**DATA PREPARTION AND ANALYSIS INSTRUCTION**

**for the manuscript**

**Phytoplankton group classification by integrating trait information and observed environmental thresholds**

Hoang Vuong Danga, \*, Kermode Stephanieb, Peisheng Huanga, Cayelan C. Careyc, Matthew R. Hipseya

aCentre for Water and Spatial Science, UWA School of Agriculture and Environment, The University of Western Australia, Crawley, WA, Australia

bWaterNSW, Parramatta, NSW, Australia

cDepartment of Biological Sciences, Virginia Tech, Blacksburg, VA, USA

Corresponding author: Hoang Vuong Dang ([hoangvuong.dang@research.uwa.edu.au](mailto:hoangvuong.dang@research.uwa.edu.au))

1. **Data, code and software installation**

Three main elements were sourced for this case study:

* **Data**: Phytoplankton and water quality data were provided by **Sydney Water Corporation (SWC)**. The datasets were uploaded to Zenodo.org via DOI link: **10.5281/zenodo.15209350.** The datasets are **not openly available for direct download** due to access restrictions. However, they can be made available upon reasonable request. To request access, please use the **“Request access”** on this Zenodo and submit **data sharing agreement form** to the data sharing license conditions.
* **Code**: R scripts are available on GitHub at the link provided in the manuscript.
* Software: Microsoft Excel, R/RStudio were used.

1. **Data analysis**

With the date items in place, users can proceed with Data Preprocessing (2.1) and the Data Analysis (2.2) from Stage 1 through Stage 4.

* 1. **Data Preprocessing**
* **Phytoplankton Data**: The phytoplankton dataset is provided in Excel format (.xlsx or .csv) under the file name **Data\_HN\_Phytoplankton\_SWC\_13\_3\_2025.csv**.

Data example in the below table:

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Algal Group Description** | **Genus Name** | **Toxic** | **Calendar Date** | **Year** | **Sampling Point Identifier** | **Site Code** | **Cells/mL** | **ASU** | **BioVolume (mm3/L)** |
| Bacillariophyta (Diatom) | Cocconeis | Non-Toxic | 3/01/2006 | 2006 | N42\_SURFAC | N42 | 7.00 | 5.46 | 0.01 |
| Bacillariophyta (Diatom) | Cymbella | Non-Toxic | 3/01/2006 | 2006 | N42\_SURFAC | N42 | 3.00 | 1.44 | 0.00 |
| Bacillariophyta (Diatom) | Gomphonema | Non-Toxic | 3/01/2006 | 2006 | N42\_SURFAC | N42 | 7.00 | 9.77 | 0.02 |
| Bacillariophyta (Diatom) | Navicula | Non-Toxic | 3/01/2006 | 2006 | N42\_SURFAC | N42 | 8.00 | 3.29 | 0.00 |
| Bacillariophyta (Diatom) | Synedra | Non-Toxic | 3/01/2006 | 2006 | N42\_SURFAC | N42 | 5.00 | 3.03 | 0.00 |
| Chlorophyta (Green) | Ankistrodesmus | Non-Toxic | 3/01/2006 | 2006 | N42\_SURFAC | N42 | 74.00 | 23.61 | 0.01 |
| Chlorophyta (Green) | Chlamydomonas | Non-Toxic | 3/01/2006 | 2006 | N42\_SURFAC | N42 | 34.00 | 5.81 | 0.01 |
| Chlorophyta (Green) | Closterium | Non-Toxic | 3/01/2006 | 2006 | N42\_SURFAC | N42 | 2.00 | 4.11 | 0.00 |
| Chlorophyta (Green) | Coelastrum | Non-Toxic | 3/01/2006 | 2006 | N42\_SURFAC | N42 | 10.00 | 1.37 | 0.00 |
| Chlorophyta (Green) | Scenedesmus | Non-Toxic | 3/01/2006 | 2006 | N42\_SURFAC | N42 | 21.00 | 1.45 | 0.00 |
| Chlorophyta (Green) | Sphaerocystis | Non-Toxic | 3/01/2006 | 2006 | N42\_SURFAC | N42 | 47.00 | 3.62 | 0.00 |
| Chlorophyta (Green) | Staurastrum | Non-Toxic | 3/01/2006 | 2006 | N42\_SURFAC | N42 | 2.00 | 19.71 | 0.01 |
| … | … | … | … | … | … | … | … | … | … |

* **Water Quality Data**: This dataset is provided in Excel (csv) format under the file name **Data\_HN\_water\_quality\_SWC\_13\_3\_2025.csv**.
* **Next Step – Prepare Input File for Multi-Correlation Analysis**:
  + Open Data\_HN\_Phytoplankton\_SWC\_13\_3\_2025.csv in Excel.
  + Use a **Pivot Table** to calculate the **total daily biovolume** by summing the biovolume of all species for each **date** and **site**. (**Rows**: Date, Site; **Columns**: Site (optional, if you want separate columns for each site), **Values**: Sum of Biovolume (BIOVOL).
  + Export the result as **Total\_BIOVOL**.csv.

|  |  |  |
| --- | --- | --- |
| **Site Code** | **Calendar Date** | **BIOVOL** |
| N18 | 29/07/2008 | 2.374012882 |
| N18 | 12/08/2008 | 1.989492 |
| N18 | 3/09/2008 | 1.435298701 |
| N18 | 23/09/2008 | 1.988031257 |
| N18 | 14/10/2008 | 1.927863749 |
| N18 | 5/11/2008 | 2.066298704 |
| N18 | 25/11/2008 | 1.301655155 |
| N18 | 16/12/2008 | 2.98333389 |
| N18 | 6/01/2009 | 2.549035051 |
| N18 | 27/01/2009 | 3.149 |
| N18 | 17/02/2009 | 11.52599229 |
| N18 | 10/03/2009 | 1.666555795 |
| N18 | 1/04/2009 | 1.9569595 |
| N18 | 21/04/2009 | 0.775864032 |
| N18 | 12/05/2009 | 1.222236847 |
| N18 | 24/06/2009 | 0.694767527 |
| … | … | … |

* + **Then add Water Quality Data to the Table**: Use the **VLOOKUP** (or **XLOOKUP** for newer Excel versions) function to match and import environmental variables from **Data\_HN\_water\_quality\_SWC\_13\_3\_2025.csv** into the **Total\_BIOVOL.csv** table.
  + **Note**: Ensure both files (Total\_BIOVOL.csv and Data\_HN\_water\_quality\_SWC\_13\_3\_2025.csv) have matching columns: **Calendar Date** and **Site Code.** User might need to covert Calendar Date to Datenum, then add Site to create column Site\_Datenum for more accurate adding function.
  + After all parameters are added, save the updated file as **MCA\_data.csv (**See below table).

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Site Code** | **Calendar Date** | **BIOVOL** | **Datenum** | **Site\_Datenum** | **TN** | **TP** | **SAL** | **TEMP** | **AMM** | **NIT** | **SIL** | **TURB** | **CHLA** |
| N18 | 29/07/2008 | 2.374012882 | 39658 | N1839658 | 6.00E-01 | 3.40E-02 | NaN | 1.22E+01 | 1.00E-02 | 2.50E-01 | 26 | 1.19E+01 | 26 |
| N18 | 12/08/2008 | 1.989492 | 39672 | N1839672 | 4.80E-01 | 1.70E-02 | NaN | 1.21E+01 | 5.00E-03 | 1.60E-01 | 1.71E+01 | 5.12E+00 | 1.71E+01 |
| N18 | 3/09/2008 | 1.435298701 | 39694 | N1839694 | 3.70E-01 | 2.60E-02 | NaN | 1.38E+01 | 5.00E-03 | 8.00E-02 | 1.24E+01 | 6.30E+00 | 1.24E+01 |
| N18 | 23/09/2008 | 1.988031257 | 39714 | N1839714 | 5.90E-01 | 2.10E-02 | NaN | 1.79E+01 | 1.00E-02 | 3.00E-01 | 1.73E+01 | 7.44E+00 | 1.73E+01 |
| N18 | 14/10/2008 | 1.927863749 | 39735 | N1839735 | 4.70E-01 | 2.60E-02 | NaN | 2.01E+01 | 4.00E-02 | 1.40E-01 | 2.66E+01 | 6.12E+00 | 2.66E+01 |
| N18 | 5/11/2008 | 2.066298704 | 39757 | N1839757 | 3.20E-01 | 2.00E-02 | NaN | 2.24E+01 | 5.00E-03 | 5.00E-03 | 2.04E+01 | 4.06E+00 | 2.04E+01 |
| N18 | 25/11/2008 | 1.301655155 | 39777 | N1839777 | 2.90E-01 | 1.70E-02 | NaN | 23 | 5.00E-03 | 4.00E-02 | 1.34E+01 | 5.50E+00 | 1.34E+01 |
| N18 | 16/12/2008 | 2.98333389 | 39798 | N1839798 | 4.40E-01 | 3.40E-02 | NaN | 2.45E+01 | 5.00E-03 | 5.00E-03 | 2.92E+01 | 8.48E+00 | 2.92E+01 |
| N18 | 6/01/2009 | 2.549035051 | 39819 | N1839819 | 3.60E-01 | 3.60E-02 | NaN | 2.63E+01 | 5.00E-03 | 5.00E-03 | 1.68E+01 | 9.66E+00 | 1.68E+01 |
| N18 | 27/01/2009 | 3.149 | 39840 | N1839840 | 3.30E-01 | 2.40E-02 | NaN | 2.71E+01 | 5.00E-03 | 3.00E-02 | 1.36E+01 | 1.12E+01 | 1.36E+01 |
| N18 | 17/02/2009 | 11.52599229 | 39861 | N1839861 | 5.30E-01 | 3.60E-02 | NaN | 2.43E+01 | 3.00E-02 | 4.00E-02 | 4.33E+01 | 5.36E+00 | 4.33E+01 |
| N18 | 10/03/2009 | 1.666555795 | 39882 | N1839882 | 5.30E-01 | 4.60E-02 | NaN | 2.47E+01 | 1.00E-02 | 8.00E-02 | 2.17E+01 | 1.54E+01 | 2.17E+01 |
| N18 | 1/04/2009 | 1.9569595 | 39904 | N1839904 | 4.00E-01 | 3.20E-02 | NaN | 2.46E+01 | 1.00E-02 | 1.00E-02 | 1.79E+01 | 7.48E+00 | 1.79E+01 |
| N18 | 21/04/2009 | 0.775864032 | 39924 | N1839924 | 5.50E-01 | 4.60E-02 | NaN | 2.05E+01 | 3.00E-02 | 1.90E-01 | 1.25E+01 | 1.07E+01 | 1.25E+01 |
| N18 | 12/05/2009 | 1.222236847 | 39945 | N1839945 | 4.70E-01 | 5.30E-02 | NaN | 1.82E+01 | 5.00E-03 | 6.00E-02 | 2.66E+01 | 1.44E+01 | 2.66E+01 |
| N18 | 24/06/2009 | 0.694767527 | 39988 | N1839988 | 4.30E-01 | 3.40E-02 | NaN | 1.41E+01 | 5.00E-03 | 1.80E-01 | 1.13E+01 | 1.25E+01 | 1.13E+01 |

* + In your table, create a new column for **TN:TP molecular ratio** using the following formula: TNTP = (TN / 14) / (TP / 31).
  + Add a new column called **Site\_Quarter\_Year (SQY),** then use the IF function (or TEXT function) to assign the quarter and year for each site. For example: “N182006Q1” is from site N18, Jan to March 2006 in 2006.
  + **Use Pivot Table to Calculate Mean**: Set **SQY** as the **Row**, Set the **values** to calculate the **mean** of **BIOVOL** and all other environmental variables (e.g., TN, TP, Temperature, etc.).
  + **Output File**: Save the output as **MCA\_data.csv (**See below table).

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| SQY | TN | TP | TNTP | SAL | TEMP | AMM | NIT | FRP | TURB | SIL | CHLA | VOL |
| N422006 Q3 | 0.82 | 0.018 | 100.873 | 0.154 | 16.3 | 0.005 | 0.56 | 0.007 | 0.83 | 3.39 | 8.7 | 1.248129 |
| N422006 Q4 | 0.38 | 0.01 | 84.14286 | 0.178 | 22.9 | 0.02 | 0.12 | 0.005 | 0.62 | 1.7 | 3.1 | 0.249592 |
| N422007 Q1 | 0.72 | 0.062 | 25.71429 | 0.143 | 23.8 | 0.01 | 0.27 | 0.014 | 3.01 | 4.96 | 12.3 | 1.190617 |
| N422007 Q2 | 0.61 | 0.01 | 135.0714 | 0.115 | 13.8 | 0.01 | 0.35 | 0.004 | 3.97 | 1.62 | 3.6 | 0.184 |
| N422007 Q3 | 0.81 | 0.029 | 61.55187 | 0.1285 | 14.55 | 0.0225 | 0.43 | 0.0085 | 2.365 | 3.39 | 19.25 | 2.142718 |
| N422008 Q1 | 0.51 | 0.028 | 40.33163 | 0.102 | 25.7 | 0.01 | 0.16 | 0.009 | 4.72 | 5.82 | 11.4 | 2.099816 |
| N422009 Q1 | 0.48 | 0.019 | 55.93985 | 0.0648 | 23.2 | 0.03 | 0.17 | 0.008 | 2.54 | 1.24 | 5.5 | 0.9303 |
| N422010 Q1 | 0.45 | 0.022 | 45.29221 | 0.121 | 25.9 | 0.01 | 0.18 | 0.008 | 2.33 | 2.46 | 8.4 | 1.74763 |
| … | … | … | … | … | … | … | … | … | … | … | … | … |

* **Prepare Input Data for Principal Component Analysis (PCA)**:
  + **Open** the **Data\_HN\_Phytoplankton\_SWC\_13\_3\_2025.csv** (original file) and **Data\_HN\_water\_quality\_SWC\_13\_3\_2025.csv** in Excel.
  + **Join the Water Quality Data to Phytoplankton Data**: Use the **H\_VLOOKUP** (or VLOOKUP if using a standard lookup function) to merge the water quality data with the phytoplankton data based on **matching Date** and **Site** columns (Similar to join data for MCA). The resulting table should now contain both the phytoplankton data (from Data\_HN\_Phytoplankton\_SWC\_13\_3\_2025.csv) and water quality data (from Data\_HN\_water\_quality\_SWC\_13\_3\_2025.csv) for each matching **Date** and **Site**.
  + **Output File**: Save the merged dataset as **PHY\_WQ\_dataset\_PCA.csv** for PCA analysis (See below table).

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Calendar Date** | **Year** | **Site Code** | **Cells/mL** | **CellVolume(mm3)** | **Genus Name** | **Algal Group Description** | **TN** | **TP** | **SAL** | **TEMP** | **AMM** | **NIT** | **FRP** | **SIL** | **TURB** |
| 23/03/2006 | 2006 | N75 | 398 | 0.0000002 | Thalassiosira | Bacillariophyta (Diatom) | 4.18 | 0.092 | 0.231 | 22.4 | 0.19 | 2.4 | 0.009 | 2.56 | 4.73 |
| 23/03/2006 | 2006 | N35 | 2 | 0.0000057 | Merotrichia | Chloromonadophyta | NaN | NaN | 0.15 | 24.4 | NaN | NaN | NaN | 1.05 | NaN |
| 23/03/2006 | 2006 | N48 | 7 | 5.67E-06 | Merotrichia | Chloromonadophyta | 1.01 | 0.138 | 0.121 | 22.8 | 0.01 | 0.06 | 0.019 | 44.9 | 13.7 |
| 23/03/2006 | 2006 | N18 | 197 | 6.65E-08 | Actinastrum | Chlorophyta (Green) | NaN | NaN | NaN | 25.5 | NaN | NaN | NaN | 14.1 | NaN |
| 23/03/2006 | 2006 | N48 | 27 | 6.67E-08 | Actinastrum | Chlorophyta (Green) | 1.01 | 0.138 | 0.121 | 22.8 | 0.01 | 0.06 | 0.019 | 44.9 | 13.7 |
| 23/03/2006 | 2006 | N75 | 145 | 6.62E-08 | Actinastrum | Chlorophyta (Green) | 4.18 | 0.092 | 0.231 | 22.4 | 0.19 | 2.4 | 0.009 | 2.56 | 4.73 |
| 23/03/2006 | 2006 | N39 | 14 | 3.14E-07 | Actinotaenium | Chlorophyta (Green) | 0.5 | 0.02 | NaN | NaN | 0.005 | 0.1 | 0.006 | 23.8 | NaN |
| 23/03/2006 | 2006 | N18 | 81 | 1.15E-07 | Ankistrodesmus | Chlorophyta (Green) | NaN | NaN | NaN | 25.5 | NaN | NaN | NaN | 14.1 | NaN |
| 23/03/2006 | 2006 | N26 | 462 | 1.15E-07 | Ankistrodesmus | Chlorophyta (Green) | NaN | NaN | NaN | 25.1 | NaN | NaN | NaN | 37.5 | NaN |
| 23/03/2006 | 2006 | N35 | 29 | 1.14E-07 | Ankistrodesmus | Chlorophyta (Green) | NaN | NaN | 0.15 | 24.4 | NaN | NaN | NaN | 1.05 | NaN |
| 23/03/2006 | 2006 | N39 | 179 | 1.15E-07 | Ankistrodesmus | Chlorophyta (Green) | 0.5 | 0.02 | NaN | NaN | 0.005 | 0.1 | 0.006 | 23.8 | NaN |
| 23/03/2006 | 2006 | N48 | 61 | 1.15E-07 | Ankistrodesmus | Chlorophyta (Green) | 1.01 | 0.138 | 0.121 | 22.8 | 0.01 | 0.06 | 0.019 | 44.9 | 13.7 |
| 23/03/2006 | 2006 | N75 | 36 | 1.14E-07 | Ankistrodesmus | Chlorophyta (Green) | 4.18 | 0.092 | 0.231 | 22.4 | 0.19 | 2.4 | 0.009 | 2.56 | 4.73 |
| 23/03/2006 | 2006 | N18 | 7 | 1.14E-07 | Ankistrodesmus falcatus | Chlorophyta (Green) | NaN | NaN | NaN | 25.5 | NaN | NaN | NaN | 14.1 | NaN |
| 23/03/2006 | 2006 | N26 | 11 | 1.18E-07 | Ankistrodesmus falcatus | Chlorophyta (Green) | NaN | NaN | NaN | 25.1 | NaN | NaN | NaN | 37.5 | NaN |
| 23/03/2006 | 2006 | N35 | 16 | 1.13E-07 | Ankistrodesmus falcatus | Chlorophyta (Green) | NaN | NaN | 0.15 | 24.4 | NaN | NaN | NaN | 1.05 | NaN |
| 23/03/2006 | 2006 | N75 | 5 | 0.00000012 | Ankistrodesmus falcatus | Chlorophyta (Green) | 4.18 | 0.092 | 0.231 | 22.4 | 0.19 | 2.4 | 0.009 | 2.56 | 4.73 |
| 23/03/2006 | 2006 | N39 | 7 | 6.29E-06 | Botryococcus colony | Chlorophyta (Green) | 0.5 | 0.02 | NaN | NaN | 0.005 | 0.1 | 0.006 | 23.8 | NaN |
| 23/03/2006 | 2006 | N75 | 72 | 5.97E-08 | Carteria | Chlorophyta (Green) | 4.18 | 0.092 | 0.231 | 22.4 | 0.19 | 2.4 | 0.009 | 2.56 | 4.73 |
| 23/03/2006 | 2006 | N48 | 61 | 2.15E-07 | Chlamydomonas | Chlorophyta (Green) | 1.01 | 0.138 | 0.121 | 22.8 | 0.01 | 0.06 | 0.019 | 44.9 | 13.7 |
| 23/03/2006 | 2006 | N75 | 18 | 2.17E-07 | Chlamydomonas | Chlorophyta (Green) | 4.18 | 0.092 | 0.231 | 22.4 | 0.19 | 2.4 | 0.009 | 2.56 | 4.73 |
| 23/03/2006 | 2006 | N26 | 23 | 0.0000001 | Chodatella | Chlorophyta (Green) | NaN | NaN | NaN | 25.1 | NaN | NaN | NaN | 37.5 | NaN |
| 23/03/2006 | 2006 | N26 | 4 | 0.0000093 | Closteriopsis | Chlorophyta (Green) | NaN | NaN | NaN | 25.1 | NaN | NaN | NaN | 37.5 | NaN |
| … | … | … | … | … | … | … | … | … | .. | … | … | … | … | … | … |

* **Prepare Input Data for Taxa Indicator Threshold Analysis (TITAN)**:
  + **Convert Phytoplankton Data from Long to Wide Format**: Open **R/RStudio** and run the script **Stage\_1\_Convert\_long\_to\_wide\_phyto\_data.R** to transform the phytoplankton data from long to wide format and calculate **log(abundance + 1)** for each species. Export the result as **Phyto\_HN\_wide\_log.csv.**
  + **Count Values and Remove Low Occurrence Species**: Open **Phyto\_HN\_wide\_log.csv** in Excel., Count the number of values greater than 0 in each column (for each species). Then remove any species (columns) with fewer than **4 occurrences** greater than 0.
  + **Add Water Quality Data**: Use **VLOOKUP** to merge the **water quality data** from **Data\_HN\_water\_quality\_SWC\_13\_3\_2025.csv** into **Phyto\_HN\_wide\_log.csv**. Match the species data with corresponding **Date** and **Site** values.
  + **Save the Final Dataset**: Save the updated file as **HN\_PHY\_WQ\_DATA\_log\_1.csv** (see below table).

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Site** | **Date** | **TN** | **TP** | **TN\_TP\_Ratio** | **SAL** | **TEMP** | **AMM** | **NIT** | **FRP** | **SIL** | **TURB** | **T.Anabaena.aphanizomenoides** | **T.Anabaena.bergii** | **T.Anabaena.circinalis** | **T.Anabaenopsis** | **T.Aphanizomenonaceae** | **T.Heterocapsa** | **T.Microcystis** | … |
| N18 | 10/07/2008 | 0.63 | 0.028 | 49.82 | NA | 12.7 | 0.005 | 0.38 | 0.007 | 17 | 8.15 | 0 | 0 | 0 | 0 | 0 | 0 | 2.39794 | … |
| N18 | 29/07/2008 | 0.6 | 0.034 | 39.08 | NA | 12.2 | 0.01 | 0.25 | 0.006 | 26 | 11.9 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | … |
| N18 | 12/08/2008 | 0.48 | 0.017 | 62.52 | NA | 12.1 | 0.005 | 0.16 | 0.004 | 17.1 | 5.12 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | … |
| N18 | 3/09/2008 | 0.37 | 0.026 | 31.51 | NA | 13.8 | 0.005 | 0.08 | 0.006 | 12.4 | 6.3 | 0 | 0 | 0 | 0 | 0 | 0 | 0.60206 | … |
| N18 | 23/09/2008 | 0.59 | 0.021 | 62.21 | NA | 17.9 | 0.01 | 0.3 | 0.006 | 17.3 | 7.44 | 0 | 0 | 0 | 0 | 0 | 0 | 0.60206 | … |
| N18 | 14/10/2008 | 0.47 | 0.026 | 40.03 | NA | 20.1 | 0.04 | 0.14 | 0.006 | 26.6 | 6.12 | 0 | 0 | 0 | 0 | 0 | 0 | 2.017033 | … |
| N18 | 5/11/2008 | 0.32 | 0.02 | 35.43 | NA | 22.4 | 0.005 | 0.005 | 0.004 | 20.4 | 4.06 | 0 | 0 | 2.716003 | 0 | 0 | 0 | 2.658965 | … |
| N18 | 25/11/2008 | 0.29 | 0.017 | 37.77 | NA | 23 | 0.005 | 0.04 | 0.004 | 13.4 | 5.5 | 0 | 0 | 0 | 0 | 0 | 0 | 2.017033 | … |
| N18 | 16/12/2008 | 0.44 | 0.034 | 28.66 | NA | 24.5 | 0.005 | 0.005 | 0.005 | 29.2 | 8.48 | 0 | 0 | 2.80618 | 0 | 0 | 0 | 2.515874 | … |
| N18 | 6/01/2009 | 0.36 | 0.036 | 22.14 | NA | 26.3 | 0.005 | 0.005 | 0.005 | 16.8 | 9.66 | 2.658965 | 0 | 0 | 0 | 0 | 0 | 2.017033 | … |
| N18 | 27/01/2009 | 0.33 | 0.024 | 30.45 | NA | 27.1 | 0.005 | 0.03 | 0.004 | 13.6 | 11.2 | 0 | 0 | 0 | 0.60206 | 0 | 0 | 0.60206 | … |
| N18 | 17/02/2009 | 0.53 | 0.036 | 32.6 | NA | 24.3 | 0.03 | 0.04 | 0.005 | 43.3 | 5.36 | 0 | 0 | 0 | 2.737193 | 0 | 0 | 2.39794 | … |
| N18 | 10/03/2009 | 0.53 | 0.046 | 25.51 | NA | 24.7 | 0.01 | 0.08 | 0.011 | 21.7 | 15.4 | 0 | 0 | 2.462398 | 2.271842 | 0 | 0 | 2.462398 | … |
| N18 | 1/04/2009 | 0.4 | 0.032 | 27.68 | NA | 24.6 | 0.01 | 0.01 | 0.009 | 17.9 | 7.48 | 0 | 0 | 2.39794 | 2.017033 | 0 | 0 | 2.271842 | … |
| N18 | 21/04/2009 | 0.55 | 0.046 | 26.48 | NA | 20.5 | 0.03 | 0.19 | 0.015 | 12.5 | 10.7 | 0 | 0 | 0 | 0 | 0 | 0 | 2.271842 | … |
| N18 | 12/05/2009 | 0.47 | 0.053 | 19.64 | NA | 18.2 | 0.005 | 0.06 | 0.009 | 26.6 | 14.4 | 0 | 0 | 0 | 0.60206 | 0 | 0 | 2.748188 | … |
| N18 | 24/06/2009 | 0.43 | 0.034 | 28 | NA | 14.1 | 0.005 | 0.18 | 0.011 | 11.3 | 12.5 | 0 | 0 | 0.60206 | 0 | 0 | 0 | 0.60206 | … |
| N18 | 13/07/2009 | 0.41 | 0.024 | 37.83 | 0.103 | 12.8 | 0.01 | 0.14 | 0.007 | 19.8 | 9.69 | 0 | 0 | 0 | 0 | 0 | 0 | 2.583199 | … |
| N18 | 7/08/2009 | 0.36 | 0.02 | 39.86 | 2.114 | 12.9 | 0.005 | 0.1 | 0.004 | 14.3 | 6.97 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | … |
| N18 | 27/08/2009 | 0.36 | 0.016 | 49.82 | 0.696 | 15.4 | 0.005 | 0.06 | 0.004 | 14.2 | 5.62 | 0 | 0 | 2.39794 | 0 | 0 | 0 | 2.271842 | … |
| …. | … | … | … | … | … | … | … | … | … | … | … | … | … | … | … | … | … | … | … |

* 1. **Data Analysis**

Once the data preparation is complete, the input data is ready for the next step. The following analyses are then conducted:

* + 1. **MCA**
  + Open **RStudio**.
  + Input data: **MCA\_data.csv**.
  + Run the R script: **Stage\_2\_MCA\_Final**. This script will show the correlation matrix and correlation plot.
    1. **Conduct Principal Component Analysis (PCA)**
  + Open **RStudio**.
  + Input data: **PHY\_WQ\_dataset\_PCA.csv**.

Run the R script: **Stage\_2\_PCA\_Final**.

* + 1. **TITAN**
  + Open **RStudio**.
  + **Input Data**: **HN\_PHY\_WQ\_DATA\_log\_1.csv**.
  + Open the R Script: **Stage\_3\_TITAN**.
    - **Run lines 1–692** to perform **TITAN** for 10 variables and save the results. This process may take up to 4 days for each variable.
    - **Run lines 694–907** to export **“Change Point 10th and 90th”** and create input files for **K-prototype clustering**: Clustering\_data\_4.csv (with **TN, TP, TNTP, and TEMP**) and Clustering\_data\_all.csv (with **all variables**)
    - **Run lines 913–1062** to perform **TITAN** for 2 species (as per review suggestions) and update the new thresholds for the species in **Clustering\_data\_4.csv**.
    - Next, **check and plot the purity of TITAN** by running code from **lines 1103–1479**.
    - Add **species traits** to **Clustering\_data\_4.csv**. Species traits can be seen in the Supplementary Table S1.
    - The final output file is **K\_prototype\_input.csv** (**Supplementary Table S1).**
    1. **Classification – K-prototype**
  + Open **RStudio**
  + Input data: **K\_prototype\_input.csv; Classification\_trait\_thresholds.RData** (for traits and revised species name after 1st revision)
  + **Open the R Script**: **Stage\_4\_Classification**.
    - **Run the lines 1- 31** to read the input data and assign categorical variables.
    - **Run the line 37 - 42** to select the optimal number of clusters based on the **Elbow plot**.
    - **K-Prototype Classification**: Since **5 clusters** were selected, run the K-prototype classification (line **48**).
    - **Assign new names** to the clusters (line **51 - 61**).
    - **Summary of Clusters**: Use **summary(cluster)** to display the summary of the cluster characteristics.
    - **Visualize Clusters**: Add other thresholds and visualize the clusters using plotting (**run line 62 – 258**).
    - **Test Normality**: To determine which groups differ significantly from each other, first, run **shapiro.test** (line **262 - 283**) to check the normality of the thresholds.
    - Since the p-value of all Shapiro tests is **< 0.05**, indicating non-normal distribution, proceed with non-parametric tests.
    - Then **Kruskal-Wallis Test**: Run **kruskal.test** to assess whether there are statistically significant differences between the thresholds of the clusters (**Line 288 – 304)**.
    - **Post-hoc Tests**: If significant differences are observed, run **kwAllPairsConoverTest** with the **"holm"** method, and **dunnTest** with the **method="bonferroni"** to identify which groups differ (line **306 - 367**).
    - **Taxonomic Group Differences**: Run **kruskal.test** (Line **370 – 599)** to assess differences among taxonomic groups contributing to different functional groups (see **Supplementary Table S5, Text S2**).
    - **Levene's Test**: Use **Levene's Test** (line **601 – 799)** to quantify group homogeneity (see **Supplementary Text S3**).
    1. **Sensitivity Analysis (for Classification)**
  + Open **RStudio**.
  + **Input Data**:
    - **K\_prototype\_input.csv** (same as classification with 4 variables).
    - **K\_prototype\_classification\_input\_all\_trait.csv** (file containing other threshold values same as Clustering\_data\_all.csv).
    - **Trait\_check.csv** (file containing revised species names and traits in Supplementary Table S1).
  + Open the R script: **Stage\_4\_Sensitivity.R**. This script will export results of 4 test in csv file.