55) / (8.64 - 3.69) = 0.22$; $z_2(7.55) = (7.55 - 3.69) / (8.64 - 3.69) = 0.78$, and $z_3(7.55) = 0$.

The advantage of this coding is that it is invertible – we can recover the original value from the fuzzy values as a linear combination of the hinge values (in fact, a weighted average since the fuzzy values add up to 1):

$$x = z_1 \, m_1 + z_2 \, m_2 + z_3 \, m_3 \tag{3.7}$$

for example, $0.22 \times 3.69 + 0.78 \times 8.64 + 0 \times 19.65 = 7.55$. This reverse process of going from the fuzzy values back to the original data is called *defuzzification.* The fact that the fuzzy coding is reversible means that we have conserved all the information in the coded values, while gaining the advantage of converting the continuous variable to a form similar to the categorical dummies. However, there is still a problem of balancing the variances, which we now discuss.

Standardization

Standardization is an important issue in structural methods of multivariate analysis. Variables on different scales have natural variances which depend mostly on their scales. For example, suppose we measure the length of a dorsal fin of a sample of fish in centimeters – the variance of our measurements across the sample might be 0.503 cm^2, and the standard deviation 0.709 cm (the square root of the variance). Then we decide to express the lengths in millimeters, because most of the other measurements are in millimeters; so the variance is now 50.3 mm^2, a hundred times the previous value, while the standard deviation is 7.09 mm, ten times more. The fact that some variables can have high variances just because of the chosen scale of measurement causes problems when we look for structure amongst the variables. The variables with high variance will dominate our search because they appear to contain more information, while those with low variance are swamped because of their small differences between values.

The answer is clearly to balance out the variances so that each variable can play an equal role in our analysis – this is exactly what standardization tries to achieve. The simplest form of standardization is to make all variances in the data set exactly the same. For a bunch of continuous variables, for example, we would divide the values of each variable by its corresponding sample standard deviation so that each variable has variance (and also standard deviation) equal to 1. Often this is accompanied by *centering* the variable as well, that is, subtracting its mean, in which case we often refer to the standardized variable as a *Z-score*. This terminology originates in the standardization of a normally distributed variable X, which after subtracting its mean and dividing by its standard deviation is customarily denoted by the letter Z and called a *standard normal variable,* with mean 0 and variance 1.

Standardization can also be thought of as a form of weighting. That is, by dividing variables with large variances by their large standard deviations, we are actually multiplying them by small numbers and reducing their weight. The variables with small variances, on the other hand, are divided by smaller standard deviations and thus have their weight increased relative to the others.

Other forms of standardization are:

- by the range: each variable is linearly transformed to lie between 0 and 1, where 0 is its minimum and 1 its maximum value;

- by chosen percentiles: because the range is sensitive to outliers, we can "peg" the 0 and 1 values of the linearly transformed variable to, say, the 5th and 95th percentile of the sample distribution;

- by the mean: the values of a variable are divided by their mean, so that they have standard deviations equal to what is called their *coefficient of variation*.

There are various forms of standardization which rely on the assumed theoretical characteristics of the variable. For example, count data are often assumed to come from a *Poisson distribution*. This distribution has the property that the variance is theoretically equal to the mean. Thus, dividing by the square root of the mean would be like dividing by the standard deviation (this is, in fact, the standardization inherent in correspondence analysis – see Chapter 13). Another theoretical result is that, while a Poisson variable has variance that increases as the average count increases, its square root has a variance tending to a constant value of ¼. Hence, an alternative form of standardization that is regularly used to "stabilize" the variance of count data is simply to square root transform them.

Finally, coming back to the handling of continuous and categorical variables jointly, where the continuous variables have been coded into fuzzy dummy variables and the categorical variables into "crisp" (zero-one) dummies, we could standardize by calculating the collective variance of each set of dummies corresponding to one variable and then weighting the set accordingly. That is, we do not standardize individual dummy variables, which would be incorrect, but each group as a whole.

SUMMARY: Measurement scales, transformation and standardization

1. Variables can be either categorical or continuous, although all measurements are categorical in the sense of being discretized. Continuous variables are those that have very many categories, for example a count variable, or are discretized versions of a variable which could, at least theoretically, be measured on a continuous scale, for example a length or a concentration.

2. Categorical variables can be either ordinal or nominal, depending on whether the categories have an inherent ordering or not.

3. Continuous variables can be either ratio or interval, depending on whether we compare two observations on that variable multiplicatively (as a ratio) or additively (as a difference).

4. The logarithmic transformation is a very useful transformation for most positive ratio measurements, because multiplicative comparisons are converted to additive ones and because high values are pulled in, making the distribution of the variable more symmetric.

5. Box-Cox transformations are a very flexible class of power transformations which include the log-transformation as a limiting case.

6. Categorical variables are usually coded as dummy variables in order to be able to judge the effect or relationship of individual categories.

7. Continuous variables can also be dummy coded but this loses a lot of information. A better option is to fuzzy code them into a small number of categories,

which allows continuous variables to be analysed together with categorical ones more easily, especially in the case of structural multivariate methods.

8. In structural methods standardization is a major issue for consideration. Variances of the variables being analysed need to be balanced in some way that gives each variable a fair chance of being involved in the determination of the latent structure. Results should not depend on the scale of measurement. Standardization is not an issue for functional methods because the effect of a variable on a response is measured independently of the scale.

List of Exhibits