Lab 1. Ecological Data

Smit, A. J.
University of the Western Cape

2022-08-01

Table of contents

BCB743

This material must be reviewed by BCB743 students in Week 1 of Quantitative Ecology.

This Lab Accompanies the Following Lecture

- Lecture 3: Ecological Gradients
- Lecture 4: Biodiversity Concepts

Data For This Lab

The Doubs River (Verneaux 1973, Borcard et al. 2011) and toy data are at the links below:

- The environmental data – `DoubsEnv.csv`
- The species data – `DoubsSpe.csv`
- The spatial data – `DoubsSpa.csv`
- Example xyz data – `Euclidean_distance_demo_data_xyz.csv`

Stuff

> *"A scientific man ought to have no wishes, no affections, – a mere heart of stone."*
>
> — Charles Darwin

1 About Macroecology

This course is about community ecology across different spatial and temporal scales. Community ecology underpins the vast fields of biodiversity and biogeography and concerns spatial scales from square meters to all of Earth. We can look at historical, contemporary, and future processes implicated in shaping the distribution of life on our planet.

Ecologists tend to analyse how multiple environmental factors act as drivers that influence the distribution of tens or hundreds of species. These data often are messy and statistical considerations need to be understood within the context of the available data.

Up to 20 years ago, ecologists focused on populations (the dynamics of individuals of one species interacting among each other and with their environment) and communities (collections of multiple populations, how they interact with each other and their environment, and how this affects the structure and dynamics of ecosystems). This is a modern development of ecology. But ecologists have expanded their horizon regarding the questions they now seek answers for. Today, **macroecology** offers a broadened view of ecology. Macroecologists seek to find the geographical patterns and processes in biodiversity across all spatial scales, from local to global, across time scales from years to millennia, and across all taxonomic hierarchies (from genetic variability within species up to major higher-level taxa, such as families and orders). It attempts to arrive at a unifying theory for ecology across all of these

scales — e.g., one that can explain all patterns in structure and functioning from microbes to blue whales. Perhaps most importantly, it attempts to offer mechanistic explanations for these patterns. At the heart of all ecological answers are also deep insights stemming from understanding evolution (facilitated by the growth of phylogenetic datasets — see below).

On a basic data analytical level, population ecology, community ecology, and macroecology all share the same approach regarding the underlying data. We start with data representing the species and the associated environmental conditions at a selection of sites (called **species tables** and **environmental tables**). The species tables are then converted to **dissimilarity matrices** and the environmental tables to **distance matrices**. From here, basic analyses can offer insights into how biodiversity is structured, e.g., **species-abundance distributions**, **occupancy-abundance curves**, **species-area curves**, **distance decay curves**, and **gradient analyses** (Shade et al. 2018). In the Labs, we will explore some of these properties.

2 Ecological Data

2.1 Properties of Ecological Datasets

Ecological data capture properties of the environment and properties of communities. They are typically stored as separate datasets, but they are analysed together.

These data sets are usually arranged in a **matrix**. In the case of community composition, a matrix has **species (or higher level taxa whose resolution depends on the research question) arranged down columns** and **samples (typically the sites, stations, transects, time, plots, etc.) along rows**. We call this a **sites × species table**. In the case of environmental data, a matrix is a **site × environment table**. The term 'sample' denotes the basic unit of observation. Samples on a map may be quadrats, transects, stations, locations, traps, seine net tows, trawls, grids cells, etc. It is essential to be unambiguous about the basic unit of the samples.

2.2 The Doubs River Data

An obvious example of environmental and species datasets is the Doubs River dataset. Please refer to David Zelený's website for an explanation of these data. The primary publication outlining this study is Verneaux (1973), and an example analysis is provided by Borcard et al. (2011). These data demonstrate how one of the basic mechanisms of biodiversity patterning — gradients — can be seen operating in a real-world case study. It offers keen insight also into the properties of species and environmental tables and the dissimilarity and distance matrices derived from them.

2.3 Looking at the Files' Content

These data are available in CSV format, but we can open and view it in MS Excel. 'CSV' means *comma separated value*. It is a plain text file that can be edited in any text editor (such

as Notepad on MS Windows, or VS Code, VIM, emacs, etc. on all platforms). Figure 1 shows what a CSV file looks like in a plain text editor, VS Code, on my computer. Once imported, it will look similar to the one seen in Figure 3.

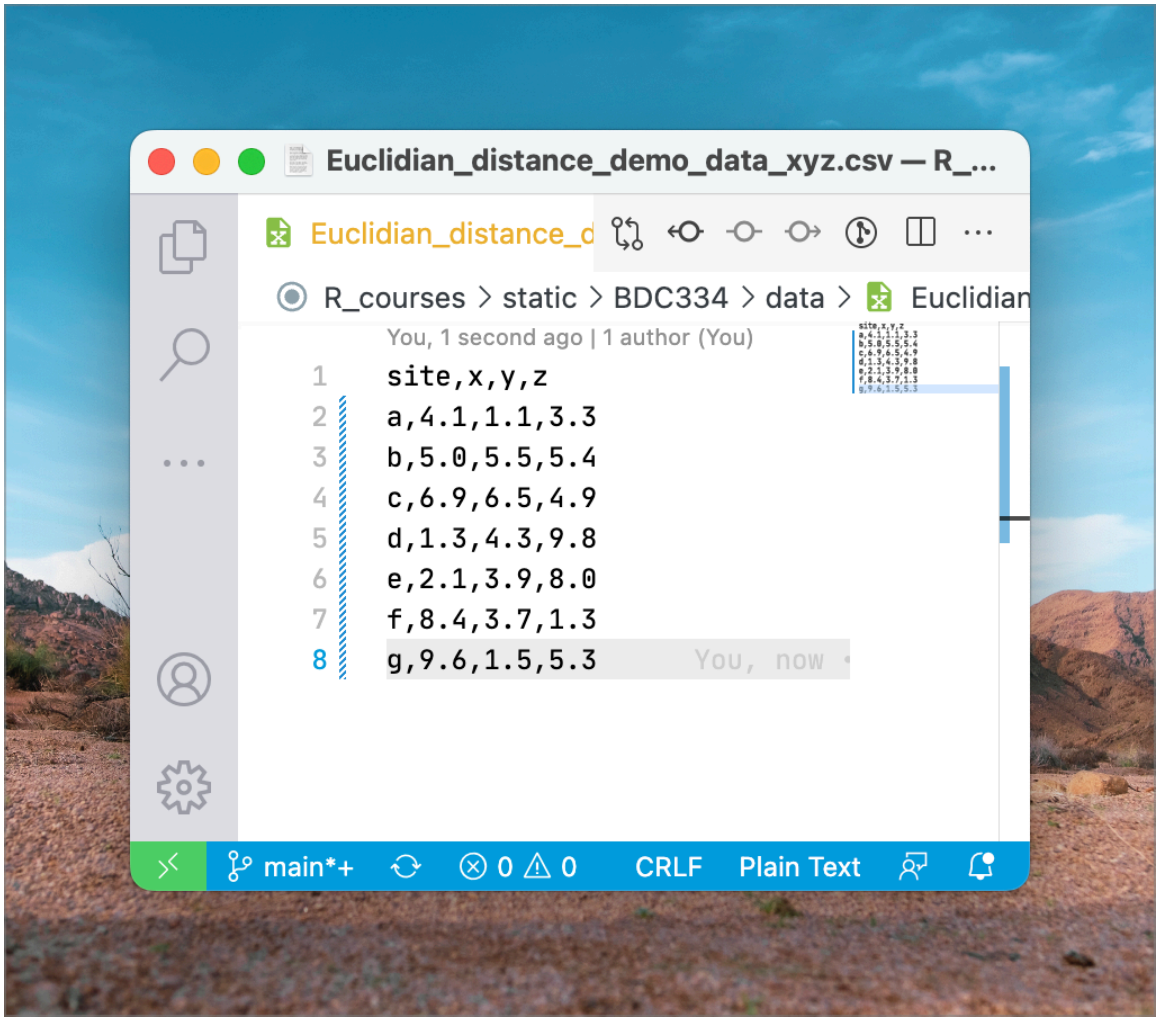


Figure 1: View of a CSV file inside VS Code.

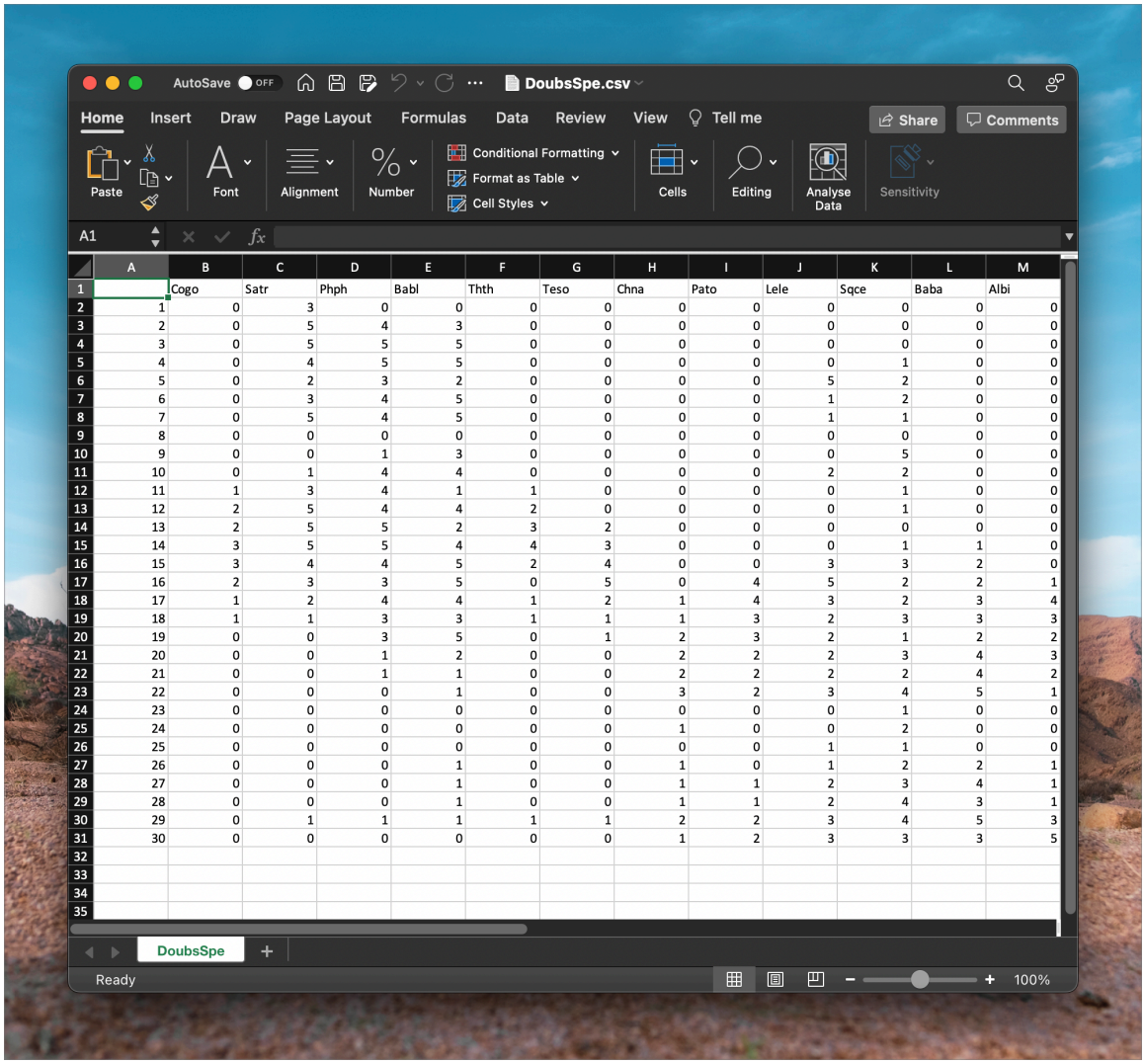
i Note About CSV Files and MS Excel

CSV is a standard format used in the scientific disciplines as it is compatible with many software. Globally, scientists use a period '.' as a decimal point separator. You can see this in the file above. Commas are used exclusively as field separators (you'll see separate columns once opened in MS Excel).

CSV files create a bit of a problem for South Africans, who are indoctrinated from a young age to use commas as a decimal point separators — this is to conform with the regional (South African) expectation that dictates commas be used as decimals. So, when you import a CSV file for the first time, you'll likely see gibberish because your computer will probably be set up to honour the regional (locale) the expectation of commas as decimal points (and 'R' for currency, metric units of measurements, etc.). So, you need to know how to fix this to prevent upsetting me (it is a pet peeve and frustrates me endlessly) and yourselves.

Fixing this annoyance is not too tricky, as is demonstrated here. Follow the instruction under '**Changing commas to decimals and vice versa by changing Excel Options**'. Better still, change the global system settings, as the same article explains. Do this before importing the CSV file.

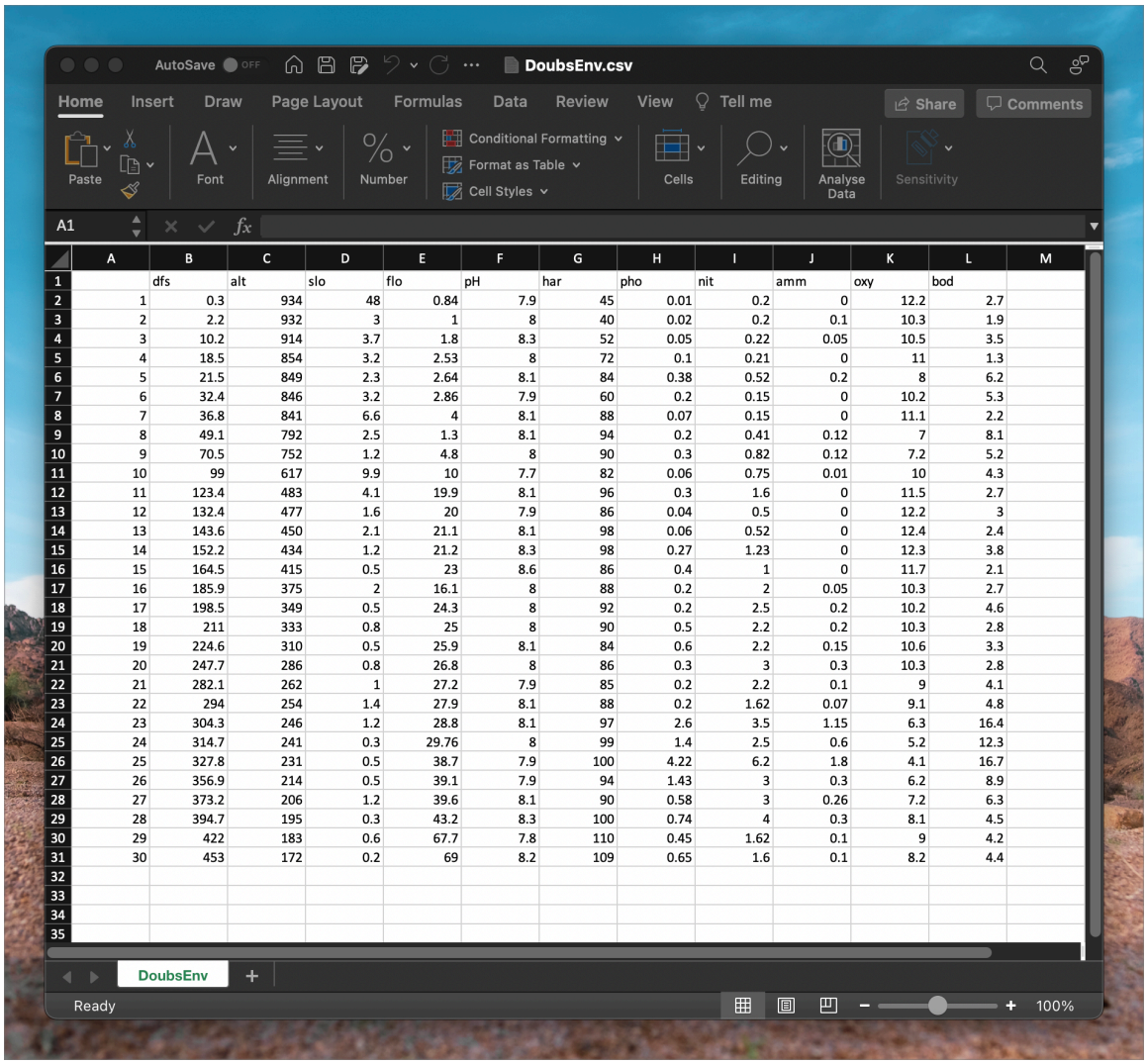
After importing the Doubs River data, we see something that resembles the following two figures. First, in `DoubsSpe.csv`, we see the table (or spreadsheet) view of the species data. The species codes for 27 species of fish appear as column headers (not all species' data are visible as the data are truncated to the right) and in rows 2 through 31 (30 rows) are each of the samples — in this case, there is one sample per site down the length of the river (Figure 2).



	A	B	C	D	E	F	G	H	I	J	K	L	M
1		Cogo	Satr	Phph	Babl	Thth	Teso	Chna	Pato	Lele	Sqce	Baba	Albi
2	1	0	3	0	0	0	0	0	0	0	0	0	0
3	2	0	5	4	3	0	0	0	0	0	0	0	0
4	3	0	5	5	5	0	0	0	0	0	0	0	0
5	4	0	4	5	5	0	0	0	0	0	1	0	0
6	5	0	2	3	2	0	0	0	0	5	2	0	0
7	6	0	3	4	5	0	0	0	0	1	2	0	0
8	7	0	5	4	5	0	0	0	0	1	1	0	0
9	8	0	0	0	0	0	0	0	0	0	0	0	0
10	9	0	0	1	3	0	0	0	0	0	5	0	0
11	10	0	1	4	4	0	0	0	0	2	2	0	0
12	11	1	3	4	1	1	0	0	0	0	1	0	0
13	12	2	5	4	4	2	0	0	0	0	1	0	0
14	13	2	5	5	2	3	2	0	0	0	0	0	0
15	14	3	5	5	4	4	3	0	0	0	1	1	0
16	15	3	4	4	5	2	4	0	0	3	3	2	0
17	16	2	3	3	5	0	5	0	4	5	2	2	1
18	17	1	2	4	4	1	2	1	4	3	2	3	4
19	18	1	1	3	3	1	1	1	3	2	3	3	3
20	19	0	0	3	5	0	1	2	3	2	1	2	2
21	20	0	0	1	2	0	0	2	2	2	3	4	3
22	21	0	0	1	1	0	0	2	2	2	2	4	2
23	22	0	0	0	1	0	0	3	2	3	4	5	1
24	23	0	0	0	0	0	0	0	0	0	1	0	0
25	24	0	0	0	0	0	0	1	0	0	2	0	0
26	25	0	0	0	0	0	0	0	0	1	1	0	0
27	26	0	0	0	1	0	0	1	0	1	2	2	1
28	27	0	0	0	1	0	0	1	1	2	3	4	1
29	28	0	0	0	1	0	0	1	1	2	4	3	1
30	29	0	1	1	1	1	1	2	2	3	4	5	3
31	30	0	0	0	0	0	0	1	2	3	3	3	5
32													
33													
34													
35													

Figure 2: The Doubs River species data seen in MS Excel.

`DoubsEnv.csv` contains the environmental data, as seen in the following figure. The names of the 11 environmental variables appear as column headers, and there are 30 rows, one for each of the samples — the samples match that of the species data (Figure 3).



	A	B	C	D	E	F	G	H	I	J	K	L	M
1		dfs	alt	slo	flo	pH	har	pho	nit	amm	oxy	bod	
2	1	0.3	934	48	0.84	7.9	45	0.01	0.2	0	12.2	2.7	
3	2	2.2	932	3	1	8	40	0.02	0.2	0.1	10.3	1.9	
4	3	10.2	914	3.7	1.8	8.3	52	0.05	0.22	0.05	10.5	3.5	
5	4	18.5	854	3.2	2.53	8	72	0.1	0.21	0	11	1.3	
6	5	21.5	849	2.3	2.64	8.1	84	0.38	0.52	0.2	8	6.2	
7	6	32.4	846	3.2	2.86	7.9	60	0.2	0.15	0	10.2	5.3	
8	7	36.8	841	6.6	4	8.1	88	0.07	0.15	0	11.1	2.2	
9	8	49.1	792	2.5	1.3	8.1	94	0.2	0.41	0.12	7	8.1	
10	9	70.5	752	1.2	4.8	8	90	0.3	0.82	0.12	7.2	5.2	
11	10	99	617	9.9	10	7.7	82	0.06	0.75	0.01	10	4.3	
12	11	123.4	483	4.1	19.9	8.1	96	0.3	1.6	0	11.5	2.7	
13	12	132.4	477	1.6	20	7.9	86	0.04	0.5	0	12.2	3	
14	13	143.6	450	2.1	21.1	8.1	98	0.06	0.52	0	12.4	2.4	
15	14	152.2	434	1.2	21.2	8.3	98	0.27	1.23	0	12.3	3.8	
16	15	164.5	415	0.5	23	8.6	86	0.4	1	0	11.7	2.1	
17	16	185.9	375	2	16.1	8	88	0.2	2	0.05	10.3	2.7	
18	17	198.5	349	0.5	24.3	8	92	0.2	2.5	0.2	10.2	4.6	
19	18	211	333	0.8	25	8	90	0.5	2.2	0.2	10.3	2.8	
20	19	224.6	310	0.5	25.9	8.1	84	0.6	2.2	0.15	10.6	3.3	
21	20	247.7	286	0.8	26.8	8	86	0.3	3	0.3	10.3	2.8	
22	21	282.1	262	1	27.2	7.9	85	0.2	2.2	0.1	9	4.1	
23	22	294	254	1.4	27.9	8.1	88	0.2	1.62	0.07	9.1	4.8	
24	23	304.3	246	1.2	28.8	8.1	97	2.6	3.5	1.15	6.3	16.4	
25	24	314.7	241	0.3	29.76	8	99	1.4	2.5	0.6	5.2	12.3	
26	25	327.8	231	0.5	38.7	7.9	100	4.22	6.2	1.8	4.1	16.7	
27	26	356.9	214	0.5	39.1	7.9	94	1.43	3	0.3	6.2	8.9	
28	27	373.2	206	1.2	39.6	8.1	90	0.58	3	0.26	7.2	6.3	
29	28	394.7	195	0.3	43.2	8.3	100	0.74	4	0.3	8.1	4.5	
30	29	422	183	0.6	67.7	7.8	110	0.45	1.62	0.1	9	4.2	
31	30	453	172	0.2	69	8.2	109	0.65	1.6	0.1	8.2	4.4	
32													
33													
34													
35													

Figure 3: The Doubs River environmental data in MS Excel.

Species data may be recorded as various kinds of measurements, such as presence/absence data, biomass, frequency, or abundance. 'Presence/absence' of species simply tells us the species is there or is not there. It is binary. 'Abundance' generally refers to the the number of individuals per unit of area, volume. 'Per cent cover' refers to the proportion of a covered by a species. Per cent cover is used for vegetation, some encrusting species of animals (e.g., sponges), or organisms such as oysters or mussels that can be too numerous to count but whose abundance can be estimated as filling a portion of a sampling unit such as a quadrat. 'Biomass' refers to the species' mass per unit of area or volume. The type of measure will depend on the taxa and the questions under consideration. The critical thing to note is that all species have to be homogeneous in terms of the metric used to quantify them (i.e., all of it as presence/absence, or abundance, or biomass, not mixtures of them). The matrix's row vectors are the species composition for the corresponding sample. That is to say, a row

runs across multiple columns, which tells us that the sample is comprised of all the species whose names are given by the column titles. Note that in the case of the data in the above figures, it is often the case that there are 0s, meaning that not all species are present at all sites. Species composition is frequently expressed in relative abundance, i.e. constrained to a constant total such as 1 or 100%, or biomass, where the upper limit might be arbitrary.

The environmental data may be heterogeneous, i.e. the units of measure may differ among the variables. For example, pH has no units, the concentration of some nutrients has a unit of (typically) μM, elevation may be in meters, etc. Because these units have different magnitudes and ranges, we may need to standardise them. To standardise data, we subtract the mean of each column from each data point in the column and then divide each of the resultant values by the standard deviation of the columns.

! Lab 1

(To be reviewed by BCB743 student but not for marks)

- 1.a) Calculate the mean and SD for each variable (column) of the "raw" data. Explain.
- 1.b) Standardise the Doubs River environmental data in MS Excel.
- 1.c) Calculate the mean and SD for each standardised variable (column). Explain.

2.4 Properties of Species Datasets

Many community data matrices share some general characteristics:

- Most species occur only infrequently. The majority of species might typically be represented at only a **few locations** (where they might be pretty abundant). Or some species are simply **rare** in the sampled region (i.e. when they are present, they are present at a very low abundance). This results in **sparse matrices** where the bulk of the entries consists of zeros.
- Ecologists tend to sample a multitude of factors that they think influence species composition, so the matching environmental data set will also have multiple (10s) columns that will be assessed in various hypotheses about the drivers of species patterning across the landscape. For example, fynbos biomass may be influenced by the fire regime, elevation, aspect, soil moisture, soil chemistry, edaphic features, etc. These datasets are called **multi-dimensional** matrices, with the 'dimensions' referring to the many species or environmental variables.
- Even though we may capture a multitude of information about many environmental factors, **the number of important ones is generally relatively low** — i.e. a few factors can explain the majority of the explainable variation, and it is our intention to find out which of them is most important.

- Much of the signal may be spurious, i.e. the matrices have **high noise**. Variability is a general characteristic of the data, which may result in emerging false patterns. This is because sampling may capture a considerable amount of stochasticity that may mask the actual pattern of interest. Imaginative and creative sampling may reveal some of the ecological patterns we are after, but this requires long years of experience and is not something that can easily be taught as part of our module.
- There is a significant amount of **collinearity**. This means that many correlated explanatory variables can explain patterning, but only a few act in a way that implies causation. Collinearity is something we will return to later on.

3 Ecological Gradients

Although there are many ways in which species can respond to their environment, one of the most striking responses can be seen along with environmental gradients. Next, we will explore this concept by discussing coenoclines and unimodal species distribution models.

3.1 The Unimodal Model

The **unimodal** model is an idealised species response curve (visualised as a coenocline) where a species has only one mode of abundance. In this species response curve, the species has one optimal environmental condition where it is most abundant (the fewest ecophysiological and ecological stressors). If any aspect of the environment is suboptimal (greater or lesser than the optimum), the species will perform more poorly and have a lower abundance. The unimodal model offers a convenient heuristic tool for understanding how species can become structured along environmental gradients.

3.2 Coenoclines, Coenoplanes, and Coenospaces

A **coenocline** is a graphical display of *all species* response curves (see definition below) *simultaneously* along one environmental gradient. This is a useful way to display the arrangement of species' *fundamental niches* along gradients. It aids our understanding of the species response curve if we imagine the gradient operating in only one geographical direction. The **coenoplane** concept extends the coenocline to cover two gradients. Again, our visual representation can be facilitated if the two gradients are visualised orthogonal (in this case, at right angles) to each other (e.g., east-west and north-south) and do not interact. A **coenospace** complicates the model substantially, as it can allow for an unspecified number of gradients to operate simultaneously on multiple species simultaneously. It will probably also capture interactions of environmental drivers on the species.

```
library(coenocliner)
set.seed(2)
M <- 20                                 # number of species
```

```
ming <- 3.5                                  # gradient minimum...
maxg <- 7                                    # ...and maximum
locs <- seq(ming, maxg, length = 100)        # gradient locations
opt  <- runif(M, min = ming, max = maxg)     # species optima
tol  <- rep(0.25, M)                         # species tolerances
h    <- ceiling(rlnorm(M, meanlog = 3))      # max abundances
pars <- cbind(opt = opt, tol = tol, h = h) # put in a matrix

mu <- coenocline(locs, responseModel = "gaussian", params = pars,
                 expectation = TRUE)

matplot(locs, mu, lty = "solid", type = "l", xlab = "pH", ylab =
"Abundance")
```

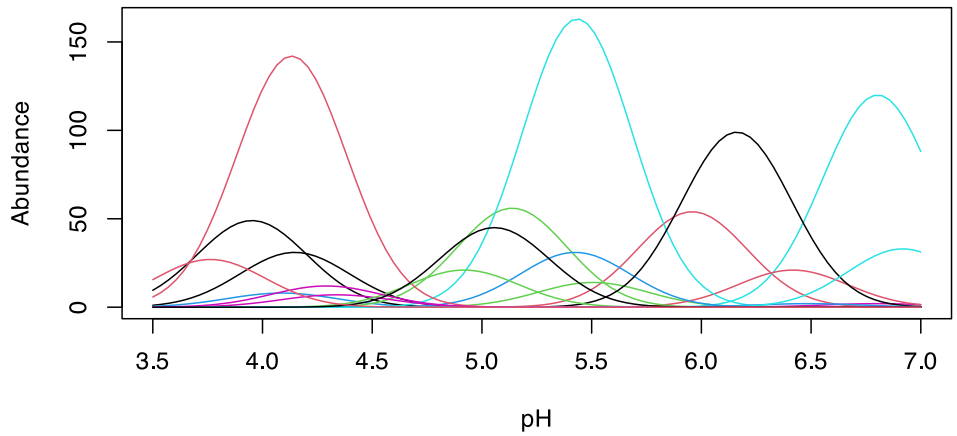


Figure 4: A coenocline.

Above is an example of a coenocline using simulated species data. It demonstrates an important idea: that of unimodal species distributions (Figure 4).

```
set.seed(10)
N <- 30                                      # number of samples
M <- 20                                      # number of species
## First gradient
ming1 <- 3.5                                 # 1st gradient minimum...
maxg1 <- 7                                   # ...and maximum
loc1 <- seq(ming1, maxg1, length = N)        # 1st gradient locations
opt1 <- runif(M, min = ming1, max = maxg1)   # species optima
```

```r
tol1 <- rep(0.5, M)                              # species tolerances
h    <- ceiling(rlnorm(M, meanlog = 3))          # max abundances
par1 <- cbind(opt = opt1, tol = tol1, h = h)     # put in a matrix
## Second gradient
ming2 <- 1                                       # 2nd gradient minimum...
maxg2 <- 100                                     # ...and maximum
loc2 <- seq(ming2, maxg2, length = N)            # 2nd gradient locations
opt2 <- runif(M, min = ming2, max = maxg2)       # species optima
tol2 <- ceiling(runif(M, min = 5, max = 50))     # species tolerances
par2 <- cbind(opt = opt2, tol = tol2)            # put in a matrix
## Last steps...
pars <- list(px = par1, py = par2)               # put parameters into a
list
locs <- expand.grid(x = loc1, y = loc2)          # put gradient locations
together

mu2d <- coenocline(locs, responseModel = "gaussian",
                   params = pars, extraParams = list(corr = 0.5),
                   expectation = TRUE)

layout(matrix(1:4, ncol = 2))
op <- par(mar = rep(1, 4))
for (i in c(2,8,13,19)) {
  persp(loc1, loc2, matrix(mu2d[, i], ncol = length(loc2)),
        ticktype = "detailed", zlab = "Abundance",
        theta = 45, phi = 30)
}
```

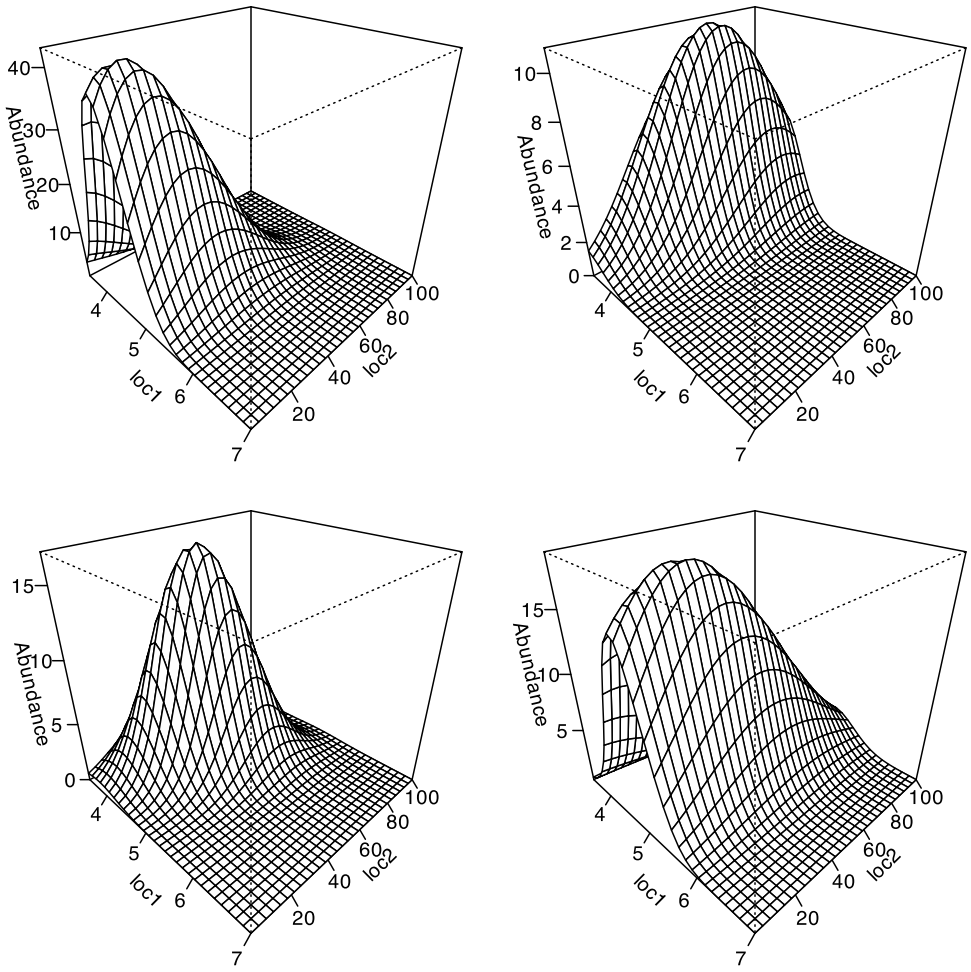


Figure 5: A smoothed coenoplane.

```
sim2d <- coenocline(locs, responseModel = "gaussian",
                    params = pars, extraParams = list(corr = 0.5),
                    countModel = "negbin", countParams = list(alpha =
1))

layout(matrix(1:4, ncol = 2))
op <- par(mar = rep(1, 4))
for (i in c(2,8,13,19)) {
  persp(loc1, loc2, matrix(sim2d[, i], ncol = length(loc2)),
        ticktype = "detailed", zlab = "Abundance",
        theta = 45, phi = 30)
}
```

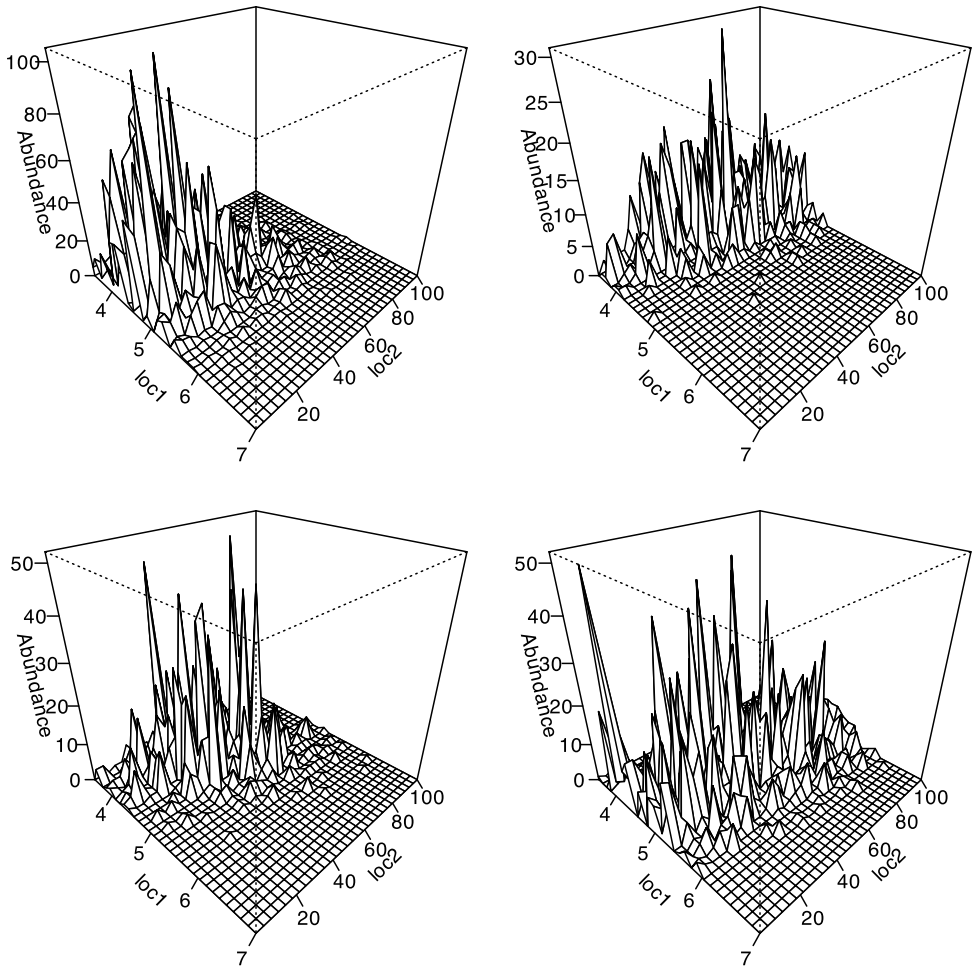


Figure 6: A 'raw' coenoplane.

A coenoplane is demonstrated above (Figure 5). We see idealised surfaces (smooth models), and the 'raw' species counts are obscured. Plotting the actual count data looks messier (Figure 6) because the measured data are not only a reflection of the underlying species response according to the unimodal model (and hence the fundamental niche), but also of the biotic processes that result in the realised niche, and the stochastic processes that generate some 'noise' seen in the data.

3.2.1 Species response curves

Plotting the abundance of a species as a function of position along a the gradient is called a **species response curve**. If a long enough the gradient is sampled, a species typically has a *unimodal* response (one peak *resembling* a Gaussian distribution) to the gradient. Although the idealised Gaussian response is desired (for statistical purposes, largely), in nature, the curve might deviate quite noticeably from what's considered ideal. It is probable that a per-

fectly normal species distribution along a gradient can only be expected when the gradient is perfectly linear in magnitude (seldom true in nature), operates along only one geographical direction (unlikely), and all other potentially additive environmental influences are constant across the ecological (coeno-) space (also not a realistic expectation). Very importantly, also, the species response curve is not a direct measure of the species' fundamental niche, but rather a reflection of the species' realised niche.

4 Exploring the Data

At the start of the analysis, before we go deeper into the patterns in the data, we need to explore the data and compute the various synthetic descriptors. This might involve calculating means and standard deviations for some of the variables we feel are most important. So, we say that we produce univariate summaries, and if there is a need we may also create some graphical summaries like line plots or frequency histograms. Be guided by the research questions as to what is required. Typically, I don't like to produce too many detailed inferential statistics of the multivariate data considered one variable at a time (there are special statistical techniques available that allow us to do so more efficiently and effectively, but we will get to it in the Honours Module Quantitative Ecology), choosing instead to see which relationships and patterns emerge from the exploratory summary plots before testing their statistical significance using multivariate approaches. But that is me. Sometimes, some hypotheses call for a few univariate inferential analyses (again, this is the topic of an Honours module on Biostatistics).

! Lab 1 (continue)

(To be reviewed by BCB743 student but not for marks)

2. Create an $x - y$ plot of the geographical coordinates in `DoubsSpa.csv`.
3. Using some graphs that plot the trends of the Doubs River environmental variables along the length of the river, describe the patterns in some of the environmental variables and offer explanations for how they might be responsible for affecting species distributions down the length of the Doubs River. Which three variables do you think will be able to explain the trends in the species data?

5 Pairwise Matrices

Although we typically start our forays into data exploration using sites × species and sites × environment tables, the formal statistical analyses usually require **pairwise association matrices**. Such matrices are symmetrical (sometimes only the lower or upper triangle is displayed) square matrices (i.e. $n \times n$). These matrices tell us how related any sample is to

any other sample in our pool of samples (i.e., relatedness among rows with respect to whatever populates the columns, be they species information of environmental information).

Let us consider various kinds of association matrices under the headings **Distances**, **Correlations**, **Associations**, **Similarities**, and **Dissimilarities**.

5.1 Distances

A frequently used distance metric in ecological and geographical studies is Euclidean distance. Euclidean distance represents the 'ordinary straight-line' distance between two points in Euclidean space. When working with geographical coordinates over small areas of Earth's surface, Euclidean distance is very similar (i.e., almost directly proportional) to the actual geographical distance, making the concept intuitive to understand.

In its simplest form, Euclidean distance is calculated in a planar Cartesian area, which is familiar as a graph with x- and y-axes. In 2D and 3D space, it gives distances in Cartesian units between points on a plane (x, y) or in volume (x, y, z). There is a linear relationship between the units in the physical realm and the units in Euclidean space, implying that short distances between pairs of points on a map or graph also represent short geographic distances on Earth.

Euclidean distance is calculated using the Pythagorean theorem and is typically applied to standardised environmental data (not species data):

$$d(a, b) = \sqrt{\sum_{i=1}^{n} (a_i - b_i)^2}$$

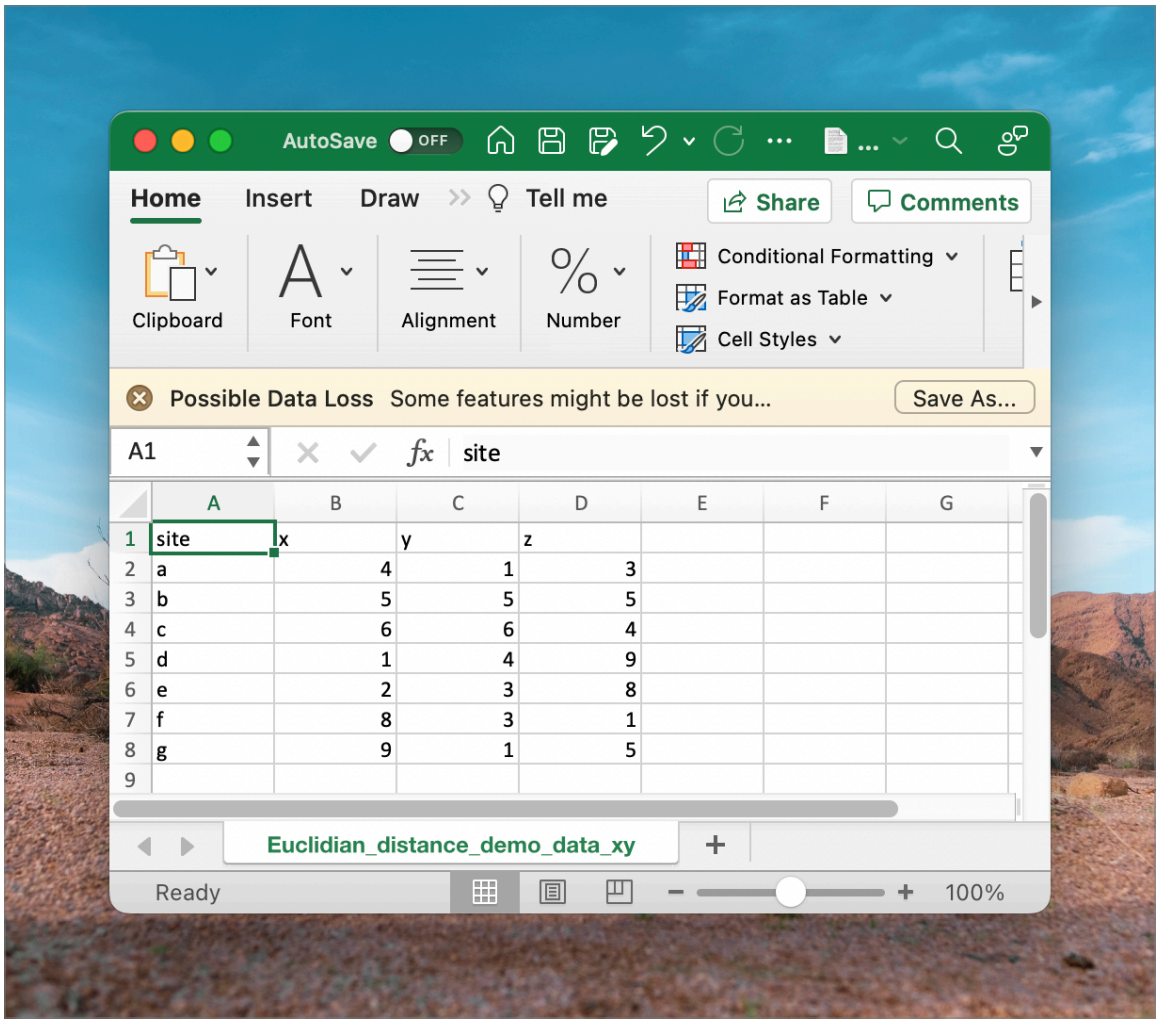
In this formula:

- a and b are two points in Euclidean space; in terms of environmental data, a and b represent two sites.
- Each element $i = 1$ through n in the vectors a and b represents a dimension or variable in the space. For example, if we have three environmental variables, $n = 3$, and the formula calculates the Euclidean distance between the two sites in a three-dimensional space.
- The summation $\sum_{i=1}^{n}$ goes over all dimensions from 1 to n.

Each coordinate or variable could represent different environmental factors such as temperature, depth, or light intensity (sometimes also called 'dimensions' of environmental space). For example, in the case of three environmental variables, the Euclidean distance would be calculated as:

$$d(a, b) = \sqrt{\left(a_{\text{temp}} - b_{\text{temp}}\right)^2 + \left(a_{\text{depth}} - b_{\text{depth}}\right)^2 + \left(a_{\text{light}} - b_{\text{light}}\right)^2}$$

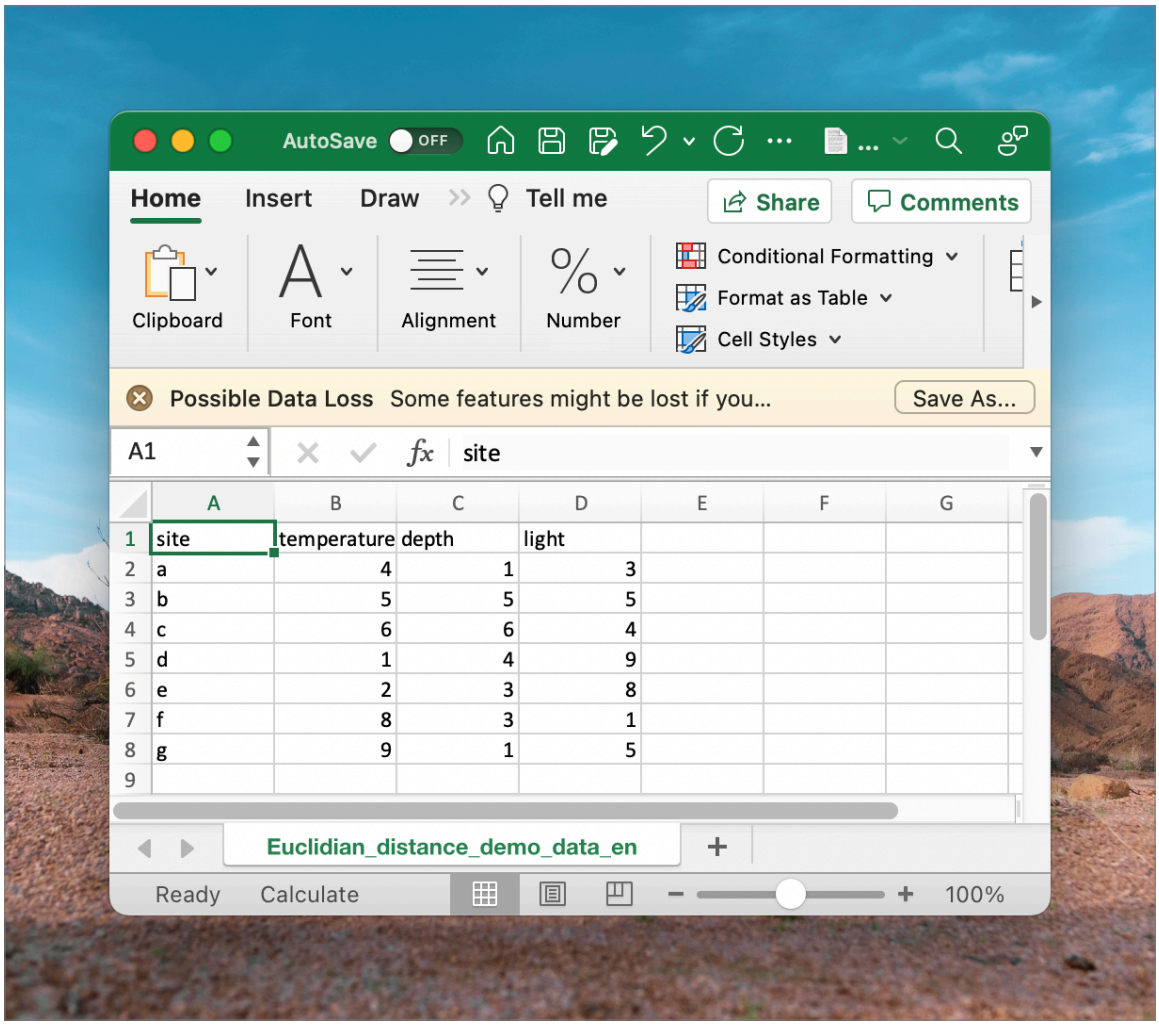
In the example dataset downloaded earlier (`Euclidean_distance_demo_data_xyz.csv`), we can calculate the distance between every pair of sites named ***a*** to ***g***. The 'raw' data representing x, y and z dimensions can be viewed in MS Excel, as we see in Figure 7.



	A	B	C	D
1	site	x	y	z
2	a	4	1	3
3	b	5	5	5
4	c	6	6	4
5	d	1	4	9
6	e	2	3	8
7	f	8	3	1
8	g	9	1	5
9				

Figure 7: Data representing three dimensions, x, y, and z.

We can substitute x, y and z for environmental 'dimensions,' and we have a set of data that resembles what we see in Figure 8. Regardless of whether we have x, y and z or environmental dimensions, the application of the Pythagorean Theorem is the same.



	A	B	C	D
1	site	temperature	depth	light
2	a	4	1	3
3	b	5	5	5
4	c	6	6	4
5	d	1	4	9
6	e	2	3	8
7	f	8	3	1
8	g	9	1	5
9				

Figure 8: Data representing three environmental 'dimensions.'

Figure 9 shows how we may calculate Euclidean distance in MS Excel using some built-in functions. The function `SUMXMY2` calculates the sum of the differences of squares between two corresponding arrays. It squares each value in array `x`, squares the corresponding value in array `y`, subtracts the `y`-square from the `x`-square, and then sums all these differences. That value is then subjected to a square-root calculation using `SQRT`.

To produce the pairwise matrix, you'd have to do this for every pair of sites. As a minimum, calculate the bottom left triangle. For completeness, calculate the diagonal, which will be all zeros in this (and every!) instance. It is a tedious process, I know!

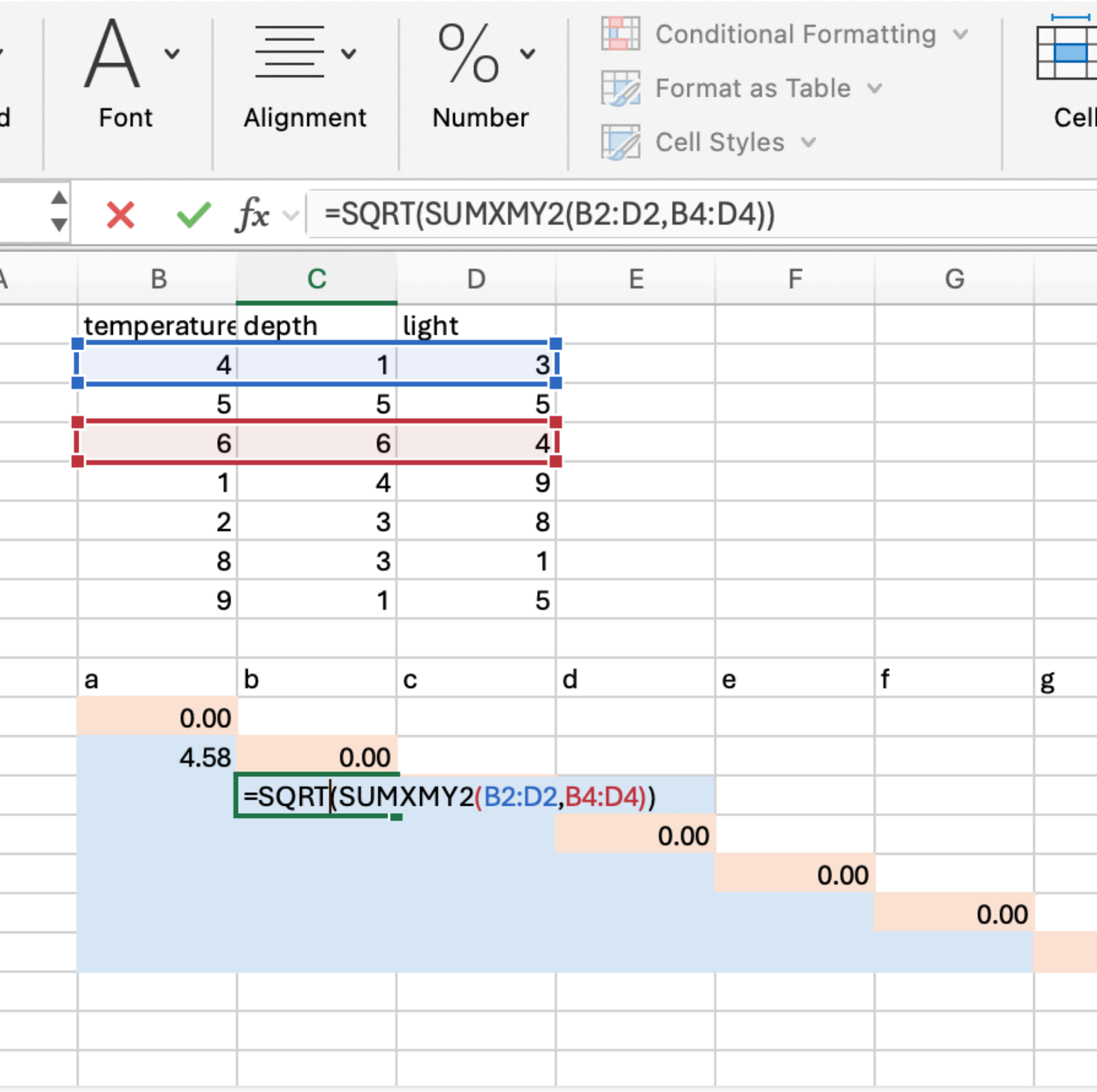


Figure 9: Calculating Euclidean distance in MS Excel. The pink shaded cells are the diagonal comprised of 0s, and the blue shaded cells are the lower triangle. The upper triangle remains unshaded but will be a mirror image of the lower triangle.

5.2 Standardisation

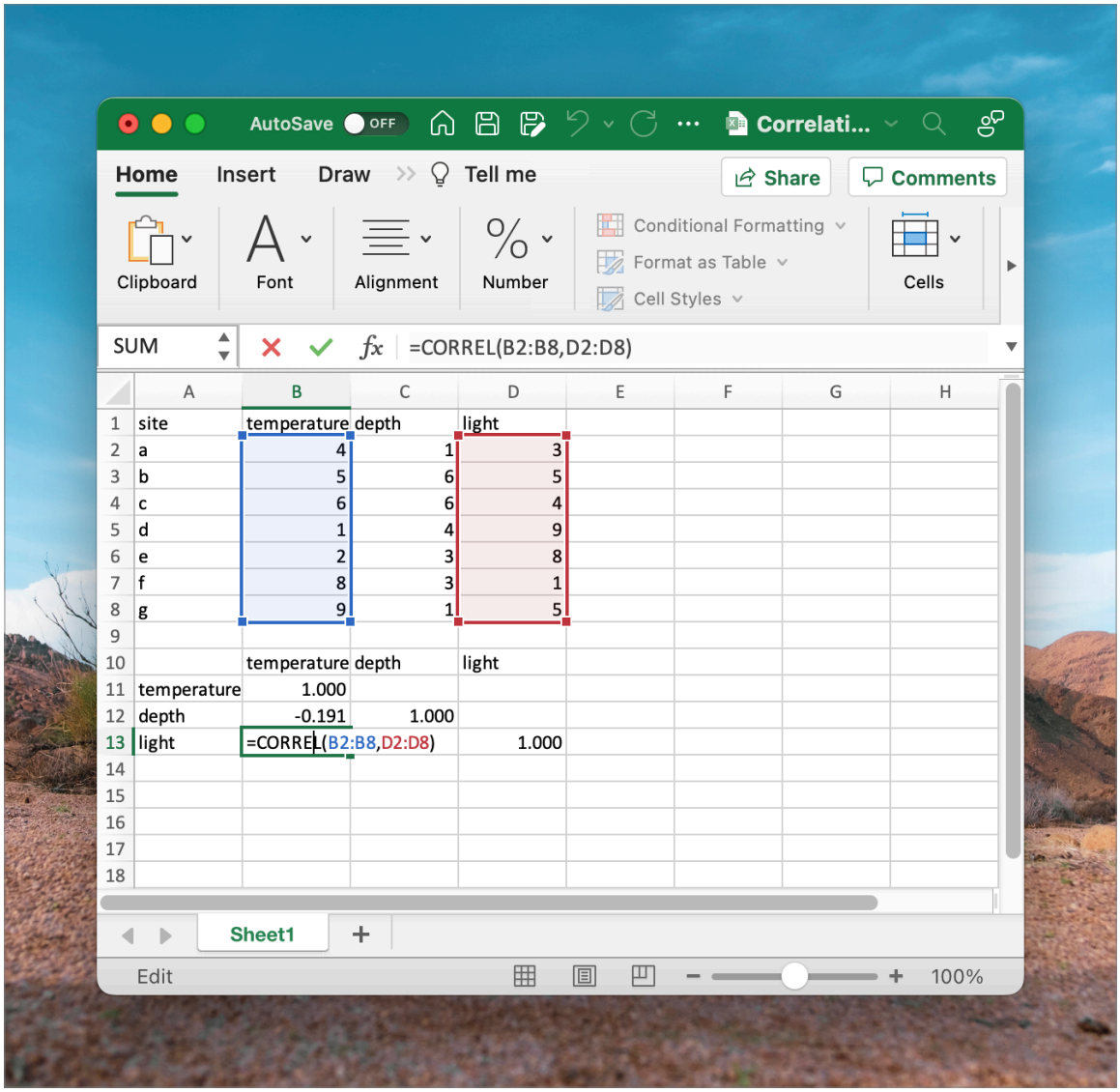
You should ensure that all your variables are standardised before applying the Euclidean distance calculations, as I have mentioned previously. This step is essential because the Euclidean distance is sensitive to the scale of the variables involved. If the variables are not standardised, those measured on larger scales may dominate the results, ultimately leading to misleading conclusions. Therefore, standardising your data enables each variable to contribute equally to the distance measure, maintaining the integrity of your subsequent analysis.

5.3 Correlations

Correlations ask whether two sets of variables, or rather, a pair of variables, exhibit any kind of relationship between them. For example, do we expect that as temperature increases, so too does humidity? In situations where an increase in temperature is associated with an increase in humidity, that is, both variables increase together, we would say that these samples are positively correlated.

Conversely, when discussing a negative correlation, we find that as one variable increases in magnitude, the variable we have paired with it demonstrates a corresponding decrease in its magnitude. In other words, there is an inverse relationship between those two variables.

So, we use correlations to establish how environmental variables relate across the sample sites. Therefore, a correlation performed to a sites × variable table is done between columns (variables), not rows, as in the Euclidean distance calculation, which compares the rows (sites). We do not need to standardise as one would for calculating Euclidean distances (but it will do no harm if you do). Correlation coefficients (so-called r-values) vary in magnitude from -1 (a perfect inverse relationship) from 0 (no relationship) to 1 (a perfect positive linear relationship).



	A	B	C	D	E	F	G	H
1	site	temperature	depth	light				
2	a	4	1	3				
3	b	5	6	5				
4	c	6	6	4				
5	d	1	4	9				
6	e	2	3	8				
7	f	8	3	1				
8	g	9	1	5				
9								
10		temperature	depth	light				
11	temperature	1.000						
12	depth	-0.191	1.000					
13	light	=CORREL(B2:B8,D2:D8)		1.000				
14								
15								
16								
17								
18								

Figure 10: Calculating pairwise correlations between environmental variables in MS Excel.

The resultant pairwise correlation matrix shows the names of the environmental variables as both column and row names. Contrast this with what is presented as row and column names in the distance matrix (Figure 10).

5.4 Associations, Similarities, and Dissimilarities

Thus far, we have worked with environmental data. Associations, similarities, and dissimilarities extend the pairwise matrix to species data. We will discuss and calculate these matrices in Lab 3.

That's it for this week, Folks! I'll leave you with some lovely exercises to take you through the rest of the week.

6 Summary of Species and Environmental Data

The diagram below (Figure 11) summarises the species and environmental data tables, and what we can do with them. These tables are the starting points of many additional analyses, and we will explore some of these ecological relationships later in this module.

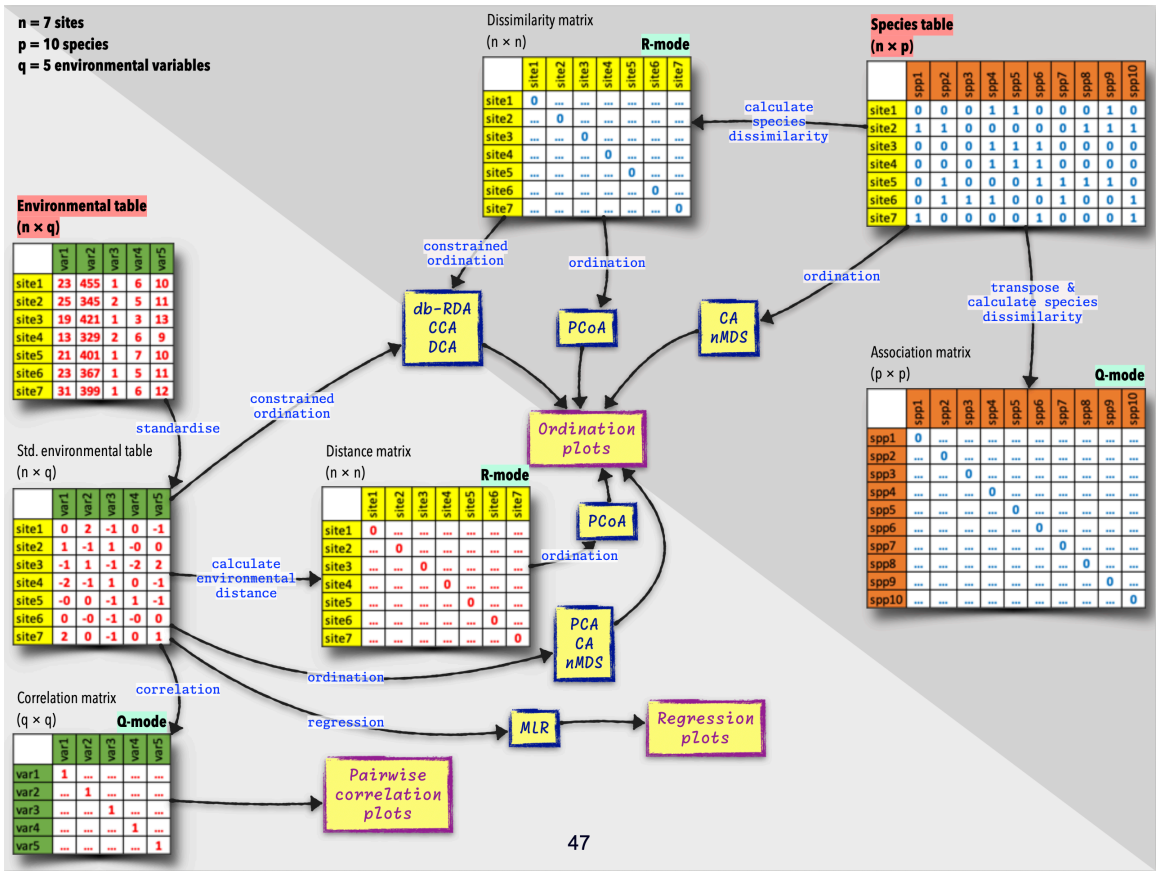


Figure 11: Species and environmental tables and what to do with them.

! Lab 1 (continue)

(To be reviewed by BCB743 student but not for marks)

4. Using the Doubs River environmental data, calculate the lower left triangle (including the diagonal) distance matrix for *every pair* of sites in Sites 1, 3, 5, …, 29 (i.e. using only every second site). Explain any patterns or trends in this resultant distance matrix regarding how similar/different sites are relative to each other. Which of the graphs you came up with in Task 3 (if any) do you think are responsible for the patterns seen in the distance matrix?
5. Using the same sites as above (Question 4), calculate a pairwise correlation matrix (lower left and including the diagonal) for the Doubs River environmental data. Explain any patterns or trends in this resultant correlation matrix and offer mechanistic explanations for why these correlations might exist.
6. Discuss in detail the properties of distance and correlation matrices.
7. If you found this exercise annoying, explain why. Or if you loved it, state why. What could be done to ease your experience of the calculations?
8. Okay, so how does all of this relate to macroecology? Please discuss the purpose of all of these approaches to what macroecology promises to accomplish. In your answer, also include consideration of the unimodal model (as in coenoclines, coenoplanes, and coenospaces) and its relevance to everything we aim to do here.

! Submission Instructions

The Lab 1 assignment on Ecological Data was discussed on Thursday 24 July and is due at **08:00 on Monday 28 July 2025**.

Provide a **neat and thoroughly annotated** MS Excel spreadsheet which outlines the graphs and all calculations and which displays the resultant distance matrix. Use separate tabs for the different questions. Written answers must be typed in an MS Word document. Please follow the formatting specifications *precisely* shown in the file **BDC334 Example essay format.docx** that was circulated at the beginning of the module. Feel free to use the file as a template.

Please label the MS Excel and MS Word files as follows:

- `BDC334_<first_name>_<last_name>_Lab_1.xlsx`, and
- `BDC334_<first_name>_<last_name>_Lab_1.docx`

(the < and > must be omitted as they are used in the example as field indicators only).

Submit your appropriately named spreadsheet and MS Word documents on iKamva when ready.

Failing to follow these instructions carefully, precisely, and thoroughly will cause you to lose marks, which could cause a significant drop in your score as formatting counts for 15% of the final mark (out of 100%).

7 References

Bibliography

Borcard D, Gillet F, Legendre P, others (2011) Numerical ecology with R. Springer

Shade A, Dunn RR, Blowes SA, Keil P, Bohannan BJ, Herrmann M, Küsel K, Lennon JT, Sanders NJ, Storch D, others (2018) Macroecology to unite all life, large and small. Trends in ecology & evolution 33:731–744.

Verneaux J (1973) Cours d'eau de Franche-Comté (Massif du Jura). Recherches écologiques sur le réseau hydrographique du Doubs.