# Analysis of Water Quality Data from European Rivers

## Part A: Analysis of River Data from Water Quality Study

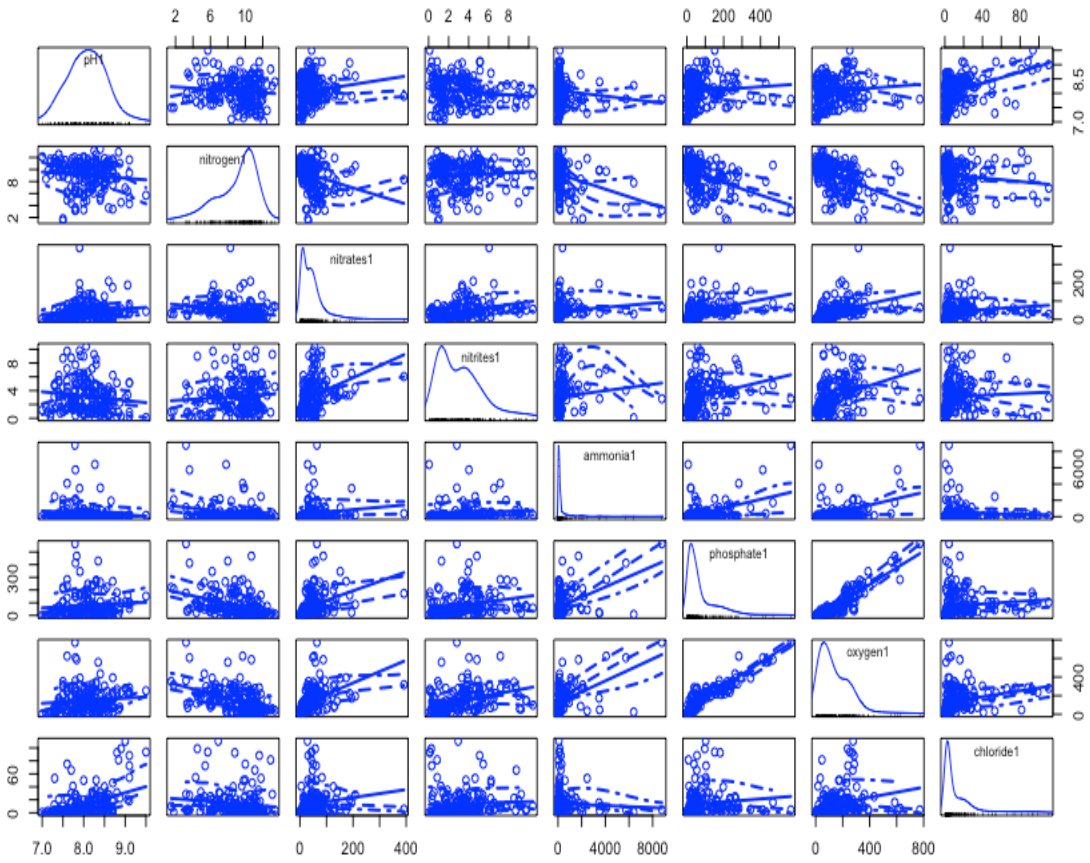
**Task 1:**

**Data Cleaning**

Data was collected from 200 rivers across Europe as part of an investigation into water quality and environmental impact on waterways. Data was collected in relation to the characteristics, chemical composition and the amounts of different species of algae in each river. A total of 18 variables are present in the data set. Initially in data cleaning, 16 cases were removed from the data set due to missing data for some variables. Data structure was then assessed and chemical variables were converted from a factor to a numeric variable for the purpose of analysis.

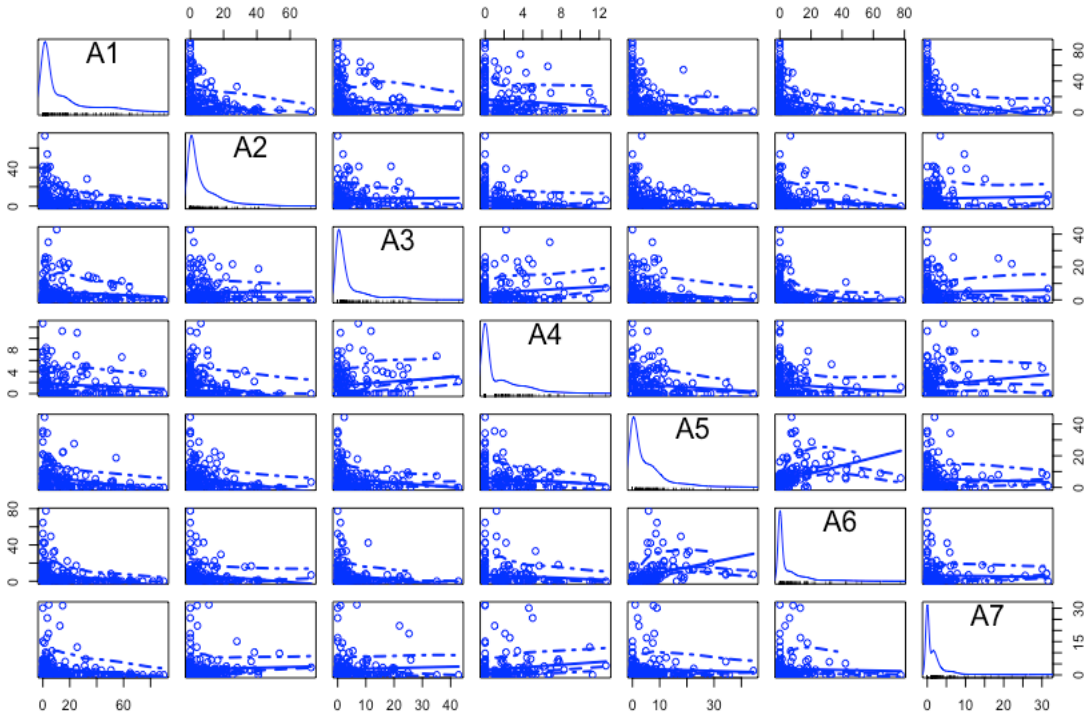
Data subsets of the chemical and algal variables respectively were then created to assess the data for initial relationships and outliers. The two scatterplot matrices below were created of each data subset and summary statistics were calculated. Given the large number of variables separating them into two groups allowed for easier visualisation of relationships.

*Fig 1. Scatterplot matrix of Chemical Variables*




From initial assessment of both scatterplots, it is evident that a number of the distributions are significantly skewed. On examination of the relationships between chemical variables, ammonia in particular has a significant right skew due to some large outliers. Right skews are also evident with chloride, phosphate, nitrates, oxygen and nitrites. Nitrogen has a left skew. The only variable with a near normal distribution is pH. There is a strong positive correlation between phosphate and oxygen however minimal to mild correlation between the other variables.

*Fig 2. Scatterplot matrix of the algae variables*




In the scatterplot matrix for the algae variables again all variables have a markedly skewed distribution and in all cases skewed to the right. There is mild to minimal correlation between variables.

Initial summary statistics of the chemical data subsets were calculated and indicated several chemical variables, ammonia, nitrates, chloride and phosphate all had large standard deviations greater than that of the mean also indicating the effect of outliers. This was also the case for all algae variables when summary statistics were calculated. Outliers that generally appeared significantly deviated from the data on assessment of scatterplots were removed. Three larger outliers evident in the plots of ammonia and one large outlier in nitrates were removed (cases 20, 34, 76, 121). In algae data one larger outlier evident in the A3 plots and one evident in the A2 plots were removed (cases 55, 124). On box plot representations of individual variables and on calculation using multivariate techniques a significant number of outliers were determined, more than the number removed from the data set. Given that these observations on a scatterplot did not appear as significantly separated from the data and due to concern for removing significant amounts of data or significant measurements from the data sets, these observations were kept for analysis.

After data cleaning 178 cases remained. Two new data subsets were created for chemical variables and algae variables. The table below indicates the number of rivers present in the data set based on size.

| **Large** | **Medium** | **Small** |
|---|---|---|
| 42 | 80 | 56 |

The next table below indicates the number of rivers measured in each season.

| **Autumn** | **Spring** | **Summer** | **Winter** |
|---|---|---|---|
| 35 | 45 | 43 | 55 |

As we can see above there are some differences between our sample sizes which may affect our interpretation of results particularly with river size. There are twice as many medium rivers as there are

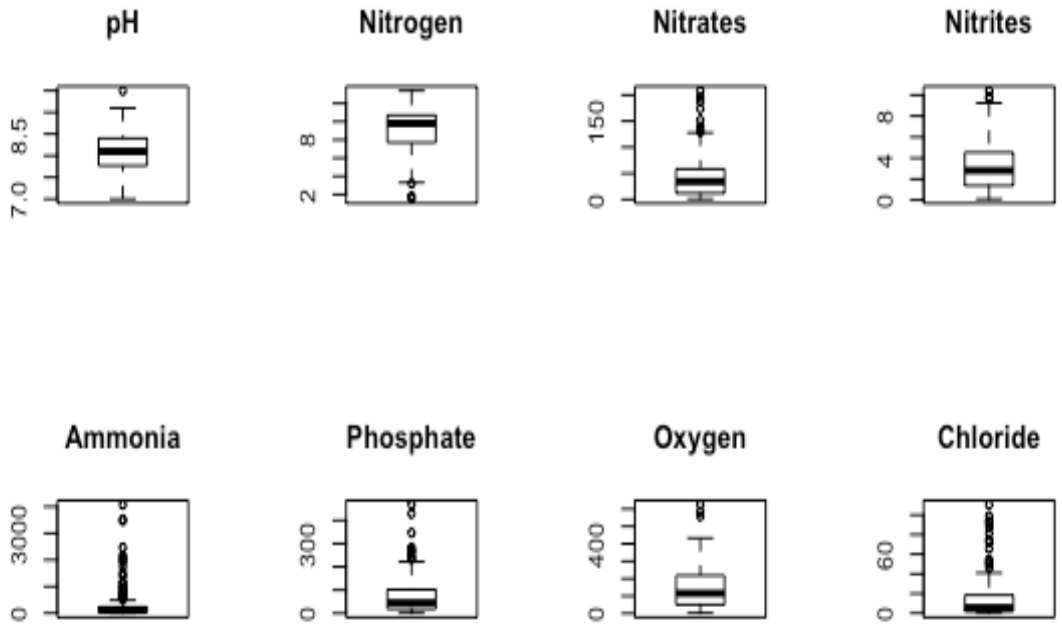
large. There is also some difference between the number of rivers sampled in autumn compared to winter which may also influence our analysis.

**Summary Statistics of Chemical Variables:**

| pH | Nitrogen | Nitrates | Nitrites | Ammonia |
|---|---|---|---|---|
| Min. :7.000 | Min. : 1.500 | Min. : 0.80 | Min. : 0.102 | Min. : 5.8 |
| 1st Qu.:7.772 | 1st Qu.: 7.700 | 1st Qu.: 11.83 | 1st Qu.: 1.378 | 1st Qu.: 49.9 |
| Median :8.100 | Median : 9.800 | Median : 34.27 | Median : 2.812 | Median : 111.4 |
| Mean :8.087 | Mean : 9.073 | Mean : 43.18 | Mean : 3.161 | Mean : 314.5 |
| 3rd Qu.:8.400 | 3rd Qu.:10.700 | 3rd Qu.: 58.21 | 3rd Qu.: 4.528 | 3rd Qu.: 228.5 |
| Max. :9.500 | Max. :13.400 | Max. :208.36 | Max. :10.416 | Max. :4073.3 |

| Phosphate1 | Oxygen1 | Chloride1 |
|---|---|---|
| Min. : 1.333 | Min. : 2.50 | Min. : 0.200 |
| 1st Qu.: 18.778 | 1st Qu.: 51.06 | 1st Qu.: 2.025 |
| Median : 45.450 | Median :115.60 | Median : 5.614 |
| Mean : 74.124 | Mean :141.30 | Mean : 14.237 |
| 3rd Qu.:101.750 | 3rd Qu.:217.38 | 3rd Qu.: 18.360 |
| Max. :467.500 | Max. :624.73 | Max. :110.456 |

*Fig 3. Boxplots of the distribution of individual chemical variables*




As was evident in the scatterplots above prior to outlier removal, apart from pH and nitrogen all the distributions for the chemical variables are skewed to the right. Ammonia in particular has a significant right skew with multiple outliers and a range of 4067mg/L. With a median of 111mg/L we can see there are some rivers that had significantly higher readings of ammonia. Chloride, phosphate and nitrates also had a significantly right skewed distribution and multiple outliers to the right. Nitrogen had a left skewed distribution and pH was roughly symmetrical. Given the distributions are skewed they are better summarised with the five summary statistics of median, Q1, Q3, maximum and minimum.

Individual distributions should be univariate normal to be multivariate normal. The chemical variables shown above and the algae variables (not shown) in the final data set have a skewed distribution. Thus, we must be mindful of this in our analysis. However, in this case, multivariate normality of our variables is assumed. In addition, for analysis a degree of correlation is required between variables.

**Cluster Analysis and Dendrogram**

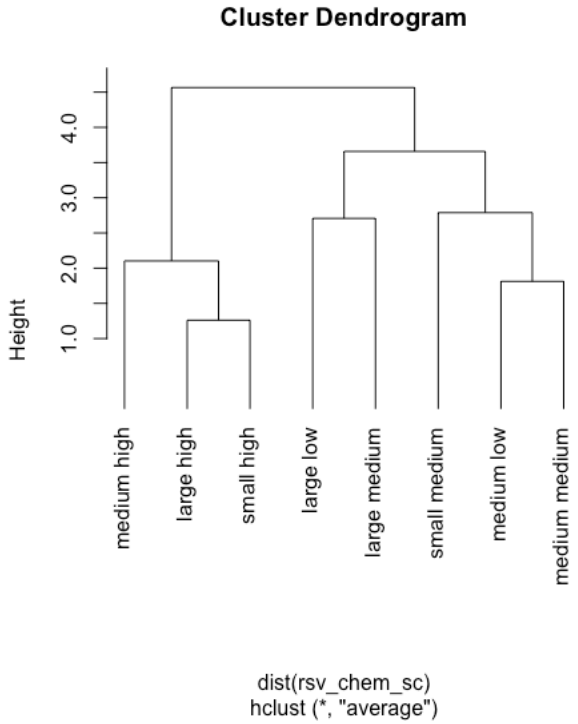
Cluster analysis was performed to investigate relationships between the combination of river size and velocity and the chemical variables. Firstly, as part of analysis a new categorical variable was created combining river size and velocity. A subset of the data was then created with this new variable and the chemical variables. A frequency table was the created to assess the sizes of each sample shown below.

| Large High | Medium High | Small High | Large low | Medium Low | Small Low | Large Medium | Medium Medium | Small Medium |
|---|---|---|---|---|---|---|---|---|
| 6 | 33 | 36 | 16 | 15 | 0 | 20 | 32 | 20 |

As we can see there are no small rivers of low velocity in our data set which may influence of the interpretation of our analysis.

Cluster analysis was performed on the average of each chemical variable for each river size/velocity group. Due to a difference in variable units and also range, variables were scaled prior to analysis. Several forms of cluster analysis were conducted, nearest neighbour linkage and group average, both with Euclidian and Manhattan distances. Of the methods tried, two clusters where determined with each analysis. The group averaging method for both distances appeared to give better results with greater distance between the two clusters and sensible grouping particularly with Euclidian distances. The dendrogram representing this analysis with Euclidian distances is below.

*Fig 4. Cluster Dendrogram*



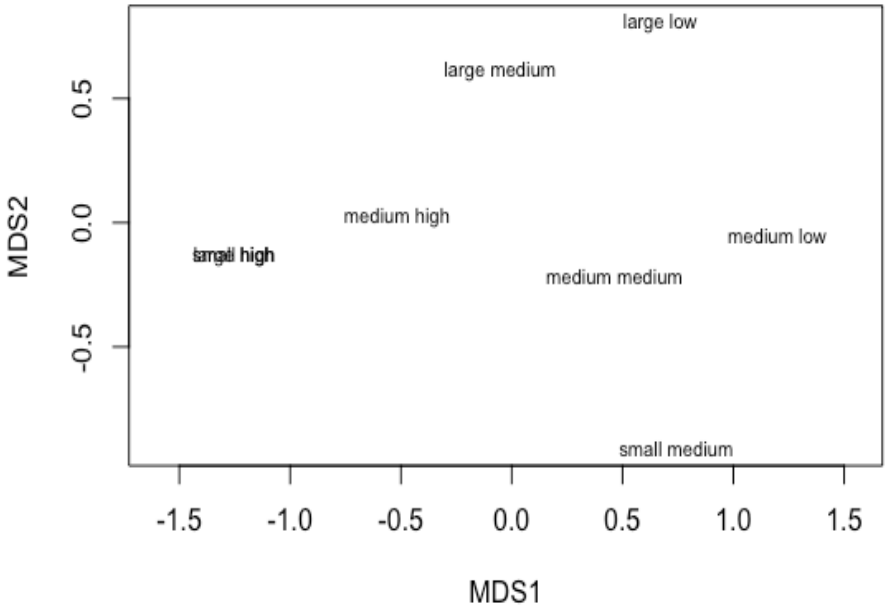

We can see from the dendrogram that river groups are grouping across chemical variables with respect to velocity rather than size. The first cluster contains only rivers of high velocity. The second cluster contains rivers of low and medium velocity.

Analysis was based on group averages and some groups were significantly different in sample size compared to others which may influence our results. We should also consider the absence of the third low velocity group, small river/low velocity and that the low velocity groups had a total number of cases that was half of the other two groups.

**MDS Analysis and Ordination Plot**

Both metric and non-metric forms of MDS analysis were performed. Metric MDS produced a reasonable goodness of fit with three dimensions however, with a very low stress level achieved with non-metric MDS in 2 dimensions, this method was chosen. Again, in the MDS ordination plot we can see that rivers of similar velocity are grouping together. Both these analyses highlight a relationship between levels of chemical variables and river velocity.

*Fig 5. MDS Ordination Plot*




**Task 2:**

To investigate differences in river health between seasons based on chemical variables and algal growth, the MANOVA technique was used. The assumptions for MANOVA require that the variables are multivariate normal and assume equal covariance matrices across populations. MANOVA does require a degree of correlation between variables. Prior to analysis the correlation matrix was examined between chemical and algae variables. There was a mix of correlations across variables with some variables having very low correlation and others small to moderate correlation. Two variables, phosphate and oxygen had a high positive correlation. Given that there is some correlation present, it was deemed appropriate to proceed with MANOVA for analysis however bearing in mind that some variables had very low levels of correlation.

*Correlation matrix*

| | pH1 | nitrogen1 | nitrates1 | nitrites1 | ammonia1 | phosphate1 | oxygen1 | chloride1 | A1 |
|---|---|---|---|---|---|---|---|---|---|
| pH1 | 1.000 | -0.121 | 0.189 | -0.139 | -0.150 | 0.134 | 0.142 | 0.431 | -0.174 |
| nitrogen1 | - 0.121 | 1.000 | -0.301 | 0.089 | -0.168 | -0.328 | -0.412 | -0.147 | 0.243 |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| nitrates1 | 0.189 | -0.301 | 1.000 | 0.368 | 0.241 | 0.436 | 0.504 | 0.189 | -0.403 |
| nitrites1 | -0.139 | 0.089 | 0.368 | 1.000 | 0.257 | 0.282 | 0.378 | 0.056 | -0.360 |
| ammonia1 | -0.150 | -0.168 | 0.241 | 0.257 | 1.000 | 0.308 | 0.389 | 0.020 | -0.171 |
| phosphate1 | 0.134 | -0.328 | 0.436 | 0.282 | 0.308 | 1.000 | 0.887 | 0.145 | -0.413 |
| oxygen1 | 0.142 | -0.412 | 0.504 | 0.378 | 0.389 | 0.887 | 1.000 | 0.302 | -0.483 |
| chloride1 | 0.431 | -0.147 | 0.189 | 0.056 | 0.020 | 0.145 | 0.302 | 1.000 | -0.279 |
| A1 | -0.174 | 0.243 | -0.403 | -0.360 | -0.171 | -0.413 | -0.483 | -0.279 | 1.000 |
| A2 | 0.365 | -0.133 | 0.150 | 0.113 | -0.062 | 0.233 | 0.250 | 0.447 | -0.278 |
| A3 | 0.032 | -0.286 | 0.137 | -0.069 | -0.183 | 0.065 | 0.102 | -0.057 | -0.118 |
| A4 | -0.243 | -0.405 | 0.100 | -0.021 | 0.253 | 0.111 | 0.201 | -0.096 | -0.076 |
| A5 | -0.120 | 0.208 | 0.196 | 0.357 | 0.137 | 0.162 | 0.195 | -0.083 | -0.282 |
| A6 | -0.186 | 0.180 | 0.227 | 0.347 | 0.291 | 0.033 | 0.087 | 0.003 | -0.273 |
| A7 | -0.179 | -0.113 | -0.054 | 0.175 | 0.040 | 0.036 | 0.096 | 0.015 | -0.197 |

| | A2 | A3 | A4 | A5 | A6 | A7 |
|---|---|---|---|---|---|---|
| pH1 | 0.365 | 0.032 | -0.243 | -0.120 | -0.186 | -0.179 |
| nitrogen1 | -0.133 | -0.286 | -0.405 | 0.208 | 0.180 | -0.113 |
| nitrates1 | 0.150 | 0.137 | 0.100 | 0.196 | 0.227 | -0.054 |
| nitrites1 | 0.113 | -0.069 | -0.021 | 0.357 | 0.347 | 0.175 |
| ammonia1 | -0.062 | -0.183 | 0.253 | 0.137 | 0.291 | 0.040 |
| phosphate1 | 0.233 | 0.065 | 0.111 | 0.162 | 0.033 | 0.036 |
| oxygen1 | 0.250 | 0.102 | 0.201 | 0.195 | 0.087 | 0.096 |
| chloride1 | 0.447 | -0.057 | -0.096 | -0.083 | 0.003 | 0.015 |
| A1 | -0.278 | -0.118 | -0.076 | -0.282 | -0.273 | -0.197 |
| A2 | 1.000 | 0.061 | -0.214 | -0.200 | -0.153 | 0.036 |
| A3 | 0.061 | 1.000 | 0.122 | -0.137 | -0.202 | 0.058 |
| A4 | -0.214 | 0.122 | 1.000 | -0.095 | -0.082 | 0.146 |
| A5 | -0.200 | -0.137 | -0.095 | 1.000 | 0.384 | -0.055 |
| A6 | -0.153 | -0.202 | -0.082 | 0.384 | 1.000 | -0.033 |
| A7 | 0.036 | 0.058 | 0.146 | -0.055 | -0.033 | 1.000 |

Four MANOVA statistics were calculated and the resulting values and p values are in the table below.

*MANOVA Statistics*

| **Statistic** | **Value** | **P value** |
|---|---|---|
| Wilk's Lambda | 0.683 | 0.034 |
| Roy's Largest Root | 0.33471 | 2.044e-5 |
| Pillai's Trace | 0.34097 | 0.054 |
| Lawes-Hotelling | 0.42971 | 0.020 |

Of the four MANOVA test statistics calculated, three statistics were significant at a level of $p < 0.05$, Wilks Lambda, Roy's Largest Root and Lawes-Hotelling. These results support a significant difference in river health between at least two seasons. Pillai's trace however was only significant to a level of $p<0.1$. We are assuming that variables are multivariate normal which is required for the above tests and although as discussed above there are some differences in sample sizes between seasons these statistics should be robust to this with the sample size we have. Pillai's however is more robust if there are variations from MVN or differences in the population covariance matrix which may be relevant here thus making it more appropriate to look at this result. Given that Pillai's is significant to a level of $p < 0.1$ does not place a lot of strength on that there is difference between river health in at least two seasons based on chemical and algae variables.

**Task 3:**

The next step in analysis was to determine if season could be predicted based on our chemical and algae variables. The data was first subset and rivers measured only in Spring, Summer and Autumn were included for this analysis. The sample sizes of these groups were all similar. Data was further subset into

training and test sets. Discriminant function analysis (DFA) was then performed on the training set of data and the model was applied to the test set to determine how effective it was in predicting season. DFA assumes that data is multivariate normal within groups and that the covariance matrix is the same for these groups.

**DFA Results:**

Coefficients of linear discriminants:

| | LD1 | LD2 |
|---|---|---|
| pH1 | -0.3306575503 | -0.4488750326 |
| nitrogen1 | -0.2159118464 | -0.2883414220 |
| nitrates1 | 0.0017081501 | -0.0163040449 |
| nitrites1 | -0.1497891964 | 0.0566064956 |
| ammonia1 | -0.0008077803 | 0.0001560751 |
| phosphate1 | -0.0053767792 | 0.0216697789 |
| oxygen1 | 0.0082960320 | -0.0118097996 |
| chloride1 | -0.0040380349 | 0.0113780406 |
| A1 | -0.0115753267 | 0.0021141404 |
| A2 | -0.0386847790 | -0.0115779432 |
| A3 | 0.0660701979 | -0.0122472898 |
| A4 | -0.0656941968 | 0.1405555258 |
| A5 | -0.0206353325 | 0.0373011914 |
| A6 | -0.0539042931 | 0.0254975486 |
| A7 | 0.0108856550 | -0.0918337340 |

Proportion of trace:

| LD1 | LD2 |
|---|---|
| 0.7365 | 0.2635 |

DFA produced two discriminant functions to represent our chemical variables. When initially looking at the proportion of trace results we can see that our discriminant functions do represent a good proportion of the between group variation with discriminant function 1 (DF1) representing 73.65% of between group variation and DF2 representing an additional 26.35%. However we must be cautious with these results as when looking at the weighting of variables on these functions. The highest weighting is of pH at -0.449 on DF2 and apart from variables pH and nitrogen with small to moderate weightings on both discriminant functions, the remaining variables all have very low weights. This indicates weak relationships of these variables with the discriminant functions.

The model produced with the training of set data was then applied to the test set of data to see how effective the model is. The results are displayed below.

*Results of DFA model predictions of seasonal groupings of river observations based on chemical and algae variables*

| | **Autumn** | **Spring** | **Summer** |
|---|---|---|---|
| **Autumn** | 5 | 0 | 3 |
| **Spring** | 1 | 5 | 5 |
| **Summer** | 7 | 2 | 1 |

We can see from the results above that the model was poor in predicting seasonal groups of river observations. Of the rivers measured in summer the model only correctly predicted the season of one river thus only 10% of cases were correctly classified. The model was improved with the group autumn where 62.5% of cases were correctly classified. With the group spring 45.45% of cases were correctly classified. So overall the model was not very effective indicating in this case it was difficult to correctly predict the season in which river observations were taken based on chemical and algae variables. It is

important to consider that if our data deviated significantly from multivariate normal within groups this may affect our results.

If we look at a plot of our data with colour differentiating seasons, we can see that there isn't a lot of separation between seasons across the chemical or the algal variables.

*Fig 6 Splom plot of the chemical variables with observations coloured by season*

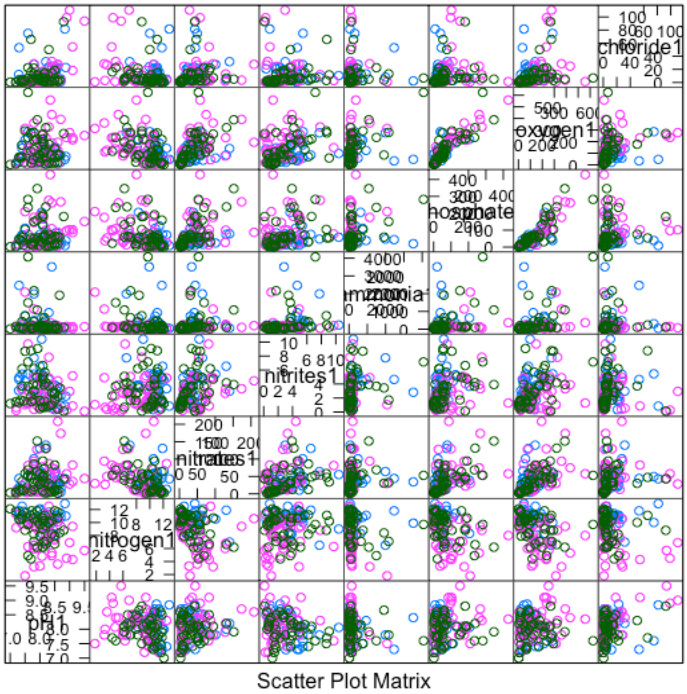
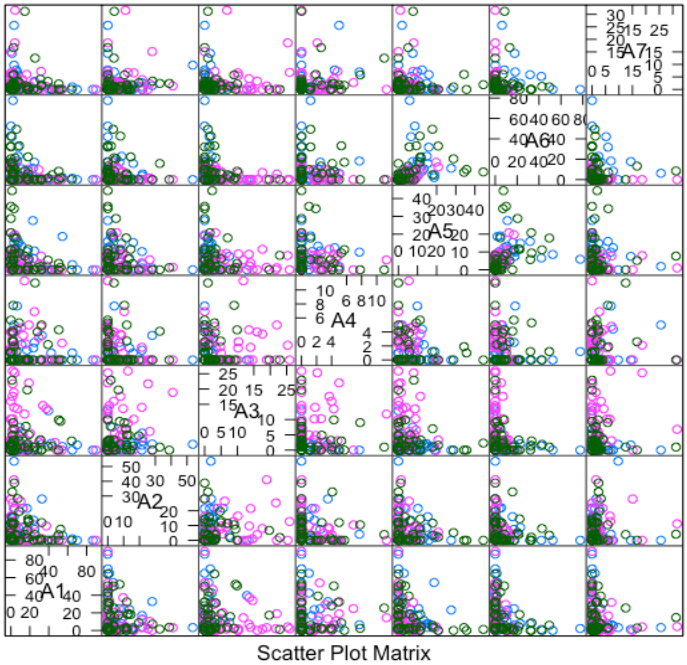



*Fig 7 Splom plot of algae variables with observations coloured by season*



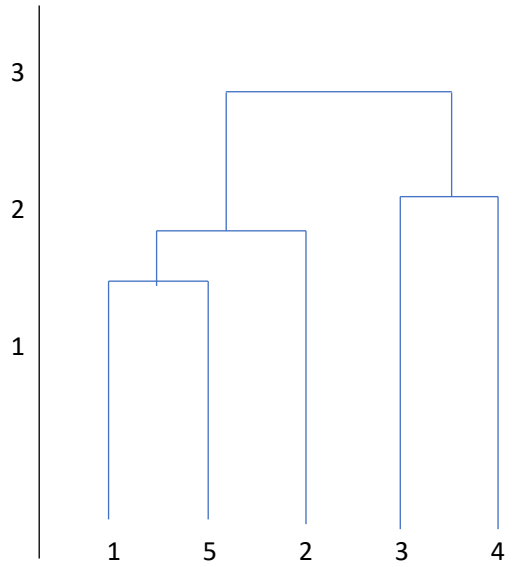
**Part B:**

**Question 1:**

| Feature | MANOVA | PCA | FA | DFA | CCA | CA | MDS |
|---|---|---|---|---|---|---|---|
| **Eigen Analysis** | ✓ | ✓ | ✓ | ✓ | ✓ | × | × |
| **Distance Matrix** | × | × | × | × | × | ✓ | ✓ |
| **Data/Dimension Reduction** | × | ✓ | ✓ | ✓ | × | ✓ | ✓ |
| **Classification** | × | × | × | ✓ | × | ✓ | × |
| **Can be used to identify structure/clusters** | × | × | × | ✓ | × | ✓ | × |
| **Need independent *a priori* categorical variable (s)** | ✓ | × | × | × | × | × | × |
| **Ordination Method** | × | ✓ | ✓ | × | × | × | ✓ |

**Question 2:**

Nearest neighbour dendrogram –




**Question 3:**

Euclidian distance between individuals 1 and 2 for variables X1 and X2

$$d_{12} = [(-0.46 + 1.41)^2 + (-0.46 + 1.79)^2]^{\frac{1}{2}}$$
$$d_{12} = [(0.95)^2 + (1.33)^2]^{1/2}$$
$$d_{12} = 1.634$$

**Question 4:**
There are limitations on multivariate analysis. Sample size is an important limitation. Larger sample sizes improve our ability to interpret data and smaller sample sizes result in larger chances of errors in interpretation. Missing values can be a problem in multivariate analysis as any individuals with missing values from variables must be removed from our analysis which can greatly reduce our sample size. Extreme values can also influence our methods and thus these may also need to be removed again reducing our sample size. Data cleaning is an important first step in analysis.

Multivariate analysis is also affected by correlation. We do need a degree of correlation between variables however variables can be too correlated.

Another important consideration with multivariate analysis when applying different techniques is that exact results are difficult to reproduce and being clear about how results are calculated is important.

**Question 5:**
An eigenvalue is a number associated with a square matrix. When this number is multiplied by the identity matrix and our original matrix is the subtracted from this new matrix, the determinant of the final calculated matrix is zero . A matrix may have several eigenvalues and each value has this property. The maximum number of eigenvalues for this matrix is the number of rows or columns of the matrix (n).

Eigenvectors are a set of values associated with an eigenvalue. If this vector of values is multiplied by our eigenvalue we achieve the same result as if we multiplied our original matrix by the eigenvector. This relationship between an eigenvalue and its associated eigenvectors is generally represented by the equation below where A = the original matrix, x = the eigenvector and $\lambda$ = the eigenvalue.

$$Ax = \lambda x$$

**Question 6:**
Using parallel analysis, we would choose to interpret the first two factors. The first four factors have eigenvalues greater than one however only the first two factors have eigenvalues greater than the 95$^{th}$ percentile thus we would choose to interpret just these two.