**PRACTICE PROBLEMS – R language**

**1.** Install the ISwR R package. Write the built-in dataset *thuesen* to a tab-separated text file. View it with a text editor. Change the NA to . (period), and read the changed file back into R.

**2.** The exponential growth of a population is described by this mathematical function:

where Nt is the population size at time t, N0 is the initial population size, and r is the rate of growth or reproductive rate. Write an exponential growth function in R that also generates a plot with time on the x axis and Nt on the y axis. Using that function create plots for 20 days assuming an initial population size of 10 individuals under three growth rate scenarios (0.5, 0.8, -0.1).

**3.** Under highly favorable conditions, populations grow exponentially. However resources will eventually limit population growth and exponential growth cannot continue indefinitely. This phenomenon is described by the logistic growth function:

where K is the carrying capacity. Write a logistic growth function in R that also generates a plot with time on the x axis and Nt on the y axis. Using that function create plots for 20 days assuming an initial population size of 10 individuals and a carrying capacity of 1,000 individuals under three growth rate scenarios (0.5, 0.8, 0.4).

**4.** Write a function (*sum\_n*) that for any given value, say n, computes the sum of the integers from 1 to n (inclusive). Use the function to determine the sum of integers from 1 to 5,000.

**5.** Write a function (*sqrt\_round*) that takes a number x as input, takes the square root, rounds it to the nearest whole number and then returns the result.

**6.** Install the R package ‘nycflights13’, and load the ‘weather’ data.

a) Explore the columns names and the top part of the dataset to get a sense of the data

b) Make a subset of the data from just the first month (1) and then save that subset as ‘weather1’

c) Using ggplot2, make a beautiful histogram of the variable ‘temp’

d) Using ggplot2, make a beautiful line plot of ‘temp’ as a function of ‘time\_hour’

e) Using ggplot2, make a beautiful boxplot of ‘temp’ as a function of ‘origin’

**7.** Write content in markdown syntax to replicate, as much as possible, the format of the following sample text.

**Graphical user interface, text, application, email

Description automatically generated**

**PRACTICE PROBLEMS – Fundamentals of statistics**

**1.** A researcher videotaped the glides of 8 tree snakes leaping from a 10-m tower. Undulation rates of the snakes measured in hertz (cycles per second) were as follows: 0.9, 1.4, 1.2, 1.2, 1.3, 2.0, 1.4, 1.6.

a) Draw a histogram of the undulation rate

b) Calculate the sample mean

c) Calculate the range

d) Calculate the standard deviation

e) Write a function to express the standard deviation as a percentage of the mean (that is, the coefficient of variation) and calculate it.

**2.** Blood pressure was measured (in units of mm Hg). Here are the measurements: 112, 128, 108, 129, 125, 153, 155, 132, 137.

a) How many individuals are in the sample?

b) What is the mean of this sample?

c) What is the variance?

d) What is the standard deviation?

e) What is the coefficient of variation?

**3.** The data in the file *DesertBirdAbundance.csv* are from a survey of the breeding birds of Organ Pipe Cactus National Monument in southern Arizona.

a) Draw a histogram of the abundance data.

b) Calculate the median and the mean of the bird abundance data.

c) In this particular case, which do you think is the best measure of center, the mean or the median?

d) Calculate the range, standard deviation, variance and coefficient of variation of the bird abundance data.

**4.** Calculate the probability of each of the following events:

a) A standard normally distributed variable is larger than 3

b) A normally distributed variable with mean 35 and standard deviation 6 is larger than 42

c) Getting 10 out of 10 successes in a binomial distribution with probability 0.8

d) X > 6.5 in a Chi-squared distribution with 2 degrees of freedom

**5.** Demonstrate graphically the central limit theorem (by sampling and calculating the mean) using a Binomial distribution with 10 trials (size=10) and 0.9 probability of success (prob=0.9) and a sample size of 5. Run it with 10 samples and 100 samples as well to show the change in distribution.

**6.** Imagine height is genetically determined by the combined (that is, the sum) effect of several genes (polygenic trait). Assume that each gene has an effect on height as a uniform distribution with min=1 and max=3. Simulate an stochastic model of height for 1,000 random people based on 1 gene, 2 genes, and 5 genes. As we increase the number of genes, what is the resulting height distribution?

**7.** Do the same as in the previous problem but now assuming that the combined effect of the genes is multiplicative (not the sum). As we increase the number of genes, what is the resulting height distribution?

**PRACTICE PROBLEMS – Statistical tests**

**1.** Normal human body temperature is 98.6 F. Researchers obtained body-temperature measurements on randomly chosen healthy people: 98.4, 98.6, 97.8, 98.8, 97.9, 99.0, 98.2, 98.8, 98.8, 99.0, 98.0, 99.2, 99.5, 99.4, 98.4, 99.1, 98.4, 97.6, 97.4, 97.5, 97.5, 98.8, 98.6, 100.0, 98.4.

a) Make a histogram of the data

b) Make a normal quantile plot

c) Perform a Shapiro-Wilk test to test for normality

d) Are the data normally distributed?

e) Are these measurements consistent with a population mean of 98.6 F?

**2.** The brown recluse spider often lives in houses throughout central North America. A diet-preference study gave each of 41 spiders a choice between two crickets, one live and one dead. 31 of the 41 spiders chose the dead cricket over the live one. Does this represent evidence for a diet preference?

**3.** Ten epileptic patients participated in a study of a new anticonvulsant drug. During the first 8-week period, half the patients received a placebo and half were given the drug, and the number of seizures were recorded. Following this, the same patients were given the opposite treatment and the number of seizures were recorded. Assuming that the distribution of the difference between the placebo and drug meets the assumption of normality, perform an appropriate test to determine whether there were differences in the number of epileptic seizures with and without the drug.

|  |  |  |
| --- | --- | --- |
| Patient | Placebo | Drug |
| 1 | 37 | 5 |
| 2 | 52 | 23 |
| 3 | 68 | 40 |
| 4 | 4 | 3 |
| 5 | 29 | 38 |
| 6 | 32 | 19 |
| 7 | 19 | 9 |
| 8 | 52 | 24 |
| 9 | 19 | 17 |
| 10 | 12 | 14 |

a) Make a boxplot of the data

b) Test the difference.

**4.** A bee biologist is analyzing whether there was an association between bee colony number and type of forest habitat. We expect that there is no habitat preference for bee colony number. Is this true based on this data?

|  |  |
| --- | --- |
| Habitat | Bee colonies |
| Oak | 33 |
| Hickory | 30 |
| Maple | 29 |
| Red cedar | 4 |
| Poplar | 4 |

a) Make a barplot of the data

b) Test the hypothesis of no-association

**5.** Perform 10 one-sample t-tests for mu=0 on simulated standard normally distributed data of 25 observations each and get the P-value. Repeat the experiment, but instead simulate samples of 25 observations from a t-distribution with 2 degrees of freedom. Find a way to automate the first experiment to do it a 1,000 times and applying a false discovery rate correction to the P-values (check ?replicate).

**PRACTICE PROBLEMS – ANOVA**

**1.** You are hired by the US Department of Agriculture to develop effective control practices for invasive plants (data\_ANOVA.xlsx). In particular, you are asked to investigate pesticide controls for kudzu which is an invasive vine that grows in thick mats that smother underlying plants. Two of the most widely used pesticides for kudzu are glyphosate and triclopyr. To determine the effectiveness of a single application of pesticide, you conduct an experiment in 18 plots that each had 50% kudzu cover. In mid-summer, you applied equal amounts of 2% glyphosate to 6 plots, 2% triclopyr to 6 plots, and water without pesticides to 6 plots. Then you returned in autumn and measured the percent cover of kudzu in the plots.

a) Transform the data from wide to long format (Google how to do it - Hint: function melt from reshape2 package)

b) Make a boxplot of the data

c) Perform an ANOVA

d) What is the variation explained (R2)?

e) Perform a post hoc test

**2.** Data in growth.txt come from a farm-scale trial of animal diets. There are two factors: diet and supplement. Diet is a factor with three levels: barley, oats and wheat. Supplement is a factor with four levels: agrimore, control, supergain and supersupp. The response variable is weight gain after 6 weeks.

a) Inspect the data using boxplots

b) Perform a two-way ANOVA

c) Assess graphically if there exists an interaction between both factors

d) Given what you learnt about the interaction, what would be a better model?

e) What is the variation explained (R2) of this new model?

f) Perform a Tukey HSD test of this new model

**PRACTICE PROBLEMS – Correlation**

**1.** In a study of hyena laughter, a researcher investigated whether sound spectral properties of hyenas’ giggles are associated with age.

Age(years): 2,2,2,6,9,10,13,10,14,14,12,7,11,11,14,20

Frequency (Hz): 840,670,580,470,540,660,510,520,500,480,400,650,460,500,580,500

a) Inspect the data using a scatterplot

b) Test the linear association between both variables

c) Assume that the data are not normally distributed, test the linear association using a non-parametric correlation coefficient

**PRACTICE PROBLEMS – Regression**

**1.** Testosterone is known to predict aggressive behavior and is involved in face shape during puberty. Does face shape predict aggression? Researchers measured the face width-to-height ratio of 21 university hockey players with the average number of penalty minutes awarded per game for aggressive infractions. The data are available in face.txt.

|  |  |
| --- | --- |
| Face ratio (x) | Penalty minutes (y) |
| 1.59 | 0.44 |
| 1.67 | 1.43 |
| 1.65 | 1.57 |
| 1.72 | 0.14 |
| 1.79 | 0.27 |
| 1.77 | 0.35 |
| 1.74 | 0.85 |
| 1.74 | 1.13 |
| 1.77 | 1.47 |
| 1.78 | 1.51 |
| 1.76 | 1.99 |
| 1.81 | 1.06 |
| 1.83 | 1.20 |
| 1.83 | 1.23 |
| 1.84 | 0.80 |
| 1.87 | 2.53 |
| 1.92 | 1.23 |
| 1.95 | 1.10 |
| 1.98 | 1.61 |
| 1.99 | 1.95 |
| 2.07 | 2.95 |

a) How steeply does the number of penalty minutes increase per unit increase in face ratio? Calculate the estimate of the intercept. Write the result in the form of an equation for the line.

b) How many degrees of freedom does this analysis have?

c) What is the t-statistic?

d) What is your final conclusion about the slope?

e) What is the variation explained, R2?

f) Make a scatterplot of the data and include a linear regression line

g) Check the model assumptions

**2.**

|  |  |
| --- | --- |
| Body mass (g) | Respiration rate (nmol/s) |
| 453 | 666 |
| 1283 | 643 |
| 695 | 1512 |
| 1640 | 2198 |
| 1207 | 2535 |
| 2096 | 4176 |
| 2804 | 3196 |
| 3528 | 3494 |
| 5940 | 7386 |
| 10000 | 10363 |

a) Make a scatterplot of the raw data

b) Make a scatterplot of the linearized relationship. Which transformation did you use?

c) Use linear regression to estimate β

d) Carry out a formal test of the null hypothesis of β=0 Respiratory rate (Y) is expected to depend on body mass (X) by the power law, Y=aXβ, where β is the scaling exponent. The data are available in respiratoryrate\_bodymass.txt.

e) What is the variation explained, R2?

f) Check the model assumptions

**PRACTICE PROBLEMS – Statistical Modeling**

**1.** Use the ozone.txt data to model the ozone concentration as a linear function of wind speed, air temperature and the intensity of solar radiation. Assume that the requirements to perform a linear regression are met.

a) Make a multiple panel bivariate scatterplot

b) Perform a multiple linear regression

c) What is the variation explained, R2?

d) Assess the collinearity of the explanatory variables using the variance inflation factor

e) Check the model assumptions

**2.** Use the diminish.txt data (xv is explanatory, yv is response variable) to:

a) Perform a simple linear regression

b) Perform a polynomial (second-degree) regression

c) Compare both models with Akaike’s Information Criterion (AIC). Which model is better?

d) Make a scatterplot of the data and include both regression lines

**3.** The data in stork.txt display the stress-induced corticosterone levels circulating in the blood of European white storks and their survival over the subsequent five years of study.

a) Make a scatterplot of the data

b) Which type of regression model is suitable for these data?

c) Perform an appropriate regression to predict survival from corticosterone

d) What is the pseudo-R2 of the model?

e) What is the p-value of the model?

f) Include the predicted curve in the scatterplot

g) Check the model assumptions

**4.** The clusters.txt dataset contains the response variable Cancers (the number of reported cancer cases per year per clinic) and the explanatory variable Distance (the distance from a nuclear plant to the clinic in kilometers).

a) Make a scatterplot of the data

b) Which regression is the more appropriate for these data? (Don’t take overdispersion into account for now)

c) Given your choice, is the trend significant?

d) What is the pseudo-R2 of the model?

e) What is the p-value of the model?

f) Include the predicted relationship from the model in the scatterplot

g) Do you think there might be some evidence of overdispersion?

h) Perform a new generalized linear model with a distribution that better accounts for overdispersion

i) Check this last model assumptions

**5.** Use the jaws.txt data to:

a) Make a scatterplot of the data (age explanatory, bone response)

b) Perform a non-linear regression assuming an asymptotic exponential relationship:

c) Perform a non-linear regression assuming a Michaelis-Menten model:

d) Estimate the percentage of variation explained by both models (comparing them with a null model with only a constant)

e) Compare both models with Akaike’s Information Criterion (AIC). Which model is better?

f) Make a scatterplot of the data and include both regression lines

**6.** In a recent paper we read: “*Linear mixed effects modelling fit by restricted maximum likelihood was used to explain the variations in growth. The linear mixed effects model was generated using the lmer function in the R package lme4, with turbidity, temperature, tide, and wave action set as fixed factors and site and date set as random effects*”. Write down the R code to recreate their model.

**7.** Researchers at the University of Arizona want to assess the germination rate of saguaros using a factorial design, with 3 levels of soil type (remnant, cultivated and restored) and 2 levels of sterilization (yes or no). The same experimental design was deployed in 4 different greenhouses. Each of the unique treatments was replicated in 5 pots. 6 seeds were planted in each pot.

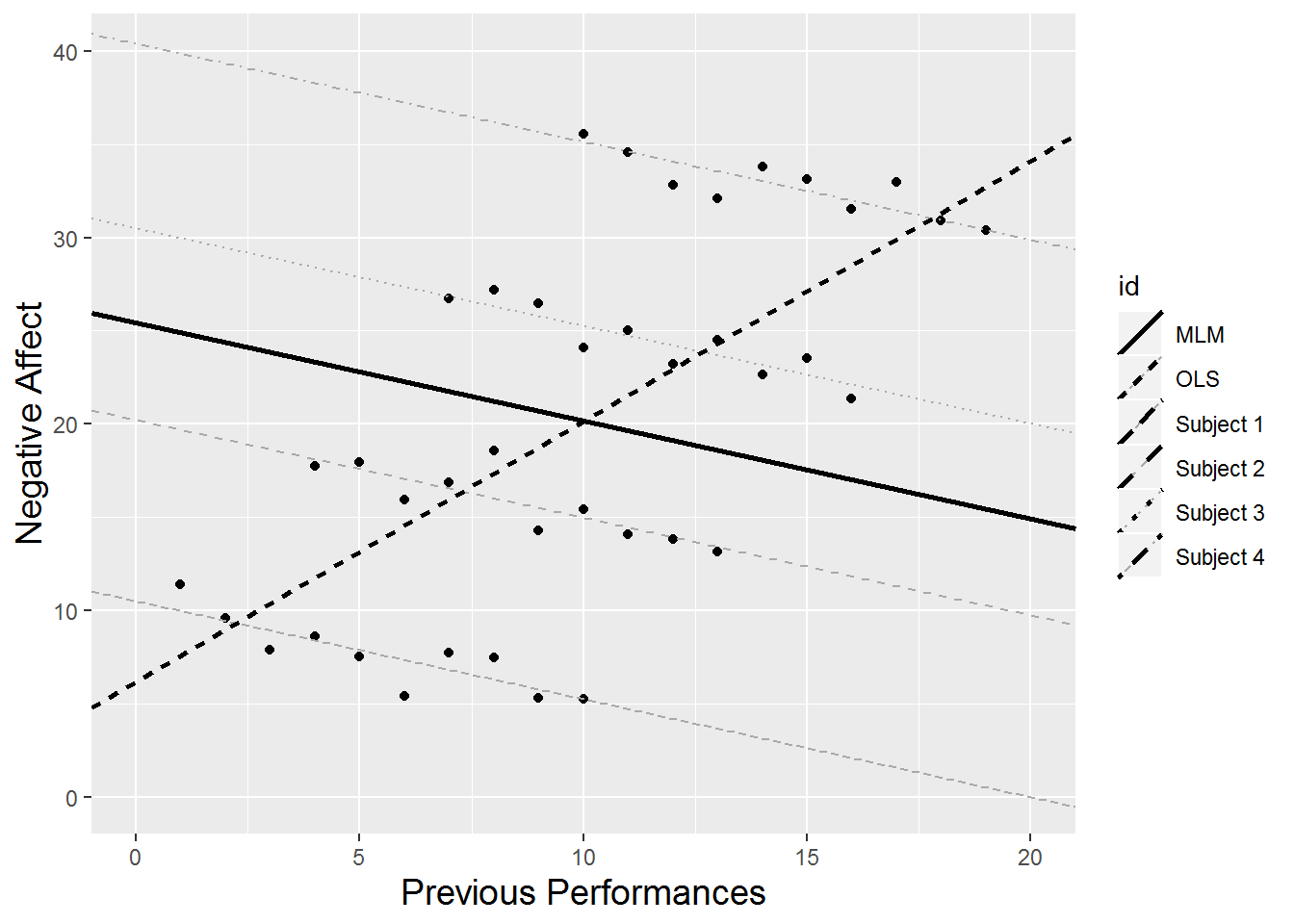
a) How many fixed treatments (unique combinations) exist?

b) What is the total number of pots?

c) What is the total number of plants measured?

d) Write the R code that you would use to analyze these data

**8.** Check these hypothetical data from 4 subjects relating number of previous performances to negative affect.



a) What does the thick dashed black line represent?

b) What is depicting the solid black line?

c) What do the 4 thin dashed lines represent?

**9.** Load dragons.RData

Load(“C:/path/to/.RData”)

a) Perform a simple linear regression with testScore as response and bodyLength as explanatory

simple <- lm(testScore ~ bodyLength, data = dragons)

summary(simple)

b) Plot the data with ggplot2 and add a linear regression line with confidence intervals. Hint: geom\_smooth()

ggplot(dragons, aes(x = bodyLength, y = testScore)) +

geom\_jitter() +

geom\_smooth(method = ‘lm’, formula = y ~ x)

c) We collected multiple samples from 8 mountain ranges. Generate a boxplot using (or not) ggplot2 to explore this new explanatory variable

ggplot(dragons, aes(x = mountainRange, y = testScore))+

geom\_jitter()+

geom\_boxplot(aes(fill=mountainRange))

d) Now repeat the scatterplot in b) but coloring by mountain range and without linear regression line

e) Instead of coloring, use facet\_wrap() to separate by mountain range

facet\_wrap(~mountainRange)

f) Perform a new linear model adding mountain range and assuming that it is fixed effects

model <- lm(testScore ~ bodyLength + mountainRange, data = dragons)

summary(model)

g) Perform the same linear model as before but now assuming mountain range is random effects

lmer(testScore

h) How much of the variation in test scores is explained by the random effect (mountain range)

Not much

i) Estimate the R2 (conditional and marginal) of the mixed-effects model

j) Check this last model assumptions

k) Include the fitted lines for each mountain range (plot from e). [Hint: predict function within geom\_line layer]

ggplot(dragons, aes(x = bodyLength, y= testScore, color = mountainRange)) +

geom\_line(data = cbind(dragons, pred = predict(model)), aes(y = pred))+

geom\_jitter()+

facet\_wrap(~mountainRange)

**10.** Data in Estuaries.csv correspond to counts of invertebrates at 3-4 sites in each of 7 (randomly chosen) estuaries.

a) Fit a linear mixed model with Total as response and Modification as explanatory, controlling for Estuary

b) Estimate the R2 (conditional and marginal) of this model

c) Plot the data with ggplot2 in a way that helps you understand the different effects

d) Include the variable Site as a random effect. Do you think this corresponds to a crossed or a nested design?

e) What are the R2 (conditional and marginal) of the model including Site

f) Check the model assumptions

g) Plot the data trying to include Site

h) Transform the variable Hydroid to presence/absence data

i) Fit a generalized linear mixed model (GLMM) with this transformed variable as Response and the same fixed and random effects as in d). [Hint: function glmer]

j) Check the model assumptions

**PRACTICE PROBLEMS – Time series and spatial analysis**

**1.** Write a function to calculate the 3-point moving average for an input vector. The formula is:

**2.** Read the temp.txt file. The data correspond to monthly average temperatures.

a) Plot the time series data. [Hint: first you need to create a monthly time series object]

b) Calculate the 5-point moving average and plot it together with the time series

c) Decompose the time series into seasonal, trend and residual error components

d) Generate a temporal correlogram to assess the autocorrelation of the time series

e) Generate a new correlogram but removing the trend and seasonal variation

f) Generate a partial correlogram plot

g) Find the best ARIMA model using the forecast package

h) Estimate future values using the previous ARIMA model and plot the results

**3.** Read the bac\_dust.txt dataset. The data corresponds to proteobacterial abundance and bacterial species in dust samples collected around the globe.

a) Make a map of the sampling points using the R libraries maps and ggplot2. Color the points based on the continent they were collected.

b) Make a map of the sampling points using the R libraries maps and ggplot2. Size the points based on bacterial richness.

c) Calculate Moran’s I autocorrelation for the variable Richness and Proteobacteria

d) Make a simple linear regression of Richness as a function of Proteobacteria. Are the residuals spatially autocorrelated?

e) Perform a generalized linear mixed model (GLMM) using the coordinates as random effect and using the Matérn covariance function (R package spaMM).

f) Report the nu and the rho values for this spatial regression model.

g) What can you say about the estimated correlation of this particular spatial regression model across increasing distances between pairs of locations.

**4.** Download the map of Spain.

a) Make a basic map of Spain

b) Get information on the population of Spanish cities and make a map of Spain including the locations and population of Spanish cities.

c) Visit Spain and enjoy

**5.** Download GBIF georeferenced occurrence data of the Pyrenean desman (*Galemys pyrenaicus*) and make a map of its geographical distribution (Hint: it is only present in Spain, Andorra, Portugal and France).

**PRACTICE PROBLEMS – Diversity analysis**

**1.** Load the dune\_bio.txt dataset. Species should be in columns and sites in rows.

a) Calculate the total number of individuals of all species

b) Calculate the total number of individuals for each species

c) Calculate the average number of individuals for each species

d) Calculate the total number of individuals for each site

e) Calculate the average number of individuals for each site

f) Write a function to report the median number of individuals for each species and the median number of individuals for each site

g) Use the vegan function decostand() to transform the dataset to relative abundances. How would you confirm the transformation worked?

h) Use the vegan function decostand() to standardize the dataset into the range 0 to 1

i) Use the vegan function decostand() to standardize the dataset to mean=0 and variance=1

**2.** Write a function to calculate the observed richness of a vector representing the abundances of different species in a community.

**3.** Write a function to calculate the Shannon-Wiener diversity index of a vector representing the abundances of different species in a community.

**4.** Write a function that calculates both the observed richness and the Shannon-Wiener diversity index for each community in a matrix, and writes an output table with the results. You can use the vegan built-in functions. Test your function with the dune\_bio.txt dataset.

**5.** Make a rank-abundance curve using the second site (13) of the dune\_bio.txt dataset. Fit a lognormal model to the data.

**6.** Create a Euclidean distance matrix after standardization using the first (A1), second (Moisture) and fifth (Manure) variables of dune\_env.txt. Then create a Bray-Curtis distance matrix using dune\_bio.txt. Perform a Mantel test on both distance matrices and plot the relationship.

**PRACTICE PROBLEMS – Cluster analysis**

**1.** Make dendrograms of the results of hierarchical clustering using the single, average and complete methods. Use the dune\_bio.txt dataset after creating a Bray-Curtis distance matrix.

**2.** Perform k-means partitions from 2 to 6 on the dune\_bio.txt dataset and select the best partition using the Calinski criterion. Visualize the results.

**PRACTICE PROBLEMS – Ordination**

**1.** Load the varechem dataset within the R package vegan. This data frame collects soil characteristics. Given the nature of this dataset perform a PCA or a CA ordination analysis.

a) How much variation is explained by the two first axes?

b) Make a screeplot of the results

c) Plot the ordination results of the sites

d) Plot both the sites scores and the soil characteristics scores focusing on the soil variables

**2.** The dune\_bio.txt dataset corresponds to species abundances. Given the nature of this dataset perform a PCA or a CA ordination analysis.

a) How much variation is explained by the two first axes?

b) Make a screeplot of the results

c) Plot the ordination results of the sites

**3.** Perform a nonmetric multidimensional scaling (NMDS) on the dune\_bio.txt dataset after calculating Bray-Curtis distances among sites.

a) What is the stress?

b) Make a Shepard plot of the NMDS results

c) Plot the ordination results of the sites

d) [Advanced – ggplot2] Plot the ordination results using ggplot2. Use the dune\_env.txt dataset to make the size of the points proportional to the site richness (number of species) and the color to represent the variable Management.

**4.** Perform a constrained correspondence analysis (CCA) on the dune\_bio.txt dataset using the first (A1), second (Moisture) and fifth (Manure) variables of dune\_env.txt as explanatory.

a) What is the variation explained by the two first constrained axes?

b) What is the adjusted R2 of the model?

c) Use a permutation test to assess the overall statistical significance of the model

d) Use a permutation test to assess the marginal statistical significance of each explanatory variable of the model. Which variables are significant?

e) Make a triplot of the ordination results focusing on the sites

**5.** Perform a permutational multivariate analysis of variance (PERMANOVA) with 9,999 permutations using the dune\_bio.txt dataset after calculating Bray-Curtis distances and the first (A1), second (Moisture) and fifth (Manure) variables of dune\_env.txt as explanatory.

a) Which explanatory variables are significant?

b) What is the explanatory power (R2) of the significant variables?

c) Perform a new PERMANOVA analysis with 9,999 permutations with A1 as the explanatory variable but constrain the permutations within the Use variable.

d) Plot a NMDS ordination to visually confirm your results in c).