Population Trends Wadden Sea

Helmut Hillebrand

2023-12-21

## 1. Introduction

The Wadden Sea is changing. Direct local human impacts such as dredging, fisheries and tourism coincide with regional stressors such as climate change and eutrophication. To understand the response of biota to these changes many projects and monitoring programs have gathered time series of communities. These data are mainly used to assess community wide metrics of biodiversity or to analyse population trends for certain species of interest. However, biodiversity metrics offer only a limited view on changes, as some only capture net changes (i.e. more immigrations than extinctions reflected by species richness), whereas others only reflect responses of dominant species (Hillebrand et al. 2018). If, for example, all species would decline at the same rate without already going extinct, most biodiversity assessments would not pick that up. On the other hand, focusing on a subset of species may not capture the full range of population trends. There is a risk we are mainly focusing on extremes, i.e. single species that are rapidly declining or have become dominant or even invasive, but do not cover the more subtle changes in the majority of species.

In order to gain a more comprehensive view of the shifts in Wadden Sea biodiversity, we need a systematic and quantitative overview of individual species trends. These can be used to identify winners and loosers under current and past environmental change by comparing the trend per species across locations. By aggregating at higher taxonomic levels such as orders, classes or phyla, we can ask whether certain groups are especially endangered profiting, which also can enable addressing vulnerability and inform management. In the first approach, we will focus on simply analysing the trends per species, but in later iterations of the synthesis we can also compare this to potential drivers and co-dependencies between species (e.g. in a trophic context).

## 2. Workflow

Raw data are species-time-site data (Fig. 1, top right), i.e, observations of abundance or biomass per species per year at a repeatedly sampled site. For the purpose of this analysis, it is not useful to take seasonal changes into account, so the idea is to use annual mean values of abundance or biomass in those cases where multiple samples were done per year. There might be cases where the coverage of the year has changed, when e.g. initially only high biomass seasons were sampled, but later also winter samples were included. This would reduce the mean even when the season-specific samples were unchanged. Therefore, we will carry the information on how many samples were used per year with us in the forthcoming analyses. The minimum information just comprises a variable “StationID” that names the site of observation, “year” as a numeric format, “species” for the species name and “abu” for the abundance of the biomass. Please see under “3. Analyses” which steps can be taken to conform to this standard if variable names differ, the data is not on a yearly basis or the data are in wide format with species being in different columns.

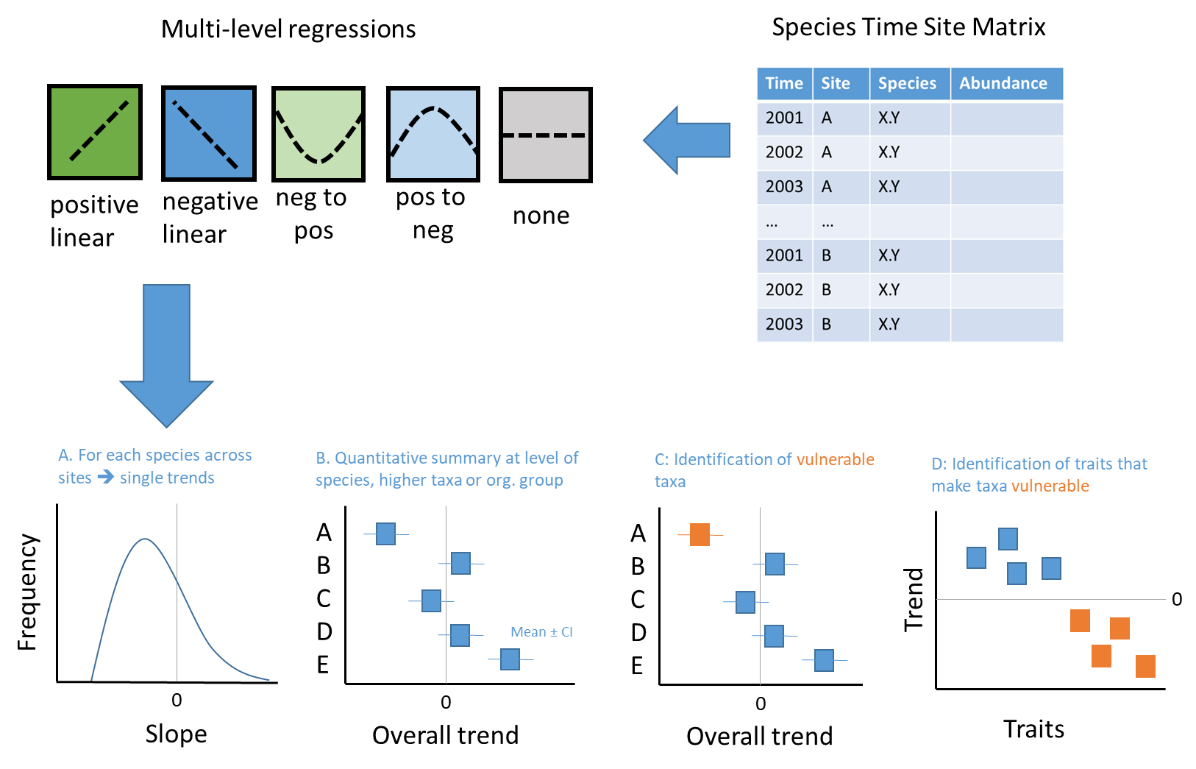


Fig. 1 Workflow of Poptend project

Population trends can be linear or non-linear of different polynomial orders (Fig. 1, top left). As we want to synthesize across species and phyla, I suggest to limit the polynomial to second order. This captures four general types of trends (next to the fifth case of “no trend”), a monotonic increase or decrease or a trend reversal from positive to negative or vice versa. To capture these trends, we will run linear and unimodal regressions on the abundance (abu, might also be biomass if this is the preferred response variable) of each species i at station j with year as the only predictor.

These will be compared using Akaike Information Criteria (AIC) and the MOS test (Mitchell-Olds & Shaw 1987). A unimodal trend is only a valid observation if three conditions apply:

1. the polynomial is the more parsimonious model, i.e. has a lower AIC than the linear model
2. the parameters b and c have different signs
3. the mode (time point of trend reversal) is in the time of observations, i.e. between the minimum and maximum year covered.

To be precise, if (i) is not true, a linear or no trend is observed. If (i) applies, but (ii) not, an accelerating trend is observed (positive when b and c are positive, negative when b and c are negative). In these cases, the slope changes with time, but the trend remains monotonic and shows no reversal of sign. If (i) and (ii) apply, but (iii) is not give, we observe decelerating trends, where depending on signs of b and c, we observe positive or negative trends that slow down. Figure 2 is an illustration of all possible cases in this decision tree

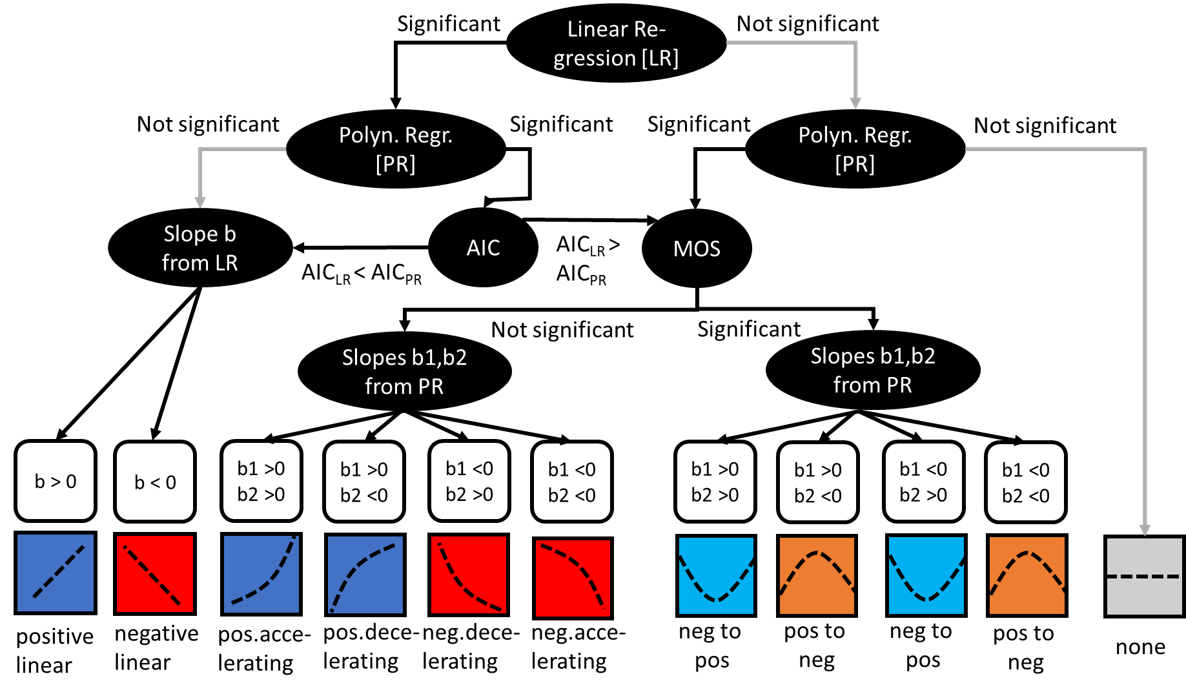


Fig. 2 Decision tree for potential outcomes of the two regressions and the corresponding trend information

For each unique combination of species and site, we will extract the information on the parameters a, b (and c) for linear and polynomical regressions together with their standard errors, plus the AIC, explained variance (R²) and significanc levels. Based on the decision tree in Fig. 2, we will use either the linear or nonlinear slopes for a weighted meta-analysis that as first step gives the average trend of species i across sites. We can analyse the overall distribution of these trends (Fig. 1, bottom panels), how much of the heterogeneity of the trends can be explained by functional groups, phylogeny or location.

## 3. Analyses

### Install libraries

library(tidyverse)  
library(vegan)  
library(reshape2)  
library(lubridate)  
library(broom)

### Explanatory case

As an explanatory case we are using the phytoplankton monitoring data from Lower Saxony and the Netherlands that were assembled during a recent EU Interreg Project. The data is ideal as it allows checking a number of potential caveats that might also occur in other data

data <- read\_delim("data/PPKTcount\_noHT\_28092022.csv", delim = ";", escape\_double = FALSE, trim\_ws = TRUE)  
names(data)

## [1] "Country" "StationID" "sampledate" "Phylum"   
## [5] "Class" "Order" "Family" "Genus"   
## [9] "Species" "abundance\_l" "cell\_biovolume" "cell\_carbon"   
## [13] "biovolume\_l" "carbon\_l"

Each of the following steps needs only to be done when required. They lead to the requested output for the actual analysis.

# STEP 1: If your variable names do not conform to our standard (StationID, year, species, abu), rename and delete empty rows   
data<-data %>% rename(species = Species, abu = biovolume\_l)   
# in our case StationID is okay, year needs to be calculated  
data<-data[!is.na(data$abu),]  
data<-data[!is.na(data$sampledate),]  
  
# STEP 2: If sample date but not year is given, extract the year. Library lubridate allows to first change the dates to numeric values and then extract the year  
data$date.numeric<-dmy(data$sampledate)   
# dmy is command if data are in format dd.mm.yyyy, ymd is for yyyy.mm.dd  
data$year <- year(data$date.numeric)   
# extracts the year  
  
# STEP 3: If your data can contain the same species more than once per sample (because different size classes or lifestages were distinguished) sum them  
data<-data %>%  
 group\_by(StationID, sampledate, year, species) %>%   
 summarise(abu= sum(abu, na.rm = T))  
# obviously sampledate and STEP 4 can be ignored if only annual data exist  
  
# STEP 4: Create annual means and the number of annual samples   
data<-data %>%  
 group\_by(StationID, year, species) %>%   
 summarise(abu= mean(abu, na.rm = T),  
 N.sample=length(abu))  
  
# STEP 5: Further reductions  
# if you need to exclude some StationID because they are not part of the WaddenSea or for other reasons  
data<-data %>%  
 filter(StationID %in% c("BOOMKDP", "DANTZGT", "DOOVBWT","GROOTGND","BOCHTVWTM", "HUIBGOT", "MARSDND", "ROTTMPT3", "TERSLG10", "Nney\_W\_2", "Bork\_W\_1", "WeMu\_W\_1", "JaBu\_W\_1"))  
  
# In other data sets it might be needed to exclude certain species or years.  
  
# Our example is already in long format, if your data is in wide format consider the "melt" command as in mentioned here, remove inital # to run  
#data <- melt(data,   
# id.vars=c("StationID", "year"), # all the variables to keep but not split apart  
# measure.vars=c(x:y), # number of the variables containing species, that shall be melted into one  
# variable.name="species", # name for new variable that contains the names of the melded variables  
# value.name="abu") # name of the new variables with the values

Not all species are suitable for this type of analysis. A recurrent problem are rare species with infrequent recordings. In years in which they have not been recorded, their abundance potentially was not 0. Relying on a few data points interspersed over a longer period increases the risk of finding significant but wrong trends. At the same time, replacing absences by zeros skews the population trends towards neutral. A generic solution can be to limit the analyses to species which have recordings for a threshold number of years (suggested here: >75%) and a minimum number of years (suggested here 5 years).

length.spec <-data %>%   
 group\_by(StationID, species) %>%   
 summarise(N.spec=length(unique(year)))  
  
length.stat <-data %>%   
 group\_by(StationID) %>%   
 summarise(N.stat=length(unique(year)))  
length<-merge(length.spec,length.stat,by="StationID")  
length$threshold<-length$N.spec/length$N.stat  
data<-merge(data,length,by=c("StationID","species"))  
  
data<-data[data$N.spec>4,]  
data<-data[data$threshold>.75,]

The following code now runs the regressions and extracts the relevant information from the regessions and merges them. I use log-transformed abundance here

#get slope b of LN(abu) ~ year from linear regression  
results.logabu = data %>%   
 group\_by(StationID, species) %>%   
 do(tidy(lm(log(abu)~year, data = .)))%>%   
 filter(term == "year")  
results.logabu<-results.logabu %>% rename(b.linear = estimate,  
 SE.b.linear = std.error,  
 pval.lin = p.value)  
  
results.logabu<-results.logabu[,c("StationID","species","b.linear","SE.b.linear", "pval.lin")]  
glance.logabu = data %>%   
 group\_by(StationID, species) %>%   
 do(glance(lm(log(abu)~year, data = .)))   
glance.logabu<-glance.logabu %>% rename(AIC.linear = AIC,  
 nobs.linear = nobs,  
 r2.lin = adj.r.squared)  
  
glance.logabu<-glance.logabu[,c("StationID","species","AIC.linear","nobs.linear","r2.lin")]  
results.logabu<-merge(results.logabu,glance.logabu, by=c("StationID","species"))  
  
results.year = data %>%   
 group\_by(StationID, species) %>%   
 summarize(min.year=min(year),  
 max.year=max(year))  
  
results.logabu<-merge(results.logabu,results.year, by=c("StationID","species"))  
  
#get slope b of LN(abu) ~ poly (year,2) from quadratic regression  
  
results.logabu2.lin = data %>%   
 group\_by(StationID,species) %>%   
 do(tidy(lm(log(abu)~poly(year,2), data = .)))%>%   
 filter(term == "poly(year, 2)1")   
results.logabu2.lin<-results.logabu2.lin %>% rename(b.poly = estimate,  
 SE.b.poly = std.error,  
 pval.b.poly = p.value)  
  
results.logabu2.lin<-results.logabu2.lin[,c("StationID","species","b.poly","SE.b.poly", "pval.b.poly")]  
  
#get slope c of LN(abu) ~ poly (year,2) from quadratic regression  
results.logabu2.quad = data %>%   
 group\_by(StationID,species) %>%   
 do(tidy(lm(log(abu)~poly(year,2), data = .)))%>%   
 filter(term == "poly(year, 2)2")   
  
results.logabu2.quad<-results.logabu2.quad %>% rename(c.poly = estimate,  
 SE.c.poly = std.error,  
 pval.c.poly = p.value)  
  
  
results.logabu2.quad<-results.logabu2.quad[,c("StationID","species","c.poly","SE.c.poly", "pval.c.poly")]  
  
glance.logabu2 = data %>%   
 group\_by(StationID,species) %>%   
 do(glance(lm(log(abu)~poly(year,2), data = .)))  
glance.logabu2<-glance.logabu2 %>% rename(AIC.poly = AIC,  
 nobs.poly = nobs,  
 r2.poly = adj.r.squared,  
 pval.poly = p.value)  
  
glance.logabu2<-glance.logabu2[,c("StationID","species","AIC.poly","nobs.poly","r2.poly","pval.poly")]  
  
results.logabu2<-merge(results.logabu2.lin,results.logabu2.quad, by=c("StationID","species"))  
results.logabu2<-merge(results.logabu2,glance.logabu2, by=c("StationID","species"))  
  
  
#merge linear and quadratic regression  
output<-merge(results.logabu,results.logabu2, by=c("StationID","species"))

In order to categorize the species trends according to Figure 2, we create a new variable “vote” and fill it according to criteria i and ii (iii follows in the next step). We need to discuss which significance level we want to use for benchmarking (suggested here p = 0.1).

#create the votes according to Figure 2  
  
output$vote<-NA  
# none of the regressions is significant  
output$vote[output$pval.lin>=0.1 & output$pval.poly>=0.1]<-"no"  
# only linear is significant  
output$vote[output$pval.lin<0.1 & output$pval.poly>0.1 &  
 output$b.linear<0]<-"negative linear"  
output$vote[output$pval.lin<0.1 & output$pval.poly>0.1 &  
 output$b.linear>0]<-"positive linear"  
  
# only quadratic is significant  
output$vote[output$pval.lin>0.1 & output$pval.poly<0.1 &  
 output$b.poly<0 & output$c.poly<0]<-"negative accelerating"  
output$vote[output$pval.lin>0.1 & output$pval.poly<0.1 &  
 output$b.poly>0 & output$c.poly>0]<-"positive accelerating"  
output$vote[output$pval.lin>0.1 & output$pval.poly<0.1 &  
 output$b.poly<0 & output$c.poly>0]<-"negative to positive"  
output$vote[output$pval.lin>0.1 & output$pval.poly<0.1 &  
 output$b.poly>0 & output$c.poly<0]<-"positive to negative"  
  
#both are significant, requiring involving AIC, first if AIC is lower for the polynomial  
output$vote[output$pval.lin<0.1 & output$pval.poly<0.1 &  
 output$AIC.linear>output$AIC.poly & output$b.poly<0 & output$c.poly<0]<-"negative accelerating"  
output$vote[output$pval.lin<0.1 & output$pval.poly<0.1 &  
 output$AIC.linear>output$AIC.poly & output$b.poly>0 & output$c.poly>0]<-"positive accelerating"  
output$vote[output$pval.lin<0.1 & output$pval.poly<0.1 &  
 output$AIC.linear>output$AIC.poly & output$b.poly<0 & output$c.poly>0]<-"negative to positive"  
output$vote[output$pval.lin<0.1 & output$pval.poly<0.1 &  
 output$AIC.linear>output$AIC.poly & output$b.poly>0 & output$c.poly<0]<-"positive to negative"  
# same but when AIC is lower for linear regression  
output$vote[output$pval.lin<0.1 & output$pval.poly<0.1 &  
 output$AIC.linear<output$AIC.poly & output$b.linear<0]<-"negative linear"  
output$vote[output$pval.lin<0.1 & output$pval.poly<0.1 &  
 output$AIC.linear<output$AIC.poly & output$b.linear>0]<-"positive linear"  
output$vote[output$pval.lin<0.1 & is.na(output$pval.poly) &  
 output$b.linear<0]<-"negative linear"  
output$vote[output$pval.lin<0.1 & is.na(output$pval.poly) &  
 output$b.linear>0]<-"positive linear"  
  
#sometimes, the polynomial or even the linear do not produce results, such that the p-value is NA, make these neutral trends as well  
output$vote[is.na(output$vote) & output$pval.lin>0.1 & is.na(output$pval.poly)]<-"no"  
output$vote[is.na(output$vote) & is.na(output$pval.lin)] <-"no"

The MOS test is part of the vegan package, it only needs to be done on the votes that passed criteria i and ii.

#create a unique case identifier  
output$UCI<-paste(output$StationID,output$species)   
data$UCI<-paste(data$StationID,data$species)   
UCI<-unique(output$UCI[output$vote=="negative to positive"|  
 output$vote=="positive to negative"])  
  
# as there might be cases when there is no such case, i do the loop as in if-else clause  
if(length(UCI)>0){  
 MOS.logabu<-data.frame()  
 for(i in 1:length(UCI)){  
 temp<-data[data$UCI==UCI[i], ]  
 if(dim(temp)[1]>2){  
 MOS<-MOStest(temp$year,log(temp$abu))  
 MOSout<-MOS$isBracketed  
 MOS.logabu<-rbind(MOS.logabu,data.frame(UCI=temp$UCI[1],MOSout))  
 rm(temp)  
 }  
 }  
} else{  
 MOS.logabu<-data.frame(matrix("", ncol = 1, nrow = dim(output)[1]))  
 MOS.logabu$UCI<-output$UCI  
 MOS.logabu$MOSout<-NA}  
  
  
output<-merge(output,MOS.logabu,by="UCI",all=T)  
  
# change votes if MOS Test was FALSE  
  
output$vote[output$vote=="positive to negative" &  
 output$MOSout=="FALSE"]<-"positive decelerating"  
output$vote[output$vote=="negative to positive" &  
 output$MOSout=="FALSE"]<-"negative decelerating"  
  
summary(as.factor(output$vote))

## negative accelerating negative decelerating negative linear   
## 10 1 39   
## negative to positive no positive accelerating   
## 15 358 29   
## positive linear positive to negative   
## 58 16

## 4. Meta-information to be used for synthesis

This analysis stream gives us 526 unique cases for this data set, i.e. species i at site j. The maximum of regressions for a single species is 13. Roughly 1/3 of the cases show some significant trend on their biovolume over time.

Based on analyses of additional Wadden Sea data we could expand this to more locations and other organism groups. This synthesized dataset would allow two approaches to a statistical analysis of the hypotheses.

We can do a formal weighted meta-analysis. Slopes are useful effect size metrics, their sampling variance is their squared standard error. We would need to decide whether we take the linear slope (b) or the slopes b and c from the quadratic regression. Additionally we can als use the correlation coefficient as a measure of the strength of the temporal trend, here sampling variances are based on the number of observations.

The first statistical test would simply address the hypotheses that there is a general temporal trend across all data sets.

Then, we would ask what can explain the (likely observed) heterogeneity in trends. Using a multi-level meta-analysis We can use phylogeny and organism group as fixed effects (explanatory variables), the identity of the data set and the StationID can be added as random effects.

In addition to the meta-analysis, we can ask whether there is a species, higher taxonomic group or oganism group that sticks out from the overall distribution of slopes, either positively (“Winners”) or negatively (“Loosers”).

Alternatively, we can also do weighted vote count, where we count the votes as delivered in output$vote but each case is weighted by the number of years with observations. Here we can also use a moving window approach that we tried in a different project to see whether the likelihood to find a positive or negative trend change with time.

Hoping that we all agree on the potential this might have, there are a few more steps to be done to be able to aggregate the data. If we do this in a standardized way, we can simply deposit the data set we worked on in an output-folder and read them one-by-one through “rbind”. For this, the order of variables needs to be exactly the same. If you follow the steps above, the “output” dataframe should look the same. If you follow the next steps in the same order for your data, our outputs will match and can be added.

# Rename the output (otherwise we overwrite each other)  
  
phyto1<-output  
  
# the data set needs a unique identifier (see above, will use this as random term)  
# suggest do use initials XXX and numbers that you can then allocate yourself  
phyto1$UDI<-"HLH001"  
  
# Information on the organism group  
phyto1$organism.group <-"phytoplankton"  
# entries can be: mammals, birds, macroinvertebrates, fish, zooplankton, phytoplankton, macrophytes, bacteria  
  
#Information whether species-biomass or -abundance were used  
phyto1$measure<-"biomass"  
  
  
# Information on the phylogeny  
# In the exemplary case this can be taken from the original data  
phylo <- read\_delim("data/PPKTcount\_noHT\_28092022.csv",   
 delim = ";", escape\_double = FALSE, trim\_ws = TRUE)  
phylo<-phylo %>% rename(species = Species)   
phylo<-phylo[,4:9]  
phylo<-phylo %>% distinct()  
phyto1<-merge(phyto1,phylo,by="species",all=F)  
  
# If the phylogeny is not available, you need to add all five variables via mydata$Phylum<-NA, same for Class, Order, Family, Genus. Instead of NA, you can if you have also add text, e.g. mydata$Phylum = "Mollusca"  
  
# Finally, a few variables for the data origin  
# where did the data come from  
phyto1$origin<-"Rijkswaterstaat NL"  
phyto1$origin[phyto1$StationID=="Nney\_W\_2"]<-"NLWKN"  
phyto1$origin[phyto1$StationID=="JaBu\_W\_1"]<-"NLWKN"  
phyto1$origin[phyto1$StationID=="Bork\_W\_1"]<-"NLWKN"  
phyto1$origin[phyto1$StationID=="WeMu\_W\_1"]<-"NLWKN"  
# who entered them into our synthesis  
phyto1$enter<-"Helmut Hillebrand"  
#is there a DOI where we can read more on this data  
phyto1$DOI<-"10.1007/s12526-023-01382-9"  
  
#We can discuss at a later point if we also want to map things, then we need to add lat and long for the StationID, currently I only add placeholders  
phyto1$lat<-NA  
phyto1$long<-NA  
  
# The final data set has the following names  
names(phyto1)

## [1] "species" "UCI" "StationID" "b.linear"   
## [5] "SE.b.linear" "pval.lin" "AIC.linear" "nobs.linear"   
## [9] "r2.lin" "min.year" "max.year" "b.poly"   
## [13] "SE.b.poly" "pval.b.poly" "c.poly" "SE.c.poly"   
## [17] "pval.c.poly" "AIC.poly" "nobs.poly" "r2.poly"   
## [21] "pval.poly" "vote" "MOSout" "UDI"   
## [25] "organism.group" "measure" "Phylum" "Class"   
## [29] "Order" "Family" "Genus" "origin"   
## [33] "enter" "DOI" "lat" "long"

# and needs to be saved like this  
write.csv(phyto1,file="output/phylo1.csv")

## 5. Literature

Hillebrand H., Blasius B., Borer E.T., Chase J.M., Downing J.A., Eriksson B.K., Filstrup C.T., Harpole W.S., Hodapp D., Larsen S., Lewandowska A.M., Seabloom E.W., Van de Waal D.B. & Ryabov A.B. (2018). Biodiversity change is uncoupled from species richness trends: consequences for conservation and monitoring. J. Appl. Ecol., 55, 169-184.

Mitchell-Olds T. & Shaw R.G. (1987). Regression Analysis of Natural Selection: Statistical Inference and Biological Interpretation. Evolution, 41, 1149-1161