Preparation of Villefranche EMODNET data

Peter M.J. Herman

29 mei 2018

Villefranche data

I selected all zooplankton data from the Villefranche permanent zooplankton station from EMODNET. There are two files, as there are additional measurements and facts.

The different data sets are mixed into one download file because I selected them all together. I will unmix them because they are methodologically incomparable.

The EMODNET data set contains (usually weekly) averages of abundances of different species. There is some hassle with starting and ending dates. Not all data have a correct starting and ending day, as some appear to either be monthly averages or some information is missing. I will check for these separately.

Reading in the basic data files (first the main file, then the measurements and facts)

```
bd<-read.csv("20180425_155218_15ae08812c7c53.csv",header = T,
stringsAsFactors = F)
mf<-read.csv("20180425_155240_15ae088289e138.csv",header = T,
stringsAsFactors = F)
```

Inspection of the data shows that there is a problem with dates, but also with abundance (fields are empty) for all data that have “1994Cnid” in measurementremarks. We temporarily drop this part of the files to avoid those problems. First we have to merge the two files.

```
bd<-merge(bd,mf,"occurrenceid")
bd<-bd[bd$measurementremarks!="1994Cnid",]
```

There are a few records that do not have startdaycollected, but do have enddaycollected, and a few where it is the other way round. We will give the available value to the other, nonavailable one.

```
bd[is.na(bd$startdaycollected),"startdaycollected"]<-
   bd[is.na(bd$startdaycollected),"enddaycollected"]
bd[is.na(bd$enddaycollected),"enddaycollected"]<-
   bd[is.na(bd$enddaycollected),"startdaycollected"]
```

We want a starting year for all observations. When it is unavailable, we take yearcollected as the starting year

```
bd$startyearcollected[is.na(bd$startyearcollected)]<-
   bd$yearcollected[is.na(bd$startyearcollected)]
```

```
bd$endyearcollected[is.na(bd$endyearcollected)]<-
   bd$yearcollected[is.na(bd$endyearcollected)]
```

After checking, we now have startyearcollected and endyearcollected for all records.

We do the same trick with months

```
bd$startmonthcollected[is.na(bd$startmonthcollected)]<-
   bd$monthcollected[is.na(bd$startmonthcollected)]
bd$endmonthcollected[is.na(bd$endmonthcollected)]<-
   bd$monthcollected[is.na(bd$endmonthcollected)]
```

Here the check shows we have no missing endmonthcollected, but some missing startmonthcollected. We assign them the value of endmonthcollected.

```
bd$startmonthcollected[is.na(bd$startmonthcollected)]<-
   bd$endmonthcollected[is.na(bd$startmonthcollected)]
```

We now calculate a startdate, an enddate, and a mean date for all observations

```
bd$startdate<-
as.Date(paste(bd$startdaycollected,bd$startmonthcollected,bd$startyearcollect
ed,sep="/"),format="%d/%m/%Y")
bd$enddate<-
as.Date(paste(bd$enddaycollected,bd$endmonthcollected,bd$endyearcollected,sep
="/"),format="%d/%m/%Y")
bd$middate<-
as.Date(apply(cbind(bd$startdate,bd$enddate),1,FUN=mean),origin="1970-01-01")
```

Further checks. We remove all old date columns, one of the identical columns occurrenceid and catalognumber and all meaningless columns (single value).

```
bd<-bd[,-c(1,(13:21))]
ul<-apply(bd,2,FUN=function(x) length(unique(x)))
ll<-which(ul==1)
bd<-bd[,-ll]
```

We have to tackle measurement units. There are three units used: ind/m3, ind/60m3 and ind/10m3. We want to express everything as ind/m3 for comparison and consistency.

```
mu<-unique(bd$measurementunit)
print(mu)

##  [1] "Individual/m^3.  Number of Fishery by week: 4"
##  [2] "Individual/m^3.  Number of Fishery by week: 5"
##  [3] "Individual/m^3.  Number of Fishery by week: 2"
##  [4] "Individual/m^3.  Number of Fishery by week: 3"
##  [5] "Individual/m^3.  Number of Fishery by week: 1"
##  [6] "Ind/60m3"
##  [7] "Abundance is a weekly average, expressed in individus/10m^3."
##  [8] "Abundance is a weekly average, expressed in individus/10m^3"
##  [9] "Individual/m^3.  Number of Fishery by week: 8"
## [10] "Individual/m^3.  Number of Fishery by week: 7"
## [11] "Individual/m^3.  Number of Fishery by week: 6"
## [12] "Individual/m^3.  Number of Fishery by week: 9"
## [13] "Individual/m^3.  Number of Fishery by week: 10"
## [14] "Individual/m^3.Number of Fishery by week: 6"
## [15] "Individual/m^3.Number of Fishery by week: 10"
## [16] "Individual/m^3.Number of Fishery by week: 8"
## [17] "Individual/m^3.Number of Fishery by week: 4"
## [18] "Individual/m^3.Number of Fishery by week: 9"
## [19] "Individual/m^3.Number of Fishery by week: 7"
## [20] "Individual/m^3.Number of Fishery by week: 5"
## [21] "Individual/m^3.Number of Fishery by week: 3"
## [22] "Individual/m^3.Number of Fishery by week: 1"
## [23] "Individual/m^3.Number of Fishery by week: 2"

bd$measurementvalue[bd$measurementunit==mu[6]]<-
  bd$measurementvalue[bd$measurementunit==mu[6]]/60
bd$measurementvalue[bd$measurementunit%in%mu[7:8]]<-
  bd$measurementvalue[bd$measurementunit%in%mu[7:8]]/10
```

The measurement unit contains two pieces of information: the unit itself, but also the number of samples that have been taken within the week for which the data stand. We do not want to loose the latter information, and store it in a field calle measurementfreq. For each of the unique units we store this frequency in freqs, then determine which unit was used, and give the appropriate element of freqs to the record. For all records, we now specify individuals/m3 as the unit, which is true after our transformations.

```
freqs<-c(4,5,2,3,1,NA,NA,NA,8,7,6,9,10,6,10,8,4,9,7,5,3,1,2)
wmu<-sapply(bd$measurementunit,FUN=function(x) which(mu==x),USE.NAMES=FALSE)
bd$measurementfreq<-freqs[wmu]
bd$measurementunit<-"Individual/m3"
```

This is a good time to save the mother data frame to a binary file.

Save the bd file

```
save(bd,file="bd.Rdata")
```

Subdatasets; Further analysis of bd1

The data set contains a field called datasetid, with three different values. Apparently I have downloaded all three datasets. We intend to analyse each of these three data sets separately. This analysis starts with the first one, which we call bd1. We collect the data of this data set in data frame bd1

```
udid<-unique(bd$datasetid)
bd1<-bd[bd$datasetid==udid[1],-c(15,16,20)]
```

After inspection of the data a number of questions can be posed. 1. Is the depth range always 0-75 m as it seems? In that case we do not further take it into account 2. sex/lifestage seems to be a further specification of species, but the impression is that within a species or group no further distinction in different lifestages is made. We check if any species occurs more than once in a single sample as a first test (if not, question closed)

```
print(unique(bd1$minimumdepth))

## [1] 0

print(unique(bd1$maximumdepth))

## [1] 75
```

Depth is OK, no further problems. We will now check if all species only occur once in a sample

```
require(reshape2)

## Loading required package: reshape2

bd1c<-dcast(bd1,middate~scientificname_accepted,value.var="measurementvalue",
            fun.aggregate = function(x) length(x[x>0]))
print(max(bd1c[,2:ncol(bd1c)]))

## [1] 2
```

Some species occur twice in a sample. We find out what varies (sex, lifestage)

```
print(unique(bd1$sex))

## [1] ""

print(unique(bd1$lifestage))

##  [1] ""                    "Nauplii larva"       "Euphausiacea larva"
##  [4] "Decapoda larva"      "Echinodermata larva" "Polychaeta larva"  
##  [7] "Veliger larva"       "Mollusca clutch"     "Fish larva"        
## [10] "Fish eggs"           "oozoids"             "blastozooids"
```

There is variation in lifestage, not in sex. What species are affected, and what dates?

```
tt<-apply(bd1c[,2:ncol(bd1c)],2,FUN=max)
ttt<-apply(bd1c[,2:ncol(bd1c)],1,FUN=max)
print(tt[tt>1])

##     Ihlea punctata           Mollusca             Pisces 
##                  2                  2                  2 
##   Salpa fusiformis Thalia democratica 
##                  2                  2

print(bd1c[ttt>1,1])

##   [1] "1992-01-15" "1992-01-22" "1992-02-05" "1992-02-12" "1992-02-19"
##   [6] "1992-03-18" "1992-03-25" "1992-04-08" "1992-04-15" "1992-04-22"
##  [11] "1992-04-29" "1992-05-06" "1992-06-03" "1992-06-10" "1992-07-15"
##  [16] "1992-08-26" "1992-10-21" "1993-03-17" "1993-04-28" "1993-05-05"
##  [21] "1993-07-28" "1993-08-04" "1993-08-18" "1993-08-25" "1993-09-01"
##  [26] "1993-09-08" "1993-09-15" "1993-09-22" "1993-09-29" "1993-10-06"
##  [31] "1993-10-13" "1993-10-20" "1993-10-27" "1993-11-03" "1994-02-02"
##  [36] "1994-03-23" "1994-03-30" "1994-04-06" "1994-04-13" "1994-04-20"
##  [41] "1994-04-27" "1994-05-04" "1994-05-11" "1994-05-18" "1994-05-25"
##  [46] "1994-06-01" "1994-06-08" "1994-06-15" "1994-06-22" "1994-06-29"
##  [51] "1994-07-06" "1994-09-07" "1994-10-12" "1994-10-19" "1994-10-26"
##  [56] "1994-11-02" "1994-11-23" "1994-11-30" "1994-12-07" "1994-12-14"
##  [61] "1994-12-21" "1995-01-18" "1995-02-01" "1995-02-22" "1995-03-08"
##  [66] "1995-03-15" "1995-03-22" "1995-03-29" "1995-04-05" "1995-04-12"
##  [71] "1995-04-19" "1995-04-26" "1995-05-10" "1995-06-14" "1995-08-23"
##  [76] "1995-08-30" "1996-02-07" "1996-04-03" "1996-04-17" "1996-04-24"
##  [81] "1996-05-01" "1996-05-08" "1996-05-15" "1996-05-22" "1996-05-29"
##  [86] "1996-06-05" "1996-06-12" "1996-06-19" "1996-06-26" "1996-07-03"
##  [91] "1996-07-10" "1996-07-17" "1996-07-24" "1996-07-31" "1996-08-07"
##  [96] "1996-08-28" "1996-09-04" "1996-09-11" "1996-09-18" "1996-09-25"
## [101] "1996-10-02" "1996-10-09" "1996-10-16" "1996-10-23" "1996-10-30"
## [106] "1996-11-06" "1996-11-13" "1996-11-20" "1996-12-04" "1996-12-11"
## [111] "1996-12-18" "1997-01-08" "1997-01-15" "1997-02-12" "1997-02-19"
## [116] "1997-04-16" "1997-04-23" "1997-04-30" "1997-05-14" "1997-05-21"
## [121] "1997-05-28" "1997-06-04" "1997-06-11" "1997-06-18" "1997-06-25"
## [126] "1997-07-02" "1997-07-09" "1997-07-16" "1997-07-23" "1997-07-30"
## [131] "1997-08-13" "1997-08-20" "1997-08-27" "1997-09-03" "1997-09-17"
## [136] "1997-09-24" "1997-10-01" "1997-10-08" "1997-10-15" "1997-10-22"
## [141] "1997-10-29" "1997-11-05" "1997-11-12" "1997-11-19" "1997-11-26"
## [146] "1998-01-07" "1998-02-04" "1998-02-11" "1998-03-04" "1998-03-18"
## [151] "1998-03-25" "1998-04-01" "1998-04-08" "1998-04-22" "1998-04-29"
## [156] "1998-05-20" "1998-05-27" "1998-06-24" "1998-07-01" "1998-07-08"
## [161] "1998-07-15" "1998-07-22" "1998-07-29" "1998-08-12" "1998-08-19"
## [166] "1998-08-26" "1998-09-02" "1998-09-09" "1998-09-16" "1998-09-23"
## [171] "1998-09-30" "1998-10-07" "1998-10-14" "1998-10-21" "1998-10-28"
## [176] "1998-11-04" "1998-11-11" "1998-11-18" "1998-11-25" "1998-12-02"
## [181] "1998-12-09" "1999-01-06" "1999-01-20" "1999-01-27" "1999-02-17"
## [186] "1999-03-10" "1999-03-17" "1999-03-24" "1999-03-31" "1999-04-07"
## [191] "1999-04-14" "1999-04-21" "1999-04-28" "1999-05-05" "1999-08-18"
## [196] "1999-08-25" "1999-09-08" "1999-09-15" "1999-09-22" "1999-09-29"
## [201] "1999-10-06" "1999-10-13" "1999-11-17" "1999-11-24" "1999-12-01"
## [206] "1999-12-08" "1999-12-15" "1999-12-22" "2005-01-26" "2005-03-09"
## [211] "2005-05-04" "2005-06-15" "2005-07-27" "2005-08-03" "2006-03-01"
## [216] "2006-10-18" "2006-11-15" "2006-11-29" "2007-02-07" "2007-02-28"
## [221] "2007-04-04" "2007-04-11" "2007-05-23" "2007-05-30" "2007-06-06"
## [226] "2007-06-13" "2007-06-27" "2007-08-01" "2007-08-22" "2007-09-05"
## [231] "2007-12-05" "2008-02-27" "2008-04-09" "2010-06-16"
```

There is distinction in lifestages for some samples, but not for all. We have to lump all lifestages together to make our basic crosstable

```r
bd1c<-dcast(bd1,middate~scientificname_accepted,value.var="measurementvalue",
            fun.aggregate = sum)
```

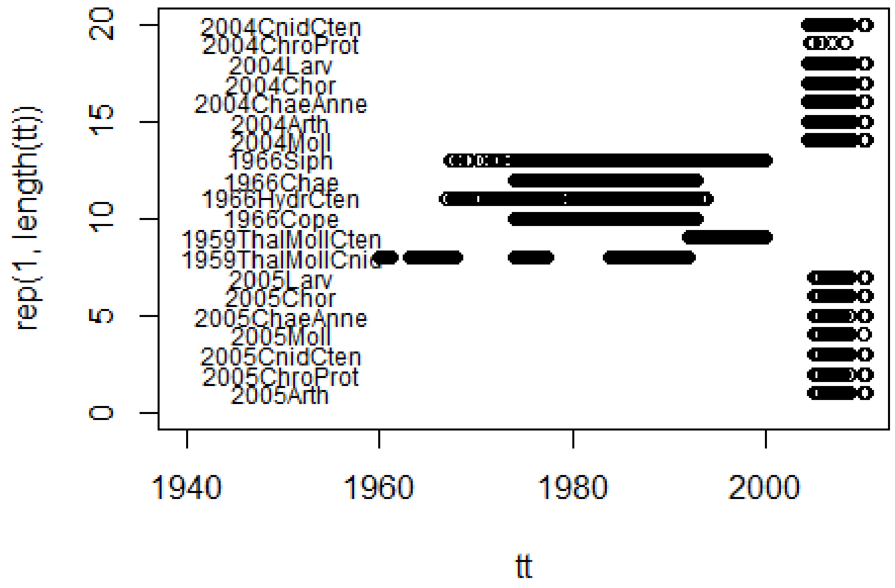
Basic inspection of the data. We make a plot of the time series for all species. It is stored in speciesplotsbd1.pdf

```
pdf("speciesplotsbd1.pdf")
for (i in 2:ncol(bd1c)){
   plot(bd1c[,1],bd1c[,i],main=names(bd1c)[i],xlab="Time",ylab="Density
(n/m3)")
}
dev.off()

## png
##   2
```

Disaster striking. This “dataset” appears to be split in subdatasets that do not overlap in time and have quite different species sets. Maybe the field measurementremarks can help us. It is available for all records and has a limited number of values. We plot the presence of samples in the different categories in time. We do it for the entire file bd, not just for bd1.

```
mrs<-unique(bd$measurementremarks)
for (i in 1:length(mrs)){
  tt<-unique(bd$middate[bd$measurementremarks==mrs[i]])
  if(i==1){plot(tt,rep(1,length(tt)),ylim=c(0,length(mrs)),
               xlim=c(as.Date("1940-01-01"),as.Date("2010-01-01")))
    }else{
           points(tt,rep(i,length(tt)))
    }
  text(as.Date("1950-01-01"),i,labels=mrs[i],cex=0.75)
}
```



This sheds new light on the subdivision of the dataset. Let's first look how species distribute over the different classes of measurementremarks, doing this for the entire dataset bd

```
tt<-
dcast(bd,measurementremarks~scientificname_accepted,value.var="measurementval
ue",
           fun.aggregate = function(x) length(x[x>0]))
```

I added taxonomic information to this table, and further ordered it to get a better insight. This was done offline, and has resulted in the spreadsheet "species-by-campaigns.xlsx". From this table the following actions were derived: 1. the xx and 2005xx datasets have the same set of species (which subtly differs from the species set in the other datasets), but the relative occurrence of different species in the two sets differs. It is better to keep them separate, but show them together for the different species 2. The two 1959xx datasets do not cover the same set of species. They differ in time coverage, without overlap. For some species, a complete series may be shown, but for other species that are exclusive for the later 1959ThalMollCten, it should be indicated that they have only been identified in the last period. This applies to all the Hydrozoa. The reverse (only in first set) is true for Doliolum nationalis. In addition lumping at higher level is needed for Pterotracheidae (family level) and Pyrosoma (genus level). 3. The 1966xxx set should be treated separately. As it largely overlaps in time with the 1959xx datasets, it may be displayed on the same page with the same x-axis. 4. The datasets 2004Larv and 2005Larv only contain information on larvae of larger taxonomic groups. They are comparable between them. We can better add "larvae" to the scientificname_accepted, so as to make clear this concerns larvae.

Treatment of the 2004xx and 2005xx datasets

First, we take the 2004xxx and 2005xxx datasets. As we want to show them together, we need the same set of species in both. We extract them together from the dataset, then do some manipulations that are common, and finally treat them separately.

```
bd200<-bd[substr(bd$measurementremarks,1,3)=="200",]
```

Check variability of lifestages. There is only variation on lifestage for samples of larvae. We use the lifestage as species indication here.

```
unique(bd200$lifestage)

##  [1] ""                    "Nauplii larva"       "Euphausiacea larva"
##  [4] "Decapoda larva"      "Echinodermata larva" "Polychaeta larva"
##  [7] "Veliger larva"       "Mollusca clutch"     "Fish larva"
## [10] "Fish eggs"           "Polychaeta  Larva"

unique(bd200[bd200$lifestage!="","measurementremarks"])

## [1] "2005Larv" "2004Larv"

#correct typing inconsistency in lifestage
bd200$lifestage[bd200$lifestage=="Polychaeta  Larva"]<-"Polychaeta larva"
ll<-which(bd200$lifestage!="")
bd200$scientificname_accepted[ll]<-bd200$lifestage[ll]
bd200$set<-substr(bd200$measurementremarks,1,4)
```

We make a cross-table of this dataset by species and then split it into the 2004 and 2005 datasets.

```
ct200<-dcast(bd200,middate+set~scientificname_accepted,
             value.var="measurementvalue",fun.aggregate = sum)
ct2004<-ct200[ct200$set=="2004",-2]
ct2005<-ct200[ct200$set=="2005",-2]
```

We check the time differences between the samples, and add dummy observations (NA) where there is a gap, so as to have a proper weekly series. In the 2004 dataset we correct the date of sample 78. We drop the last 11 (too sparse) observations. In the 2005 dataset we drop the last 9 observations.

```
# time differences
print(diff(as.numeric(ct2004$middate)))

##   [1]  14   7   7   7   7   7   7   7   7   7   7   7   7   7   7   7   7
##  [18]   7   7   7   7   7   7   7   7   7   7   7   7   7   7   7   7   7
##  [35]   7   7   7   7   7  21   7   7   7  42   7   7   7   7  21   7   7
##  [52]   7   7   7  14  14   7   7   7   7   7   7   7   7   7   7   7   7
##  [69]   7   7  14   7  14   7   7   7   9   5   7   7   7   7   7   7   7
##  [86]   7   7   7   7   7   7   7   7   7   7   7   7   7   7   7   7   7
## [103]   7   7   7   7   7   7   7   7   7   7   7   7   7   7   7   7   7
## [120]   7   7   7   7  21   7   7   7   7   7   7   7   7   7   7   7   7
## [137]   7   7   7   7   7   7   7   7   7   7   7   7   7   7   7   7   7
## [154]   7   7   7   7   7   7   7   7   7   7   7   7   7   7   7   7  14
## [171]   7   7  21   7   7   7   7   7   7   7   7   7   7   7   7   7   7
## [188]   7   7   7   7   7   7   7   7   7   7   7   7  14  35  28  28  28
## [205]  28 406  28  28  14  28  42
```

```
print(diff(as.numeric(ct2005$middate)))
```

```
##   [1]   7   7   7   7   7   7  21   7  49   7   7   7   7   7   7   7   7
##  [18]   7   7   7   7   7 154   7   7   7   7   7   7   7   7   7   7   7
##  [35]   7   7   7   7   7   7   7   7   7   7   7   7   7   7   7   7   7
##  [52]   7   7   7   7  14   7  14   7   7   7   7   7   7   7   7   7   7
##  [69]   7   7   7  21   7   7   7   7   7   7   7  14   7   7   7   7   7
##  [86]   7   7   7   7   7   7   7   7   7   7   7   7   7   7   7   7   7
## [103]   7   7   7   7  21  28   7   7   7   7   7   7   7  21   7   7   7
## [120]   7   7   7   7   7   7   7   7   7   7 140  28 476  14  28  28  14
## [137]  28  42
```

```
# correct date 78 in 2004
ct2004$middate[78]<-as.Date(as.numeric(ct2004$middate[78])-2,origin="1970-01-
01")
# drop last 11 observations in 2004
nr<-nrow(ct2004)
ll<-(nr-10):nr
ct2004<-ct2004[-ll,]
# drop last 9 observations in 2005
nr<-nrow(ct2005)
ll<-(nr-8):nr
ct2005<-ct2005[-ll,]
# insert 'empty' weeks in 2004
tt<-ct2004[1,]
tt[2:length(tt)]<-rep(NA,(ncol(ct2004)-1))
i<-2
while (i < nrow(ct2004)){
  if(as.numeric(ct2004$middate[i])-as.numeric(ct2004$middate[i-1])>7){
    tt[1]<-as.Date(as.numeric(ct2004$middate[i])-7,origin="1970-01-01")
    ct2004<-rbind(ct2004,tt)
    ct2004<-ct2004[order(ct2004$middate),]
  }else{i<-i+1}
}
# insert 'empty' weeks in 2005
tt<-ct2005[1,]
tt[2:length(tt)]<-rep(NA,(ncol(ct2005)-1))
i<-2
while (i < nrow(ct2005)){
  if(as.numeric(ct2005$middate[i])-as.numeric(ct2005$middate[i-1])>7){
    tt[1]<-as.Date(as.numeric(ct2005$middate[i])-7,origin="1970-01-01")
    ct2005<-rbind(ct2005,tt)
    ct2005<-ct2005[order(ct2005$middate),]
  }else{i<-i+1}
}
```

```
# determine species frequency
nc<-ncol(ct2004)
spfr2004<-c(220,apply(ct2004[,2:nc],2,FUN=function(x) length(x[x>0 &
!is.na(x)])))
nc<-ncol(ct2005)
spfr2005<-c(220,apply(ct2005[,2:nc],2,FUN=function(x) length(x[x>0 &
!is.na(x)])))
# drop species with a combined frequency in both datasets <25
spfr<-spfr2004+spfr2005
ll<-which(spfr<25)
ct2004<-ct2004[,-ll]
ct2005<-ct2005[,-ll]
# and save ct2004 and ct2005 for later use
save(ct2004,file="ct2004.Rdata")
save(ct2005,file="ct2005.Rdata")
```

This finishes for now the datasets. We prepare a pdf file with the plots per species, showing for each species both datasets.

```
pdf("speciesplots200xxx.pdf")
par(mfrow=c(2,1))
for (i in 2:ncol(ct2005)){
  par(mar=c(0,4,4,2),xaxt="n")
  plot(ct2004$middate,log(ct2004[,i]+1),type="b",main=names(ct2004[i]),
        ylab="log Abundance (n/m3)",xlab="Time",
        xlim=c(as.Date("2004-01-01"),as.Date("2009-01-01")))
  par(mar=c(5,4,0.5,2),xaxt="s")
  plot(ct2005$middate,log(ct2005[,i]+1),type="b",
          ylab="log Abundance (n/m3)",
          xlab="Time",xlim=c(as.Date("2004-01-01"),as.Date("2009-01-01")))
  par(mar=c(5,4,4,2)+0.1)
}
par(mfrow=c(1,1))
dev.off()

## png
##   2
```

Treatment of the 1959xx datasets

We extract both datasets from the overall dataset. Next we check for lifestages and sex. Sometimes a lifestage is indicated for Salpa, but only in the second of the two data sets. We check whether this results in two occurrences of the species at the same date. It does. We further check for which species no lifestages are indicated. For Pyrosoma, Pterotrachea and Lampetia lifestages are never indicated, for all other species they are always indicated.

```
bd1959<-bd[substr(bd$measurementremarks,1,4)=="1959",]
unique(bd1959$lifestage)

## [1] ""             "oozoids"      "blastozooids"
```

```r
unique(bd1959$sex)

## [1] ""

unique(bd1959$measurementremarks[bd1959$lifestage!=""])

## [1] "1959ThalMollCten"

sf<-dcast(bd1959,middate~scientificname_accepted,fun.aggregate = function(x)
length(x[x>0]),
          value.var = "measurementvalue")
print(max(sf[,2:ncol(sf)]))

## [1] 2

for (i in 2:ncol(sf)) print(paste(names(sf)[i],max(sf[,i])))

## [1] "Abylopsis tetragona 1"
## [1] "Chelophyes appendiculata 1"
## [1] "Doliolum nationalis 1"
## [1] "Hippopodius hippopus 1"
## [1] "Ihlea punctata 2"
## [1] "Lampetia 1"
## [1] "Lensia conoidea 1"
## [1] "Lensia subtilis 1"
## [1] "Muggiaea atlantica 1"
## [1] "Muggiaea kochii 1"
## [1] "Pterotrachea 1"
## [1] "Pterotracheidae 1"
## [1] "Pyrosoma 1"
## [1] "Pyrosoma atlanticum 1"
## [1] "Salpa fusiformis 2"
## [1] "Thalia democratica 2"

print(unique(bd1959$scientificname_accepted[bd1959$measurementremarks=="1959T
halMollCten" & bd1959$lifestage==""]))

## [1] "Pyrosoma"     "Pterotrachea" "Lampetia"

print(which(bd1959$scientificname_accepted=="Pyrosoma"&bd1959$lifestage!=""))

## integer(0)

print(which(bd1959$scientificname_accepted=="Pterotrachea"&bd1959$lifestage!=
""))

## integer(0)

print(which(bd1959$scientificname_accepted=="Lampetia"&bd1959$lifestage!=""))

## integer(0)
```

What to do with the lifestages? They are not indicated in the first subdataset, which is the big majority of samples. We decide for now to lump them, so as to make the two datasets comparable in this respect.

Next we have to lump species at the family and genus level to make the two sets comparable.This is needed for Pterotracheidae (family level) and Pyrosoma (genus level.

Then we make a cross table, making sure we sum values so as to lump lifestages. We transform Hydrozoa and Doliolum nationalis in the second dataset into NAs, as they were only recorded in the ThalMollCnid dataset. We also add interuptions in the dataset where there are gaps in the time series.

```
bd1959$scientificname_accepted[bd1959$genus=="Pyrosoma"]<- "Pyrosoma"
bd1959$scientificname_accepted[bd1959$family=="Pterotracheidae"]<-
"Pterotracheidae"
ct1959<-dcast(bd1959,middate+measurementremarks~scientificname_accepted,
              fun.aggregate = function(x) sum(x,na.rm=F),
              value.var = "measurementvalue")
unsampledCten<-c("Doliolum nationalis",
                "Abylopsis tetragona",
                "Chelophyes appendiculata",
                "Lensia conoidea",
                "Lensia subtilis",
                "Muggiaea atlantica",
                "Muggiaea kochii",
                "Hippopodius hippopus")
ll<-which(names(ct1959)%in%unsampledCten)
ct1959[ct1959$measurementremarks=="1959ThalMollCten",ll]<-NA

tt<-ct1959[1,]
tt[3:length(tt)]<-rep(NA,(ncol(ct2005)-2))
i<-3
while (i < nrow(ct1959)){
  if(as.numeric(ct1959$middate[i])-as.numeric(ct1959$middate[i-1])>7){
    tt[1]<-as.Date(as.numeric(ct1959$middate[i])-7,origin="1970-01-01")
    tt[2]<-ct1959[i,2]
    ct1959<-rbind(ct1959,tt)
    ct1959<-ct1959[order(ct1959$middate),]
  }else{i<-i+1}
}

save(ct1959,file="ct1959.Rdata")
```

We plot all species in a pdf file, named “speciesplots1959xxx.pdf”

```
pdf("speciesplots1959xxx.pdf")
for (i in 3:ncol(ct1959)){
   plot(ct1959$middate,log(ct1959[,i]+1),type="b",main=names(ct1959[i]),
        ylab="log Abundance (n/m3)",xlab="Time"
#        ,
#        xlim=c(as.Date("2004-01-01"),as.Date("2009-01-01"))
        )
}
```

It is clear from the plots that we still have a problem. There are many species that have not always been looked for. Many zeroes are actually NAs, but it is difficult to decide which precisely. This will require furter information from the data providers. For now it looks like the only species that have really been faithfully sampled are salps. For Hydrozoa only a few years appear consistent.

Treatment of the 1966xxx datasets

We first select the subset from the overall dataset. We then determine the limit dates for the different subsubsets, as these appear to differ. We will have to turn zeroes into NAs outside these limits. This dataset has sex and lifestages for at least some species. We will determine which species and find a plotting solution for this.

```
bd1966<-bd[substr(bd$measurementremarks,1,4)=="1966",]
# limits
sss<-unique(bd1966$measurementremarks)
print(sss)

## [1] "1966Cope"     "1966HydrCten" "1966Chae"     "1966Siph"

stmin<-stmax<-vector(length=length(sss))
for (i in 1:length(sss)){
  stmin[i]<-min(bd1966$middate[bd1966$measurementremarks==sss[i]])
  stmax[i]<-max(bd1966$middate[bd1966$measurementremarks==sss[i]])
}
stmin<-as.Date(stmin,origin="1970-01-01")
stmax<-as.Date(stmax,origin="1970-01-01")
# variation in sex and lifestages
print(unique(bd1966$lifestage))

## [1] ""        "Adult"   "Juvenil" "adult"   "juvenil" "Stade 1" "Stade 2"
## [8] "Stade 3" "Stade 4"

print(unique(bd1966$sex))

## [1] "M" ""  "F"

print(unique(bd1966$scientificname_accepted[bd1966$lifestage==""]))

##  [1] "Nannocalanus minor"        "Rhopalonema velatum"
##  [3] "Solmundella bitentaculata" "Calanus helgolandicus"
##  [5] "Mesosagitta minima"        "Abylopsis tetragona"
##  [7] "Chelophyes appendiculata"  "Hippopodius hippopus"
##  [9] "Lensia conoidea"           "Centropages typicus"
## [11] "Lensia subtilis"           "Muggiaea atlantica"
## [13] "Muggiaea kochii"           "Pleurobrachia rhodopis"
```

```
print(unique(bd1966$scientificname_accepted[bd1966$lifestage!=""]))

## [1] "Aglaura hemistoma"     "Flaccisagitta enflata" "Parasagitta setosa"
## [4] "Liriope tetraphylla"

print(unique(bd1966$scientificname_accepted[bd1966$sex==""]))

##  [1] "Rhopalonema velatum"       "Solmundella bitentaculata"
##  [3] "Aglaura hemistoma"         "Flaccisagitta enflata"
##  [5] "Parasagitta setosa"        "Mesosagitta minima"
##  [7] "Abylopsis tetragona"       "Chelophyes appendiculata"
##  [9] "Hippopodius hippopus"      "Lensia conoidea"
## [11] "Lensia subtilis"           "Muggiaea atlantica"
## [13] "Muggiaea kochii"           "Liriope tetraphylla"
## [15] "Pleurobrachia rhodopis"

print(unique(bd1966$scientificname_accepted[bd1966$sex!=""]))

## [1] "Nannocalanus minor"    "Calanus helgolandicus" "Centropages typicus"
```

Lifestage is always given for Aglaura hemistoma, Flaccisagitta enflate, Parasagitta setosa and Liriope tetraphylla, and never for the other species. Likewise, sex is always given for the copepods Nannocalanus minor, Calanus helgolandicus and Centropages typicus, and never for the other species. We can take this into account when plotting. In order to do so, we will need a 'sex' and 'lifestage' field in the cross table. For all species, we will need to list what columns concern the species, and make sure these columns are in logical order so that they can be displayed cumulatively.

```
ct1966<-dcast(bd1966,middate~scientificname_accepted+lifestage+sex,
              fun.aggregate = sum,value.var = "measurementvalue")
print(names(ct1966))

##  [1] "middate"                        "Abylopsis tetragona__"
##  [3] "Aglaura hemistoma_Adult_"       "Aglaura hemistoma_Juvenil_"
##  [5] "Calanus helgolandicus__F"       "Calanus helgolandicus__M"
##  [7] "Centropages typicus__F"         "Centropages typicus__M"
##  [9] "Chelophyes appendiculata__"     "Flaccisagitta enflata_adult_"
## [11] "Flaccisagitta enflata_juvenil_" "Hippopodius hippopus__"
## [13] "Lensia conoidea__"              "Lensia subtilis__"
## [15] "Liriope tetraphylla_Stade 1_"   "Liriope tetraphylla_Stade 2_"
## [17] "Liriope tetraphylla_Stade 3_"   "Liriope tetraphylla_Stade 4_"
## [19] "Mesosagitta minima__"           "Muggiaea atlantica__"
## [21] "Muggiaea kochii__"              "Nannocalanus minor__F"
## [23] "Nannocalanus minor__M"          "Parasagitta setosa_adult_"
## [25] "Parasagitta setosa_juvenil_"    "Pleurobrachia rhodopis__"
## [27] "Rhopalonema velatum__"          "Solmundella bitentaculata__"

spcol<-data.frame(species=c("Abylopsis tetragona",
                            "Aglaura hemistoma",
                            "Calanus helgolandicus",
                            "Centropages typicus",
                            "Chelophyes appendiculata",
                            "Flaccisagitta enflata",
                            "Hippopodius hippopus",
                            "Lensia conoidea",
                            "Lensia subtilis",
                            "Liriope tetraphylla",
                            "Mesosagitta minima",
                            "Muggiaea atlantica",
                            "Muggiaea kochii",
                            "Nannocalanus minor",
                            "Parasagitta setosa",
                            "Pleurobrachia rhodopis",
                            "Rhopalonema velatum",
                            "Solmundella bitentaculata"),
            colstart=c(2,3,5,7,9,10,12,13,14,15,
                             19,20,21,22,24,26,27,28),
            colend=c(2,4,6,8,9,11,12,13,14,18,19,
                           20,21,23,25,26,27,28),
            nstages=c(1,2,2,2,1,2,1,1,1,4,1,1,1,
                            2,2,1,1,1),
            stage1=c(NA,"Adult","Female","Female",NA,"Adult",NA,NA,NA,
                           "Stage1",NA,NA,NA,"Female","Adult",NA,NA,NA),
            stage2=c(NA,"Juvenile","Male","Male",NA,"Juvenile",NA,NA,NA,
                           "Stage2",NA,NA,NA,"Male","Juvenile",NA,NA,NA),
            stage3=c(NA,NA,NA,NA,NA,NA,NA,NA,NA,
                           "Stage3",NA,NA,NA,NA,NA,NA,NA,NA),
            stage4=c(NA,NA,NA,NA,NA,NA,NA,NA,NA,
                           "Stage4",NA,NA,NA,NA,NA,NA,NA,NA)  )
for (i in 1:nrow(spcol)){
  sset<-unique(bd1966$measurementremarks[bd1966$scientificname_accepted==
                                           spcol$species[i]])
  spcol$stmin[i]<-stmin[which(sss==sset)]
  spcol$stmax[i]<-stmax[which(sss==sset)]
}
```

We have to take care of too long empty patches in the time series, substituting holes with NAs.

```
tt<-ct1966[1,]
tt[2:length(tt)]<-rep(NA,(ncol(ct1966)-1))
i<-2
while (i < nrow(ct1966)){
  if(as.numeric(ct1966$middate[i])-as.numeric(ct1966$middate[i-1])>7){
    tt[1]<-as.Date(as.numeric(ct1966$middate[i])-7,origin="1970-01-01")
    ct1966<-rbind(ct1966,tt)
    ct1966<-ct1966[order(ct1966$middate),]
  }else{i<-i+1}
}
```

We proceed by making basic plots of all species, subdivided into groups when required. We also blank all observations outside of the appropriate observations windows.

```
for(i in 1:nrow(spcol)){
  for (j in spcol$colstart[i]:spcol$colend[i]){
    ct1966[ct1966$middate<spcol$stmin[i] | ct1966$middate>spcol$stmax[i],j]<-
NA
  }
}

pdf("speciesplots1966.pdf")
cols<-c("black","red","darkgreen","darkblue")
for(i in 1:nrow(spcol)){
  plot(ct1966$middate,log(ct1966[,spcol$colstart[i]]+1),type="l",
       main=spcol$species[i],xlab="Time",ylab="log Abundance /m3",
       xlim=c(as.Date("1966-01-01"),as.Date("2000-01-01")),col=cols[1])
  j<-spcol$colstart[i]+1
  curser<-ct1966[,spcol$colstart[i]]+1
  curcol<-2
  while(j<=spcol$colend[i]){
    lines(ct1966$middate,log(ct1966[,j]+curser),type="l",col=curcol+1)
    curser<-curser+ct1966[,j]
    curcol<-curcol+1
    j<-j+1
  }
}
dev.off()
```

```
## png
##   2
```

End of analysis