Reading and processing data for use in the LTER-LIFE digital twins: the NIOZ jetty data

Karline Soetaert

first version: 01-10-2023; current version: 30 January 2024

This document prepares the biogeochemical data from the NIOZ jetty that are to be used in the LTER life dtPP (digital twin of primary production) package.

# The jetty biogeochemical data (Phillipart, NIOZ)

This is a long-term biogeochemical data set from the surface waters at the NIOZ jetty.

## Raw data

The data are presented in csv format, and are read with function *read.csv*.

jetty <- read.csv("../raw\_data/nioz/HW\_MP\_set.csv")  
summary(jetty[,-(1:2)])

## T S SecchiDepth TN   
## Min. :-2.000 Min. :18.20 Min. :0.15 Min. : 9.91   
## 1st Qu.: 7.125 1st Qu.:27.50 1st Qu.:0.75 1st Qu.: 22.02   
## Median :12.000 Median :29.04 Median :1.05 Median : 30.72   
## Mean :11.772 Mean :28.73 Mean :1.19 Mean : 36.57   
## 3rd Qu.:16.600 3rd Qu.:30.30 3rd Qu.:1.50 3rd Qu.: 46.44   
## Max. :24.400 Max. :34.30 Max. :4.00 Max. :116.45   
## NA's :49 NA's :77 NA's :90 NA's :1266   
## TP PO4 NO3 NO2   
## Min. :0.2020 Min. :0.0000 Min. : -0.031 Min. :0.0150   
## 1st Qu.:0.4870 1st Qu.:0.2320 1st Qu.: 2.672 1st Qu.:0.3832   
## Median :0.6660 Median :0.5110 Median : 11.373 Median :0.6550   
## Mean :0.6884 Mean :0.6376 Mean : 21.952 Mean :0.8465   
## 3rd Qu.:0.8570 3rd Qu.:0.9120 3rd Qu.: 38.050 3rd Qu.:1.0540   
## Max. :1.6510 Max. :3.4500 Max. :136.360 Max. :7.8100   
## NA's :1310 NA's :494 NA's :493 NA's :493   
## NH4 DON DOP Si   
## Min. :-0.084 Min. : 5.954 Min. :-0.0770 Min. :-0.696   
## 1st Qu.: 2.108 1st Qu.:12.780 1st Qu.: 0.2290 1st Qu.: 1.325   
## Median : 4.444 Median :15.162 Median : 0.3260 Median : 3.760   
## Mean : 5.144 Mean :15.661 Mean : 0.3336 Mean : 8.102   
## 3rd Qu.: 7.348 3rd Qu.:18.132 3rd Qu.: 0.4175 3rd Qu.:11.641   
## Max. :21.802 Max. :40.440 Max. : 0.8390 Max. :55.549   
## NA's :493 NA's :1417 NA's :1461 NA's :463   
## DIC TSM Chl X   
## Min. : 20.09 Min. : 2.20 Min. : 0.000 Mode:logical   
## 1st Qu.: 2157.69 1st Qu.: 8.33 1st Qu.: 4.424 NA's:2003   
## Median : 2246.89 Median : 14.30 Median : 7.550   
## Mean : 4665.91 Mean : 19.38 Mean : 11.967   
## 3rd Qu.: 2319.43 3rd Qu.: 23.98 3rd Qu.: 14.119   
## Max. :28291.89 Max. :158.52 Max. :133.428   
## NA's :1671 NA's :1143 NA's :186   
## datetime NPratio   
## Length:2003 Min. :-0.119   
## Class :character 1st Qu.: 5.774   
## Mode :character Median :23.233   
## Mean : Inf   
## 3rd Qu.:66.668   
## Max. : Inf   
## NA's :495

## Cleaning the jetty data

The following changes were done on the raw data:

1. The summary of the raw data show that the DIC data are problematic: the data should lie within 2000-2500, but have many values <100 and also values > 20000.

These data are modified so that their value is always within 2000-2500. Note that the DIC data should be used with caution; they are probably measured with different devices.

1. The negative concentrations (NO3: 2, NH4 2, Si: 41, and DOP 2) are removed.
2. The column called “X” and the NPratio are removed (the latter can easily be estimated).
3. The time is changed to *POSIXct* format.
4. Some columns are renamed
5. A description to each variable, and the station position, is added to the objects attributes.

# Cleaning the aberrant DIC data  
DIC <- jetty$DIC  
DIC[which(DIC < 200)] <- DIC[which(DIC<200)]\*100  
DIC[which(DIC > 20000)] <- DIC[which(DIC>20000)]/10  
jetty$DIC <- DIC  
  
# remove the column called "X"and NPratio   
Jetty <- jetty[,c("SampleID", "datetime", "T", "S", "SecchiDepth", "TN", "TP",  
 "PO4", "NO3", "NO2", "NH4", "DON", "DOP", "Si", "DIC", "TSM", "Chl")]  
  
# time to POSIXct format  
Jetty$datetime <- as.POSIXct(Jetty$datetime)  
  
# rename some columns  
colnames(Jetty)[1:7] <- c("sampleID","datetime", "Temperature", "Salinity", "Secchi", "Ntot", "Ptot")  
colnames(Jetty)[colnames(Jetty) == "TSM"] <- "SPM"  
  
# add attributes to the data.frame  
  
## Station position  
stat <- list(station="NIOZjetty", longitude=4.789, latitude=53.002)  
  
## variable description  
desc <- data.frame(variable=c("Temperature", "Salinity", "Secchi", "Ntot", "Ptot",  
 "PO4", "NO3", "NO2", "NH4", "DON", "DOP", "Si", "DIC", "SPM", "Chl"),  
 description=c("Water temperature", "Salinity",   
 "Max depth at which secchi disk is visible",  
 "Total nitrogen concentration",   
 "Total phosphorus concentration", "Phosphate concentration",   
 "Nitrate concentration", "Nitrite concentration",   
 "Ammonium concentration", "Dissolved organic nitrogen",  
 "Dissolved organic phosphorus concentration",   
 "Dissolved silicate concentration",   
 "Dissolved inorganic carbon concentration",   
 "Total suspended particulate matter",   
 "Chlorophyll concentration"),   
   
 unit=c("dgC", "-", "m", rep("mmol/m3", times=10), "mg/m3", "mg/m3"))  
  
attr(Jetty, "station") <- stat  
attr(Jetty, "variables") <- desc  
attr(Jetty, "datasource") <- "nioz"  
attr(Jetty, "coordinates") <- c(lon=4.789, lat=53.002)  
attr(Jetty, "format") <- "wide"  
class(Jetty) <- c("dtLife", "data.frame")

## The processed data

The final data set contains 2003 samples, ranging from 1974-02-07 till 2021-11-30.

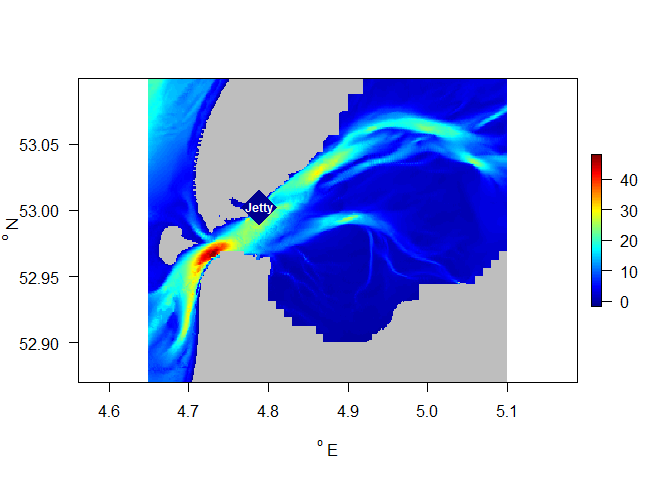
The measured variables in this data set are:

knitr::kable(desc, align='l')

| variable | description | unit |
| --- | --- | --- |
| Temperature | Water temperature | dgC |
| Salinity | Salinity | - |
| Secchi | Max depth at which secchi disk is visible | m |
| Ntot | Total nitrogen concentration | mmol/m3 |
| Ptot | Total phosphorus concentration | mmol/m3 |
| PO4 | Phosphate concentration | mmol/m3 |
| NO3 | Nitrate concentration | mmol/m3 |
| NO2 | Nitrite concentration | mmol/m3 |
| NH4 | Ammonium concentration | mmol/m3 |
| DON | Dissolved organic nitrogen | mmol/m3 |
| DOP | Dissolved organic phosphorus concentration | mmol/m3 |
| Si | Dissolved silicate concentration | mmol/m3 |
| DIC | Dissolved inorganic carbon concentration | mmol/m3 |
| SPM | Total suspended particulate matter | mg/m3 |
| Chl | Chlorophyll concentration | mg/m3 |

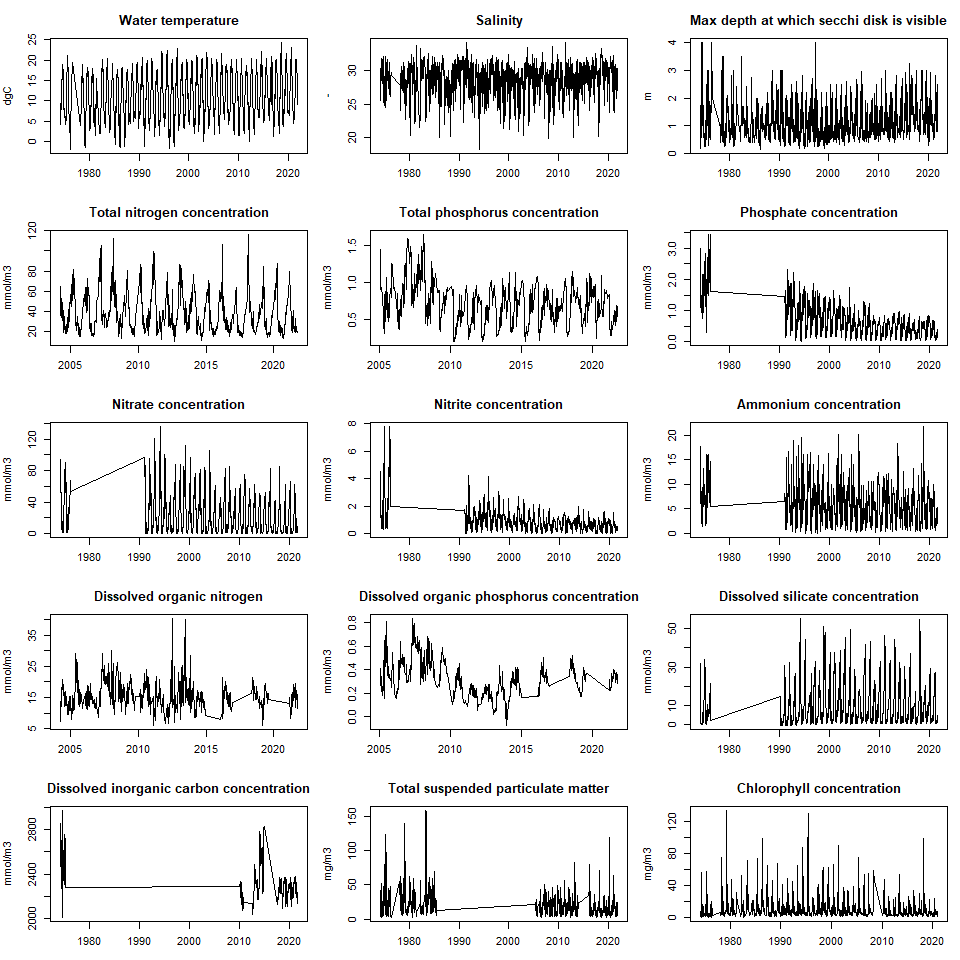
The position of the NIOZ jetty data:

stat <- attributes(Jetty)$station[c("longitude", "latitude")]  
plotBathymetry(Marsdiep,   
 pts=data.frame(longitude= stat$longitude, latitude=stat$latitude),   
 type = "image", NAcol="grey", pch=18, cex=5)  
text(x=stat$longitude, y=stat$latitude, labels="Jetty",   
 col="white", font=2, cex=0.75)



As a final check all variables are plotted versus time

par(mar=c(3,4,3,1))  
plot(Jetty, type="l", mfrow=c(5,3))



The data.frame is saved in binary R format (rda file)

save(file="../processed\_data/Jetty.rda", Jetty)

# References

The following R-sources were used for this work:

R-core:

* R Core Team (2022). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL <https://www.R-project.org/>.

R-package dtWad,

* Soetaert K (2024). dtWad: Waddensea Digital Twin: general utilities. R package version 0.0.1.