BCB743 Quantitative Ecology Exam **Total 230 marks (6 hours)**

Please provide your answers in a properly set-up .R or .Rmd file. To note:

1. Label questions and answers appropriately and clearly.
2. Marks will be deducted for failing to correctly specify the file path (i.e. all data files must reside at the same level of the folder hierarchy as the script itself).
3. Name the file as per this example: "Smit_BCB743_exam_2018.R" or "Smit_BCB743_exam_2018.Rmd"
4. Ensure that functioning code is provided for each step of your analysis, including all data transformations, summary statistics, graphs, and final analyses.
5. Provide liberal amounts of commenting throughout in the format necessary to support .R or .Rmd files.
6. All interpretative answers must be provided in the same manner.
7. There is no need to save the numerical or graphical outputs as these will be recreated by running the script on my end. For this reason, it is essential that the script is fully functional upon submission, as no attempt will be made to solve problems due to non-functioning code. A large number of marks can be lost in this way, so please pay attention!

Question 1 [60 marks]
The question concerns the file '`dataset_1.xls`', which includes the abundances of 30 fish species at each sampling location, together with the bottom depth, the temperature and the geographic coordinates (latitude and longitude).

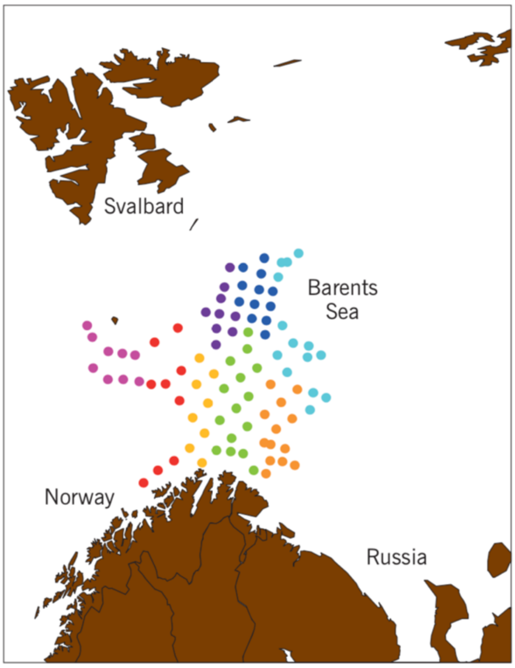

Figure 1. Locations of samples in the "dataset_1.xls" data set. At each sampling point the data consist of the abundances of 30 fish species, the bottom depth, the temperature and the spatial position (latitude and longitude). The stations have been colour coded into neighbouring groups, using great circle distances, for comparison with the ordination maps based on the abundances.

a. Specify the dimensions of the two data sets (i.e. excluding the geographic coordinates). [2]
b. Recreate the map (Figure 1), making sure to include as much of the detail displayed there as possible. Instead of showing colour coding, scale the size of the sample locations by the
    i. species richness (i.e. alpha-diversity, which is simply the number of species per location),
    ii. Shannon–Weaver index, and
    iii. Simpson's index.

    (*Hint: see the **vegan** function, '`diversity()`' for the diversity indices*) [10]
c. Make a new map as per (b), and this time, scale the symbols by
    i. temperature, and
    ii. depth. [4]
d. Based on a visual comparison between (b) and (c), are there already some patterns evident? If so, what are they? [5]
e. Provide a i) table and ii) figure(s) of the descriptive statistics of the environmental variables. What do the descriptive statistics tell us about the environment? [6]
f. This question requires you to perform a Correspondence Analysis (CA) on the species table. You will analyse the **(a) raw data**, **(b) log-transformed data**, and **(c) presence-absence data**. We will focus on

four species (*Boreogadus saida*, *Triglops murrayi*, *Notolepis rissoi krøyeri*, and *Trisopterus esmarkii*) that vary strongly along CA axes 1 and 2.

i. Show the code that produces the three CAs (i.e. on raw, presence-absence, and log-transformed data). [6]

ii. Focusing now on the output of the analysis on the raw data, provide numerical support that the four species named above are strongly influenced along CA1 and CA2 (i.e. provide the associated numerical support next to each species name). In terms of biplots, how would this visually manifest? [5]

iii. For each analysis (a-c), how much variance is associated with CA1, CA2, and CA3? What is the cumulative variance explained by these three axes (show the calculations, *and* explain where to find this information from the standard output)? Do you consider these ordinations to be very good at capturing the variation that exist across space in the community composition? [7]

iv. Now, using **vegan**'s '`ordisurf()`' function (see example code in your handouts), create panelled plots (i.e. four figures arranged in two rows and two columns per panel) for each ordination (a-c). In each of the four sub panels, focus separately on the four key species (i.e. one species per sub-panel), and superimpose the environmental vectors (temperature, depth, latitude, and longitude) on one of the sub-panels in each of the groups of plots. What are the major patterns that come out? Which are the most influential environmental drivers for each of the species? What is the effect of transforming the data on the analysis and the interpretation of the outcomes? [15]

Question 2 (110 marks)

This question is based on the Barro Colorado Island Tree Counts datasets that come with the **vegan** package. Load it as '`data("BCI")`' and '`data("BCI.env")`'.

For your analysis, you will transform the species data to presence-absence prior to starring the analysis. The question requires that you perform a Principal Components Analysis (PCA) and an CA, so go ahead and set-up the R file accordingly before proceeding with the questions.

a. Provide a full descriptive analysis of the environmental data and provide tables and/or figures as necessary. [10]

b. Create graphs that show the i) species richness, ii) Shannon–Weaver index, and the iii) Simpson's index across the landscape. Explain the patterns that are visible. [10]

c. Provide a detailed written account of the output of the PCA as per the '`summary()`' function. [15]

d. Provide a detailed written account of the output of the CA as per the '`summary()`' function. [15]

e. Using the functionality of the '`ordisurf()`' package, demonstrate the major difference between the analysis conducted using a PCA and a CA. Hint, you will have to use the argument '`family = "gaussian"`' inside of '`ordisurf()`'. In your graphic display of the differences, please use as demonstration the two species that are most heavily loaded along CA1 / PCA1 and the two species that are most heavily loaded along CA2 / PCA2. Which analysis is the better one, and why do you say so? [20]

f. Analyse these datasets using a non-Metric Multidimensional Scaling (nMDS), produce the necessary plots (of both axis 1 vs. axis 2 and axis 1 vs. axis 3), and interpret. Is any new insight possible from the nMDS that was not possible before? [20]

g. Within the context of the published literature available on the studies based on these data, how do your analyses add to that which is

   i. already known, and

   ii. what aspects of the published analysis are not captured by your own analysis? [20]

   (*Hint: yes, you will have to read papers*)

Question 3 (60 marks)

This question refers to the datasets in Questions 1 and 2, above.

a. Do a full RDA on the '`dataset_1.xls`' data, explain the output comprehensively, and provide any supporting figures you may deem necessary. What is your interpretation of the analysis? What new information can be obtained from the RDA that could not be found using the earlier analysis? [30]

b. Do a full RDA on the Barro Colorado Island Tree Counts datasets, explain the output comprehensively, and provide any supporting figures you may deem necessary. What is your interpretation of the analysis? What new information can be obtained from the RDA that could not be found using the earlier analysis? [30]