NEUS: Northeast US survey data processing summary

fishglob, Aurore A. Maureaud, Juliano Palacios Abrantes, Zoë Kitchel, Dan Forrest, & Michelle Stuart

December, 2022

Contents

General info . . . 1
Data cleaning in R . . . 1
1. Overview of the survey data table . . . 14
2. Summary of sampling intensity . . . 15
3. Summary of sampling variables from the survey . . . 16
4. Summary of biological variables . . . 17
5. Extreme values . . . 18
6. Summary of variables against swept area . . . 19
7. Abundance or Weight trends of the six most abundant species . . . 20
8. Distribution mapping . . . 21
9. Taxonomic flagging . . . 22
10. Spatio-temporal standardization: NEUS-Fall . . . 23
    a. Standardization method 1 . . . 23
    b. Standardization method 2 . . . 26
    c. Standardization summary . . . 26
11. Spatio-temporal standardization: NEUS-Spring . . . 27
    a. Standardization method 1 . . . 27
    b. Standardization method 2 . . . 30
    c. Standardization summary . . . 30

General info

This document presents the cleaning code and summary of the Northeast US bottom trawl survey provided by Sean Lucey sean.lucey@noaa.gov. It contains data from 1963 and up to 2020.

Data cleaning in R

```
##############################################################################
#### R code to clean trawl survey Northeast US (Fall and Spring)
#### Public data Ocean Adapt
#### Contacts: Sean Lucey   sean.lucey@noaa.gov Fisheries Biologist,
#### Northeast Fisheries Science Center, NOAA
#### Coding: Michelle Stuart, Dan Forrest, Zoë Kitchel November 2021
##############################################################################

#NB: Note that there was a gear and vessel swap in 2008-2009 (Albatross to Bigelow)
#this code uses conversions from NEFSC to correct data post 2009 to pre 2009
#abundance and biomass
#Sampling was based on a stratified random design using area and depth zones.
#Standard tows from 1963-2008 were 30 minutes in duration. Initially, the towing
```

```
#speed was set to approximately 3.5 knots, but in 1996, it was discovered that the
#speedlog was not working and to go
#3.5 knots by the speedlog required a speed of 3.8 knots on the Doppler. Therefore,
#speed was then changed to 3.8 knots
#for the rest of the time series. Tow direction was towards the next station unless
#wind and sea state dictated a different
#course. For tows in depths >= 183 m, a more specific depth is randomly chosen among
#four depth intervals and then
#trawling was done along that depth contour. In 2009, the tows were reduced to 20
#minutes in duration and speed was
#reduced to 3.0 knots. The direction of the tows changed to follow the depth contour.

#NB: haul_dur is raw, does not account for conversions in abundance and biomass,
#and therefore should not be used
#as the denominator for CPUE calculations without careful consideration, therefore,
#wgt_cpue, wgt_h, num_cpue, and num_h are not calculated


#Helpful documents discussing gear and  vessel transition for Northeast US
#https://www.nafo.int/Portals/0/PDFs/sc/2014/scr14-024.pdf
#http://ices.dk/sites/pub/CM%20Doccuments/CM-2007/Q/Q2007.pdf
#https://repository.library.noaa.gov/view/noaa/3726

#-----------------------------------------------------------------------------------#
#### LOAD LIBRARIES AND FUNCTIONS ####
#-----------------------------------------------------------------------------------#

library(rfishbase) #needs R 4.0 or more recent
library(tidyverse)
library(lubridate)
library(googledrive)
library(taxize) # for getting correct species names
library(magrittr) # for names wrangling
library(data.table)



source("functions/clean_taxa.R")
source("functions/write_clean_data.R")

#Data for the NEUS can be best accessed using the Pinsky Lab Ocean Adapt
#Public Git Hub Repository.

#-----------------------------------------------------------------------------------#
#### PULL IN AND EDIT RAW DATA FILES ####
#-----------------------------------------------------------------------------------#

#load conversion factors to bridge Albatross (vessel before 2008)
#data with Bigelow Data (vessel after)
NEFSC_conv <- read_csv(
"https://github.com/pinskylab/OceanAdapt/raw/master/data_raw/NEFSC_conversion_factors.csv",
  col_types = "_ddddddd")
NEFSC_conv <- data.table::as.data.table(NEFSC_conv)
```

```r
#Bigelow >2008 Vessel Conversion
#Use Bigelow conversions for Pisces as well (PC)
#Tables 56-58 from Miller et al. 2010 Biomass estimators
big_fall <- data.table::data.table(svspp =
                                     c('012', '022', '024', '027', '028',
                                      '031', '033', '034', '073', '076',
                                      '106', '107', '109', '121', '135',
                                      '136', '141', '143', '145', '149',
                                      '155', '164', '171', '181', '193',
                                      '197', '502', '512', '015', '023', '026',
                                      '032', '072', '074', '077', '078',
                                      '102', '103', '104', '105', '108',
                                      '131', '163', '301', '313', '401',
                                      '503'),
                                   season = c(rep('fall', 47)),
                         rhoW = c(
                               1.082, 3.661, 6.189, 4.45, 3.626, 1.403, 1.1, 2.12,
                               1.58, 2.088, 2.086, 3.257, 12.199, 0.868, 0.665, 1.125,
                               2.827, 1.347, 1.994, 1.535, 1.191, 1.354, 3.259, 0.22,
                               3.912, 8.062, 1.409, 2.075, 1.21,
                               2.174, 8.814, 1.95, 4.349, 1.489, 3, 2.405, 1.692,
                               2.141, 2.151, 2.402, 1.901, 1.808, 2.771, 1.375, 2.479,
                               3.151, 1.186))

big_spring <- data.table::data.table(svspp = c(
                                    '012', '022', '024', '027', '028',
                                    '031', '033', '034', '073', '076',
                                    '106', '107', '109', '121', '135',
                                    '136', '141', '143', '145', '149',
                                    '155', '164', '171', '181', '193',
                                    '197', '502', '512', '015', '023',
                                    '026', '032', '072', '074', '077',
                                    '078', '102', '103', '104', '105',
                                    '108', '131', '163', '301', '313',
                                    '401', '503'),
                                     season = c(rep('spring', 47)),
                                     rhoW = c(
                             1.082, 3.661, 6.189, 4.45, 3.626, 1.403, 1.1, 2.12,
                            1.58, 2.088, 2.086, 3.257, 12.199, 0.868, 0.665, 1.125,
                            2.827, 1.347, 1.994, 1.535, 1.191, 1.354, 3.259, 0.22,
                            3.912, 8.062, 1.409, 2.075, 1.166, 3.718, 2.786, 5.394,
                            4.591, 0.878, 3.712, 3.483, 2.092, 3.066, 3.05, 2.244,
                            3.069, 2.356, 2.986, 1.272, 3.864, 1.85, 2.861))


#read strata file
neus_strata <- read.csv(
  "https://github.com/pinskylab/OceanAdapt/raw/master/data_raw/neus_strata.csv")

neus_strata <- neus_strata %>%
    select(stratum, stratum_area) %>%
    mutate(stratum = as.double(stratum)) %>%
    distinct()
```

```r
#read and clean spp file
neus_spp_raw <- read_lines(
  "https://github.com/pinskylab/OceanAdapt/raw/master/data_raw/neus_spp.csv")
neus_spp_raw <- str_replace_all(neus_spp_raw,
                                'SQUID, CUTTLEFISH, AND OCTOPOD UNCL',
                                'Squid/Cuttlefish/Octopod (unclear)')
neus_spp_raw <- str_replace_all(neus_spp_raw,
                                'SEA STAR, BRITTLE STAR, AND BASKETSTAR UNCL',
                                'Sea Star/Brittle Star/Basket Star (unclear)')
neus_spp_raw <- str_replace_all(neus_spp_raw,
                                'MOON SNAIL, SHARK EYE, AND BABY-EAR UNCL',
                                'Moon Snail/shark eye/baby-ear (unclear)')
neus_spp_clean <- str_replace_all(neus_spp_raw, 'SHRIMP \\(PINK,BROWN,WHITE\\)',
                                  'Shrimp \\(pink/brown/white\\)')
write_lines(neus_spp_clean,"neus_spp_clean.txt")
neus_spp <- read_csv("neus_spp_clean.txt", col_types = cols(.default = col_character()))
rm(neus_spp_clean, neus_spp_raw)
file.remove("neus_spp_clean.txt")


#NEUS Fall
neus_catch_raw <- read_lines(
  "https://github.com/pinskylab/OceanAdapt/raw/master/data_raw/neus_fall_svcat.csv")
# remove comma
neus_catch_raw <- str_replace_all(
  neus_catch_raw, 'SQUID, CUTTLEFISH, AND OCTOPOD UNCL',
                                  'Squid/Cuttlefish/Octopod (unclear)')
neus_catch_raw <- str_replace_all(
  neus_catch_raw, 'SEA STAR, BRITTLE STAR, AND BASKETSTAR UNCL',
                                  'Sea Star/Brittle Star/Basket Star (unclear)')
neus_catch_raw <- str_replace_all(
  neus_catch_raw, 'MOON SNAIL, SHARK EYE, AND BABY-EAR UNCL',
                                  'Moon Snail/shark eye/baby-ear (unclear)')
neus_catch_raw <- str_replace_all(
  neus_catch_raw, 'MOON SNAIL, SHARK EYE, AND BABY-EAR UNCL',
                                  'Moon Snail/shark eye/baby-ear (unclear)')
neus_catch_clean <- str_replace_all(
  neus_catch_raw, 'SHRIMP \\(PINK,BROWN,WHITE\\)',
                                    'Shrimp \\(pink/brown/white\\)')
write_lines(neus_catch_clean, file = "neus_catch_clean.txt")
neus_fall_catch <- read_csv("neus_catch_clean.txt",
                            col_types = cols(.default = col_character()))
file.remove("neus_catch_clean.txt")

neus_fall_haul <- read_csv(
  "https://github.com/pinskylab/OceanAdapt/raw/master/data_raw/neus_fall_svsta.csv",
  col_types = cols(.default = col_character()))

#-------------------------------------------------------------------------------------#
#### REFORMAT AND MERGE DATA FILES ####
#-------------------------------------------------------------------------------------#

neus_fall <- left_join(neus_fall_catch, neus_fall_haul,
```

```
                        by = c("ID", "STATION","CRUISE6","STRATUM","TOW"))
neus_fall <- left_join(neus_fall, neus_spp, by = "SVSPP")

neus_fall <- neus_fall %>%
  rename(year = EST_YEAR,
         month = EST_MONTH,
         day = EST_DAY,
         latitude = DECDEG_BEGLAT,
         longitude = DECDEG_BEGLON,
         depth = AVGDEPTH,
         stratum = STRATUM,
         haul_id = ID,
         verbatim_name = SCINAME,
         #Expanded biomass of a species caught at a given station.
         wgt = EXPCATCHWT,
         #Expanded number of individuals of a species caught at a given station.
         num = EXPCATCHNUM,
         station = STATION,
         sst = SURFTEMP,
         sbt = BOTTEMP,
         gear = SVGEAR)

neus_fall <- neus_fall %>%
  mutate(stratum = as.double(stratum),
         latitude = as.double(latitude),
         longitude = as.double(longitude),
         depth = as.double(depth),
         wgt = as.double(wgt),
         num = as.double(num),
         year = as.double(year),
         haul_dur = as.numeric(TOWDUR)/60, #convert minutes to hours
         quarter = case_when(month %in% c(1,2,3) ~ 1,
                             month %in% c(4,5,6) ~ 2,
                             month %in% c(7,8,9) ~ 3,
                             month %in% c(10,11,12) ~ 4),
         season = "Fall",
         SVSPP = as.double(SVSPP)
 )

#apply fall conversion factors
setDT(neus_fall)

dcf.spp <- NEFSC_conv[DCF_WT > 0, SVSPP]

#test for changes due to conversion with "before" and "after"
#before <- neus_fall[year < 1985 & SVSPP %in% dcf.spp,
#.(mean_wtcpue=mean(wtcpue)), by=SVSPP][order(SVSPP)]

for(i in 1:length(dcf.spp)){
  neus_fall[year < 1985 & SVSPP == dcf.spp[i], wgt  := wgt  * NEFSC_conv[
    SVSPP == dcf.spp[i], DCF_WT]]
}
```

```
#after <- neus_fall[year < 1985 & SVSPP %in% dcf.spp,
#.(mean_wtcpue=mean(wtcpue)), by=SVSPP][order(SVSPP)]

#before <- neus_fall[SVVESSEL == 'DE' & SVSPP %in% vcf.spp,
#.(mean_wtcpue=mean(wtcpue)), by=SVSPP][order(SVSPP)]

vcf.spp <- NEFSC_conv[VCF_WT > 0, SVSPP]
for(i in 1:length(dcf.spp)){
  neus_fall[SVVESSEL == 'DE' & SVSPP == vcf.spp[i], wgt  := wgt * NEFSC_conv[
    SVSPP == vcf.spp[i], VCF_WT]]
}

#after<- neus_fall[SVVESSEL == 'DE' & SVSPP %in% vcf.spp,
#.(mean_wtcpue=mean(wtcpue)), by=SVSPP][order(SVSPP)]

spp_fall <- big_fall[season == 'fall', svspp]

#before <- neus_fall[SVVESSEL %in% c('HB', 'PC') & SVSPP %in% spp_fall,
#.(mean_wtcpue=mean(wtcpue)), by=SVSPP][order(SVSPP)]

for(i in 1:length(big_fall$svspp)){
  neus_fall[
    SVVESSEL %in% c('HB', 'PC') & SVSPP == spp_fall[i], wgt  := wgt  / big_fall[i, rhoW]]
}

#after <- neus_fall[SVVESSEL %in% c('HB', 'PC')  & SVSPP %in% spp_fall,
#.(mean_wtcpue=mean(wtcpue)), by=SVSPP][order(SVSPP)]

neus_fall <- as.data.frame(neus_fall)

# sum different sexes of same spp together
neus_fall <- neus_fall %>%
  group_by(year, latitude, longitude, depth, haul_id, CRUISE6, station,
           stratum, verbatim_name, gear, haul_dur) %>%
  mutate(wgt = sum(wgt)) %>%
  ungroup()


#join with strata
neus_fall <- left_join(neus_fall, neus_strata, by = "stratum")
neus_fall <- filter(neus_fall, !is.na(stratum_area))
neus_fall <- neus_fall %>%
  rename(stratumarea = stratum_area) %>%
  #convert square nautical miles to square kilometers
  mutate(stratumarea = as.double(stratumarea)* 3.429904)
neus_fall$survey <- "NEUS"

neus_fall<- neus_fall %>%
  mutate(
    source = "NOAA",
    timestamp = mdy("03/01/2021"),
    country = "United States",
    continent = "n_america",
```

```
    sub_area = NA,
    stat_rec = NA,
    area_swept = 0.0384, #Average tow area in km^2 for albatross
    #num_h could be calculated by dividing num by haul duration,
    #but use caution because of 2008-2009 conversion
    num_h = NA,
    #num_cpue could be calculated by dividing num by area swept,
    #but use caution because of 2008-2009 conversion
    num_cpue = NA,
    #wgt_h could be calculated by dividing wgt by haul duration,
    #but use caution because of 2008-2009 conversion
    wgt_h = NA,
    #wgt_cpue could be calculated by dividing wgt by area swept,
    #but use caution because of 2008-2009 conversion
    wgt_cpue = NA
     ) %>%
  select(survey, haul_id, source, timestamp, country, sub_area, continent, stat_rec, station,
         stratum, year, month, day, quarter, season, latitude, longitude,
         haul_dur, area_swept, gear, depth, sbt, sst,
         num, num_h, num_cpue, wgt, wgt_h, wgt_cpue, verbatim_name)

rm(neus_catch_clean, neus_catch_raw, neus_fall_catch, neus_fall_haul)

##############################################################################
#NEUS Spring


#-----------------------------------------------------------------------------------------#
#### PULL IN AND EDIT RAW DATA FILES ####
#-----------------------------------------------------------------------------------------#

neus_catch_raw <- read_lines(
  "https://github.com/pinskylab/OceanAdapt/raw/master/data_raw/neus_spring_svcat.csv")
# remove comma
neus_catch_raw <- str_replace_all(
  neus_catch_raw, 'SQUID, CUTTLEFISH, AND OCTOPOD UNCL',
                                  'Squid/Cuttlefish/Octopod (unclear)')
neus_catch_raw <- str_replace_all(
  neus_catch_raw, 'SEA STAR, BRITTLE STAR, AND BASKETSTAR UNCL',
                                  'Sea Star/Brittle Star/Basket Star (unclear)')
neus_catch_raw <- str_replace_all(
  neus_catch_raw, 'MOON SNAIL, SHARK EYE, AND BABY-EAR UNCL',
                                  'Moon Snail/shark eye/baby-ear (unclear)')
neus_catch_raw <- str_replace_all(
  neus_catch_raw, 'MOON SNAIL, SHARK EYE, AND BABY-EAR UNCL',
                                  'Moon Snail/shark eye/baby-ear (unclear)')
neus_catch_clean <- str_replace_all(neus_catch_raw, 'SHRIMP \\(PINK,BROWN,WHITE\\)',
                                    'Shrimp \\(pink/brown/white\\)')
write_lines(neus_catch_clean, file = "neus_catch_clean.txt")
neus_spring_catch <- read_csv("neus_catch_clean.txt",
                              col_types = cols(.default = col_character()))
file.remove("neus_catch_clean.txt")
rm(neus_catch_clean, neus_catch_raw)
```

```r
neus_spring_haul <- read_csv(
  "https://github.com/pinskylab/OceanAdapt/raw/master/data_raw/neus_spring_svsta.csv",
  col_types = cols(.default = col_character()))
neus_spring <- left_join(neus_spring_catch, neus_spring_haul,
                         by = c("ID", "STATION", "CRUISE6","STRATUM","TOW"))
neus_spring <- left_join(neus_spring, neus_spp, by = "SVSPP")


rm(neus_spring_catch, neus_spring_haul)
neus_spring <- neus_spring %>%
  rename(year = EST_YEAR,
         month = EST_MONTH,
         day = EST_DAY,
         latitude = DECDEG_BEGLAT,
         longitude = DECDEG_BEGLON,
         depth = AVGDEPTH,
         stratum = STRATUM,
         haul_id = ID,
         verbatim_name = SCINAME,
         #Expanded biomass of a species caught at a given station.
         wgt = EXPCATCHWT,
         #Expanded number of individuals of a species caught at a given station.
         num = EXPCATCHNUM,
         station = STATION,
         sst = SURFTEMP,
         sbt = BOTTEMP)
neus_spring <- neus_spring %>%
  mutate(stratum = as.double(stratum),
         latitude = as.double(latitude),
         longitude = as.double(longitude),
         depth = as.double(depth),
         wgt = as.double(wgt),
         num = as.double(num),
         year = as.double(year),
         quarter = case_when(month %in% c(1,2,3) ~ 1,
                             month %in% c(4,5,6) ~ 2,
                             month %in% c(7,8,9) ~ 3,
                             month %in% c(10,11,12) ~ 4),
         season = "Spring",
         SVSPP = as.double(SVSPP),
         haul_dur = as.numeric(TOWDUR)/60, #minutes to hours,
         gear = SVGEAR
)


#apply spring conversion factors
setDT(neus_spring)

dcf.spp <- NEFSC_conv[DCF_WT > 0, SVSPP]

for(i in 1:length(dcf.spp)){
  neus_spring[year < 1985 & SVSPP == dcf.spp[i],
              wgt := wgt * NEFSC_conv[SVSPP == dcf.spp[i], DCF_WT]]
```

```
}

gcf.spp <- NEFSC_conv[GCF_WT > 0, SVSPP]
for(i in 1:length(gcf.spp)){
  neus_spring[year > 1972 & year < 1982 & SVSPP == gcf.spp[i],
              wgt := wgt / NEFSC_conv[SVSPP == gcf.spp[i], GCF_WT]]
}

vcf.spp <- NEFSC_conv[VCF_WT > 0, SVSPP]
for(i in 1:length(dcf.spp)){
  neus_spring[SVVESSEL == 'DE' & SVSPP == vcf.spp[i],
              wgt := wgt* NEFSC_conv[SVSPP == vcf.spp[i], VCF_WT]]
}

spp_spring <- big_spring[season == 'spring', svspp]
#before <- neus_spring[SVVESSEL %in% c('HB', 'PC') & SVSPP %in% spp_spring,
#.(mean_wtcpue=mean(wtcpue)), by=SVSPP][order(SVSPP)]

for(i in 1:length(big_spring$svspp)){
  neus_spring[SVVESSEL %in% c('HB', 'PC') & SVSPP == spp_spring[i],
              wgt := wgt / big_spring[i, rhoW]]
}

#after <- neus_spring[SVVESSEL %in% c('HB', 'PC')  & SVSPP %in% spp_spring,
#.(mean_wtcpue=mean(wtcpue)), by=SVSPP][order(SVSPP)]


neus_spring <- as.data.frame(neus_spring)

# sum different sexes of same spp together
neus_spring <- neus_spring %>%
  group_by(year, latitude, longitude, depth, haul_id, CRUISE6, station, stratum,
           verbatim_name, gear, haul_dur) %>%
  mutate(wgt = sum(wgt)) %>%
  ungroup()

#-------------------------------------------------------------------------------------------#
#### REFORMAT AND MERGE DATA FILES ####
#-------------------------------------------------------------------------------------------#


#join with strata
neus_spring <- left_join(neus_spring, neus_strata, by = "stratum")
neus_spring <- filter(neus_spring, !is.na(stratum_area))
neus_spring <- neus_spring %>%
  rename(stratumarea = stratum_area) %>%
  #convert square nautical miles to square kilometers
  mutate(stratumarea = as.double(stratumarea)* 3.429904)
neus_spring$survey <- "NEUS"

neus_spring<- neus_spring %>%
  mutate(
    source = "NOAA",
```

```
    timestamp = mdy("03/01/2021"),
    country = "United States",
    continent = "n_america",
    sub_area = NA,
    stat_rec = NA,
    area_swept = 0.0384, #Average tow area in km^2 for albatross
    #num_h could be calculated by dividing num by haul duration,
    #but use caution because of 2008-2009 conversion
    num_h = NA,
    #num_cpue could be calculated by dividing num by area swept,
    #but use caution because of 2008-2009 conversion
    num_cpue = NA,
    #wgt_h could be calculated by dividing wgt by haul duration,
    #but use caution because of 2008-2009 conversion
    wgt_h = NA,
    #wgt_cpue could be calculated by dividing wgt by area swept,
    #but use caution because of 2008-2009 conversion
    wgt_cpue = NA
  ) %>%
  select(survey, haul_id,source, timestamp, country, sub_area, continent, stat_rec, station,
         stratum, year, month, day, quarter, season, latitude, longitude,
         haul_dur, area_swept, gear, depth, sbt, sst,
         num, num_h, num_cpue, wgt, wgt_h, wgt_cpue, verbatim_name)

rm(neus_catch_clean, neus_catch_raw, neus_spring_catch, neus_spring_haul)

#############
#bind spring and fall

neus <- rbind(neus_fall, neus_spring)

#sum abundance and wgt to fix duplicates (removes 21680 rows)
neus <- neus %>%
  group_by(survey, haul_id,source, timestamp, country, sub_area, continent, stat_rec, station,
                    stratum, year, month, day, quarter, season, latitude, longitude,
                    haul_dur, area_swept, gear, depth, sbt, sst,verbatim_name) %>%
  summarise(num=sum(num, na.rm = T),
         num_h=sum(num_h, na.rm = T),
         num_cpue=sum(num_cpue, na.rm = T),
         wgt=sum(wgt, na.rm = T),
         wgt_h=sum(wgt_h, na.rm = T),
         wgt_cpue=sum(wgt_cpue, na.rm = T)) %>%
select(survey, haul_id,source, timestamp, country, sub_area, continent, stat_rec, station,
       stratum, year, month, day, quarter, season, latitude, longitude,
       haul_dur, area_swept, gear, depth, sbt, sst,
       num, num_h, num_cpue, wgt, wgt_h, wgt_cpue, verbatim_name)

#check for duplicates, should not be any with more than 1 obs
#check for duplicates
count_neus <- neus %>%
  group_by(haul_id, verbatim_name) %>%
  mutate(count = n())
```

```
#none!

#which ones are duplicated?
unique_name_match <- count_neus %>%
  group_by(verbatim_name) %>%
  filter(count>1) %>%
  distinct(verbatim_name)

unique_name_match
#check if empty

##duplicates
#HOMARUS AMERICANUS
#SQUALUS ACANTHIAS
#MUSTELUS CANIS
#GERYON QUINQUEDENS
#OVALIPES STEPHENSONI
#MAJIDAE
#OVALIPES OCELLATUS
#MYLIOBATIS GOODEI
#CALLINECTES SAPIDUS
#PORTUNIDAE
#CANCER IRRORATUS
#LIMULUS POLYPHEMUS
#CANCER BOREALIS
#SQUATINA DUMERIL
#BATHYNECTES LONGISPINA
#PANDALUS BOREALIS
#LITHODES MAJA
#PORTUNUS SPINIMANUS
#ARENAEUS CRIBRARIUS
#UNIDENTIFIED FISH
#ACANTHOCARPUS ALEXANDRI
#CHIONOECETES OPILIO
#SCYLIORHINUS RETIFER
#LOLIGO PEALEII
#CANCRIDAE

#-------------------------------------------------------------------------------------#
#### INTEGRATE CLEAN TAXA FROM TAXA ANALYSIS ####
#-------------------------------------------------------------------------------------#

# Get WoRM's id for sourcing
wrm <- gnr_datasources() %>%
  filter(title == "World Register of Marine Species") %>%
  pull(id)

### Automatic cleaning
# Set Survey code
neus_survey_code <- "NEUS"

neus <- neus %>%
  mutate(
```

```
    taxa2 = str_squish(verbatim_name),
    taxa2 = str_remove_all(taxa2," spp.| sp.| spp| sp|NO "),
    taxa2 = str_to_sentence(str_to_lower(taxa2))
  )

# Get clean taxa
clean_auto <- clean_taxa(unique(neus$taxa2), input_survey = neus_survey_code,
                         save = F, output=NA)

#This leaves out the following species, of which 1 is a fish that needs to be added back
#Geryon quinquedens                                    no match
#Astroscopus y-graecum                                 no match      #FISH
#Portunus spinimanus                                   no match
#Pandalus propinquus                                   no match

ast_ygr <- c("Astroscopus y-graecum", 159252,3704, "Astroscopus y-graecum",
             "Animalia","Chordata","Actinopteri","Perciformes","Uranoscopidae",
             "Astroscopus","Species","NEUS")

clean_auto_missing <- rbind(clean_auto, ast_ygr)


#-------------------------------------------------------------------------------------------#
#### INTEGRATE CLEAN TAXA in NEUS survey data ####
#-------------------------------------------------------------------------------------------#

correct_taxa <- clean_auto_missing %>%
  select(-survey)

clean_neus <- left_join(neus, correct_taxa, by=c("taxa2"="query")) %>%
  filter(!is.na(taxa)) %>% # query does not indicate taxa entry that
  #were removed in the cleaning procedure
  # so all NA taxa have to be removed from the surveys because:
  #non-existing, non marine or non fish
  rename(accepted_name = taxa) %>%
  mutate(verbatim_aphia_id = NA,
         aphia_id = worms_id
         ) %>%
  select(survey, haul_id, source, timestamp, country, sub_area, continent, stat_rec,
         station, stratum,
         year, month, day, quarter, season, latitude, longitude, haul_dur, area_swept,
         gear, depth, sbt, sst, verbatim_name, num, num_h, num_cpue,
         wgt, wgt_h, wgt_cpue, verbatim_name, verbatim_aphia_id, accepted_name,
         aphia_id, SpecCode,
         kingdom, phylum, class, order, family, genus, rank)

#check for duplicates
count_clean_neus <- clean_neus %>%
  group_by(haul_id, accepted_name) %>%
  mutate(count = n())

#none!
```

```r
#which ones are duplicated?
unique_name_match <- count_clean_neus %>%
  group_by(verbatim_name, accepted_name) %>%
  filter(count>1) %>%
  distinct(verbatim_name, accepted_name)

unique_name_match
#check if empty



# ------------------------------------------------------------------------------------------#
#### SAVE DATABASE IN GOOGLE DRIVE ####
# ------------------------------------------------------------------------------------------#

# Just run this routine should be good for all
write_clean_data(data = clean_neus, survey = "NEUS", overwrite = T)
```

1. Overview of the survey data table

survey	haul_id	source	timestamp	country	sub_area	continent
NEUS	1.96307e+17	NOAA	2021-03-01	United States	NA	n_america
NEUS	1.96307e+17	NOAA	2021-03-01	United States	NA	n_america
NEUS	1.96307e+17	NOAA	2021-03-01	United States	NA	n_america
NEUS	1.96307e+17	NOAA	2021-03-01	United States	NA	n_america
NEUS	1.96307e+17	NOAA	2021-03-01	United States	NA	n_america

stat_rec	station	stratum	year	month	day	quarter	season
NA	157	1010	1963	12	10	4	Fall
NA	157	1010	1963	12	10	4	Fall
NA	157	1010	1963	12	10	4	Fall
NA	157	1010	1963	12	10	4	Fall
NA	157	1010	1963	12	10	4	Fall

latitude	longitude	haul_dur	area_swept	gear	depth
40.73333	-72.65	0.5	0.0384	11	28
40.73333	-72.65	0.5	0.0384	11	28
40.73333	-72.65	0.5	0.0384	11	28
40.73333	-72.65	0.5	0.0384	11	28
40.73333	-72.65	0.5	0.0384	11	28

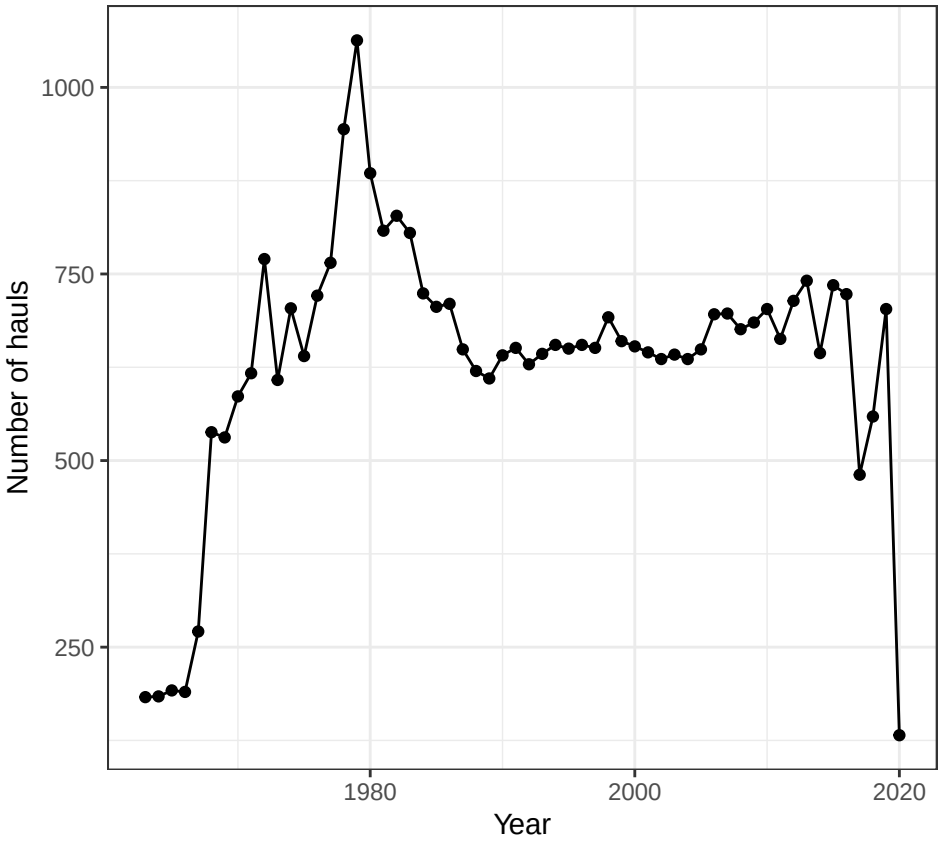
sbt	sst	verbatim_name	num	num_h	num_cpue
8.1	8.2	ALOSA PSEUDOHARENGUS	1	0	0
8.1	8.2	BREVOORTIA TYRANNUS	2	0	0
8.1	8.2	CLUPEA HARENGUS	1	0	0
8.1	8.2	GADUS MORHUA	3	0	0
8.1	8.2	LEUCORAJA ERINACEA	162	0	0

wgt	wgt_h	wgt_cpue	verbatim_aphia_id	accepted_name
0.500	0	0	NA	Alosa pseudoharengus
0.000	0	0	NA	Brevoortia tyrannus
0.000	0	0	NA	Clupea harengus
13.284	0	0	NA	Gadus morhua
91.256	0	0	NA	Leucoraja erinacea

aphia_id	SpecCode	kingdom	phylum	class	order	family
158669	1583	Animalia	Chordata	Actinopteri	Clupeiformes	Clupeidae
158691	1592	Animalia	Chordata	Actinopteri	Clupeiformes	Clupeidae
126417	24	Animalia	Chordata	Actinopteri	Clupeiformes	Clupeidae
126436	69	Animalia	Chordata	Actinopteri	Gadiformes	Gadidae
158551	2557	Animalia	Chordata	Elasmobranchii	Rajiformes	Rajidae

2. Summary of sampling intensity

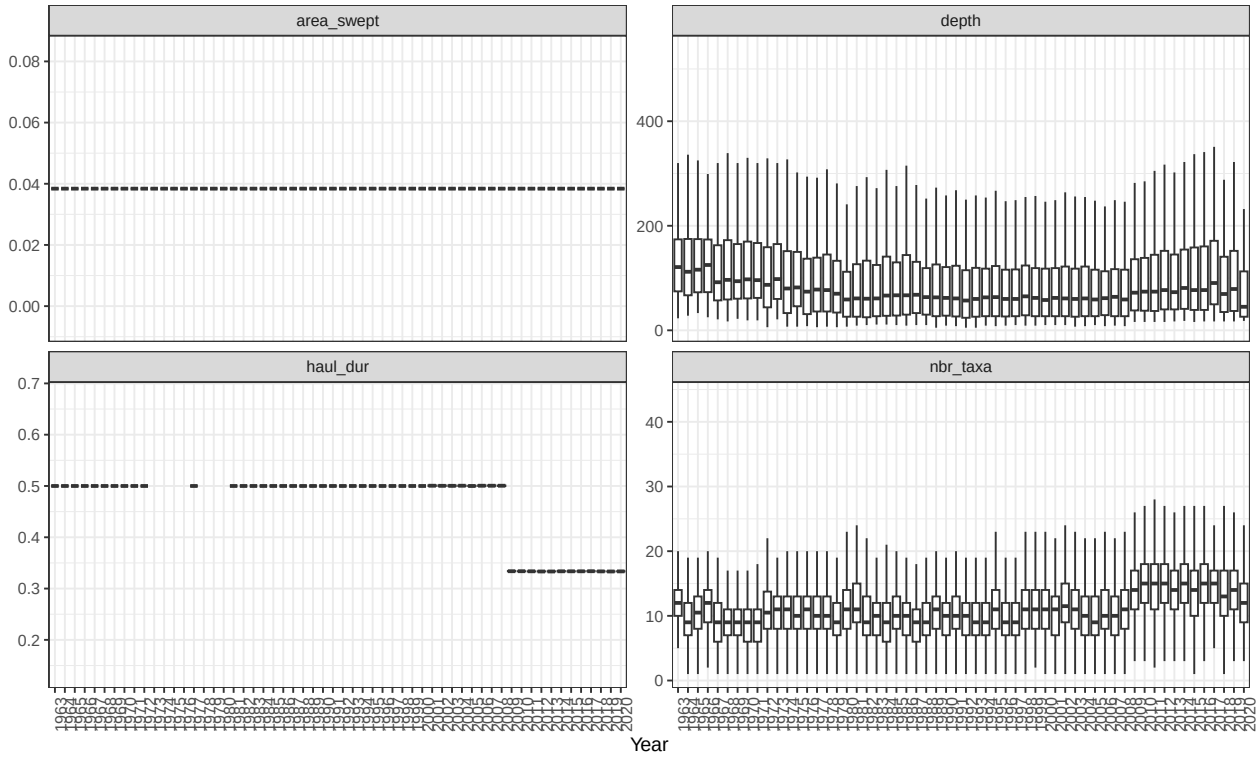
Number of hauls per year performed during the survey after data processing.



3. Summary of sampling variables from the survey

Here we show the yearly total and average of the following variables reported in the survey data:

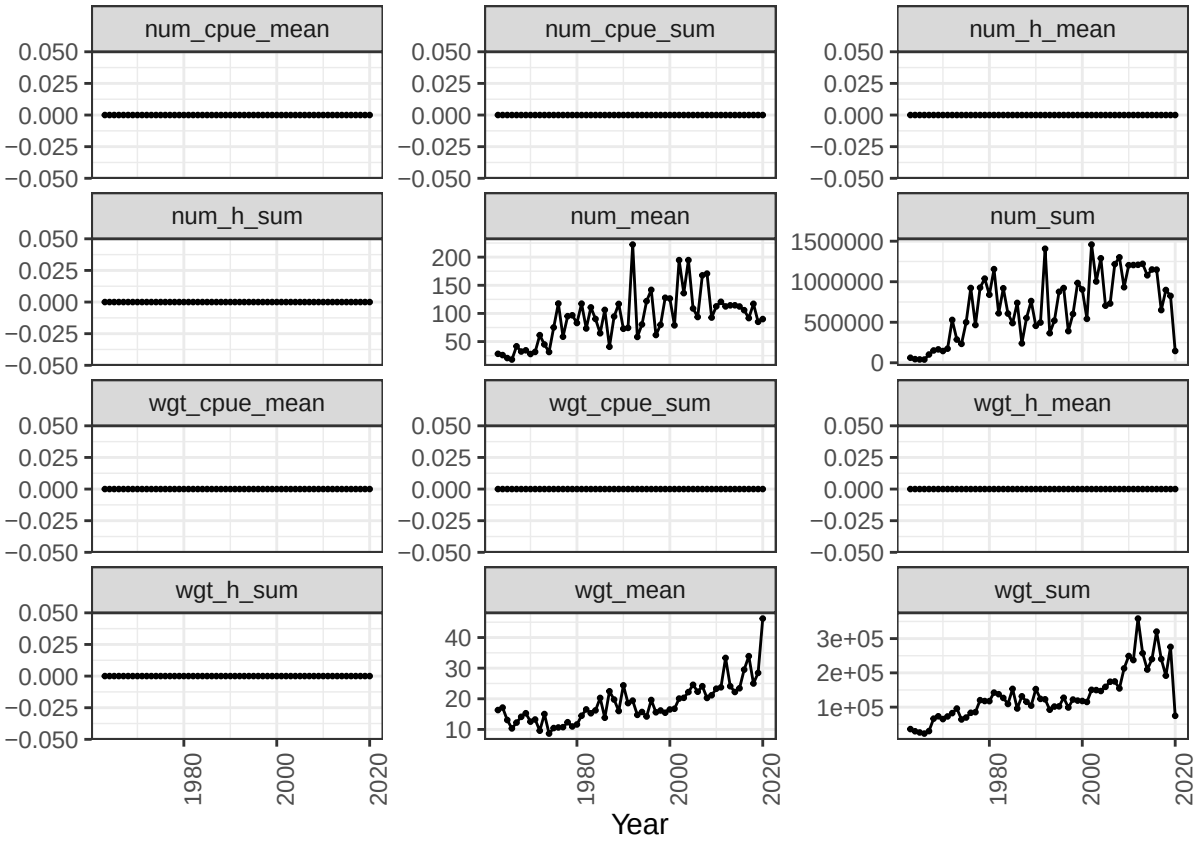
- *area_swept*, swept area by the bottom trawl gear km^2
- *depth*, sampling depth in m
- *haul_dur*, haul sampling duration *hour*
- *number of marine fish taxa*, taxa were cleaned following the last version of taxonomy from the World Register of Marine Species (https://www.marinespecies.org/, October 2021)



4. Summary of biological variables

Here we display the yearly total and average across hauls of the following variables recorded in the data:

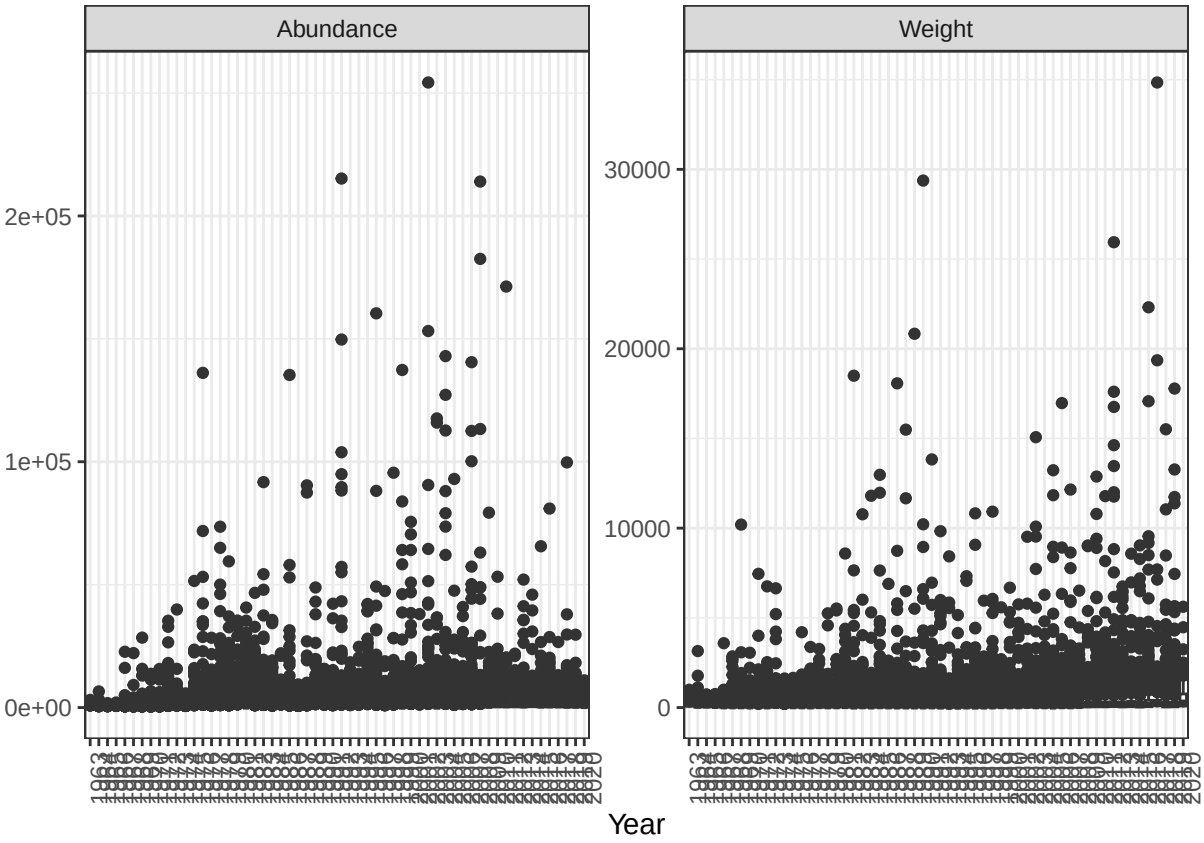
- *num_cpue*, number of individuals (abundance) in $\frac{individuals}{km^2}$
- *num_h*, number of individuals (abundance) in $\frac{individuals}{h}$
- *num*, number of individuals (abundance)
- *wgt_cpue*, weight in $\frac{kg}{km^2}$
- *wgt_h*, weight in $\frac{kg}{h}$
- *wgt*, weight in kg



5. Extreme values

Here we show a yearly total distribution of the biomass data to visualize outliers:

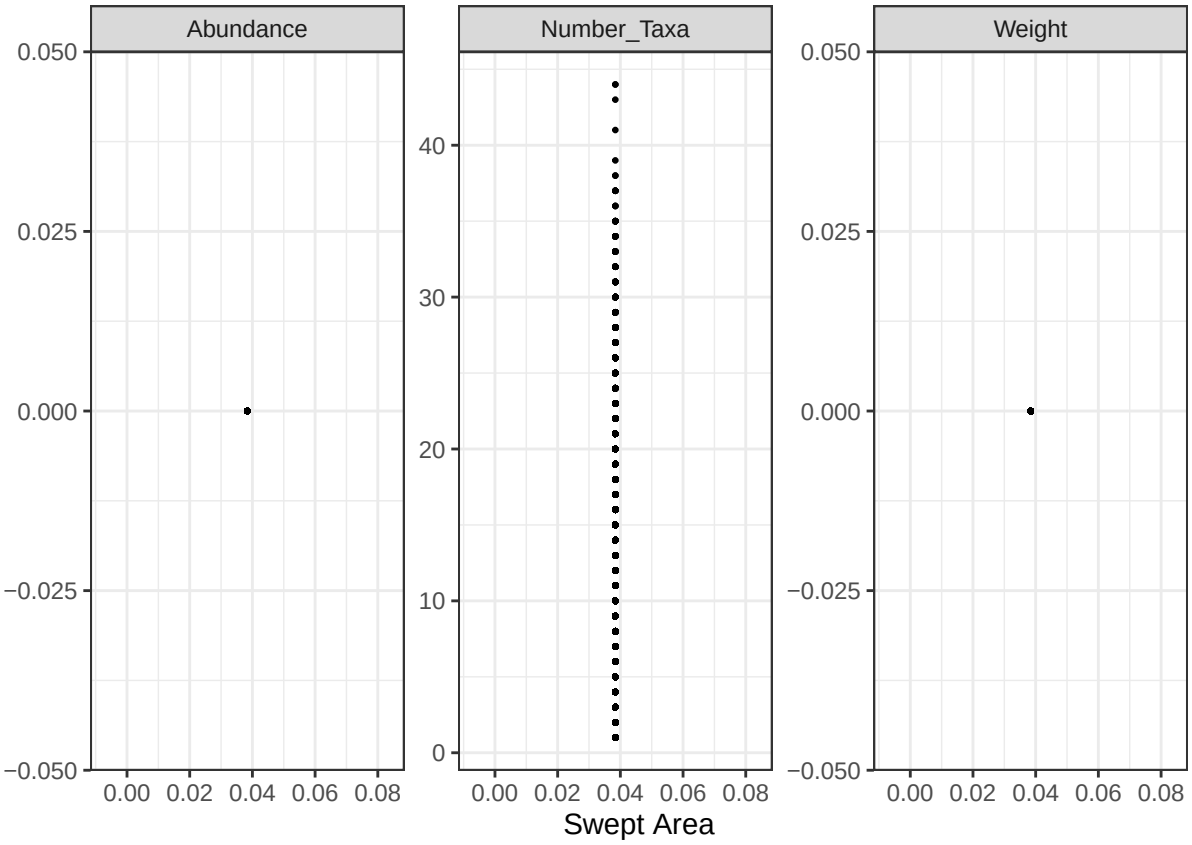
- *wgt*, total weight in *kg* per haul and year per haul and year, if available in the survey data
- *num*, total number of individuals, if available in the survey data



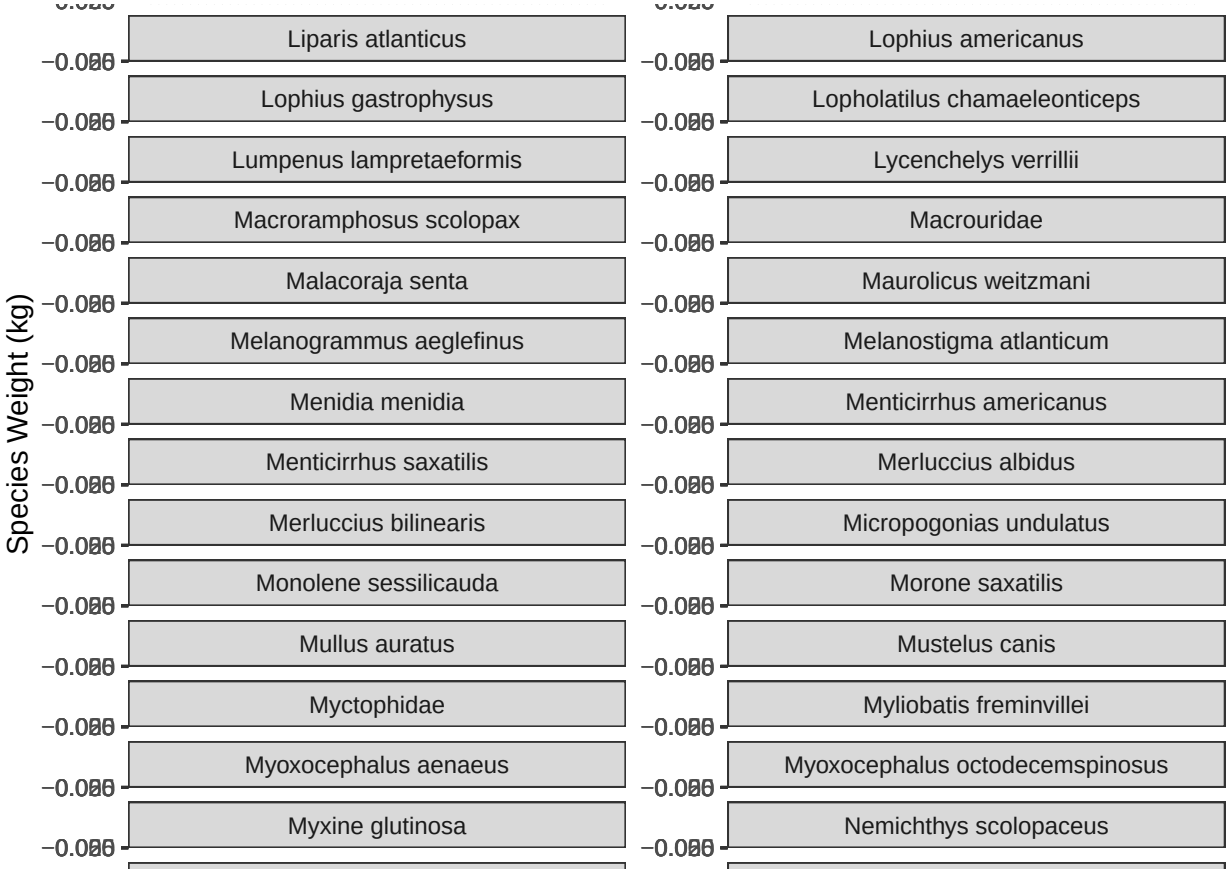
6. Summary of variables against swept area

Here we show the total abundance and number of taxa relationships with the area swept:

- *nbr_taxa*, number of marine fish taxa after taxonomic data cleaning
- *num_cpue*, number of individuals (abundance) in $\frac{individuals}{km^2}$
- *wgt_cpue*, weight in $\frac{kg}{km^2}$

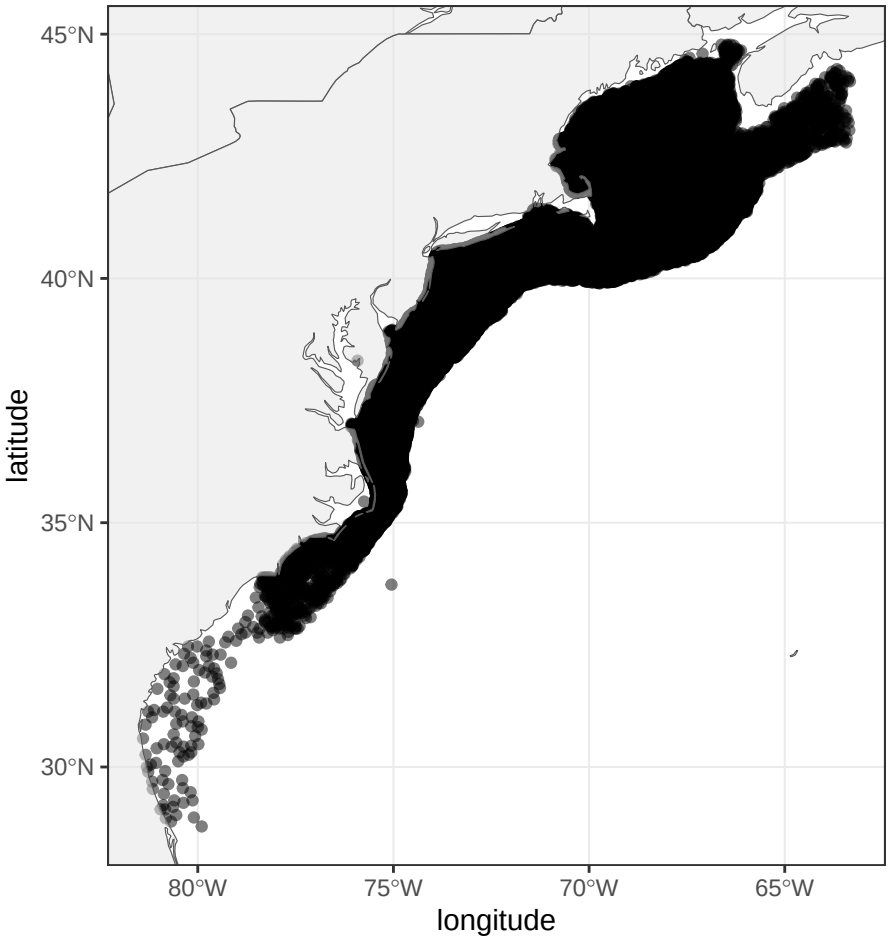


7. Abundance or Weight trends of the six most abundant species



8. Distribution mapping

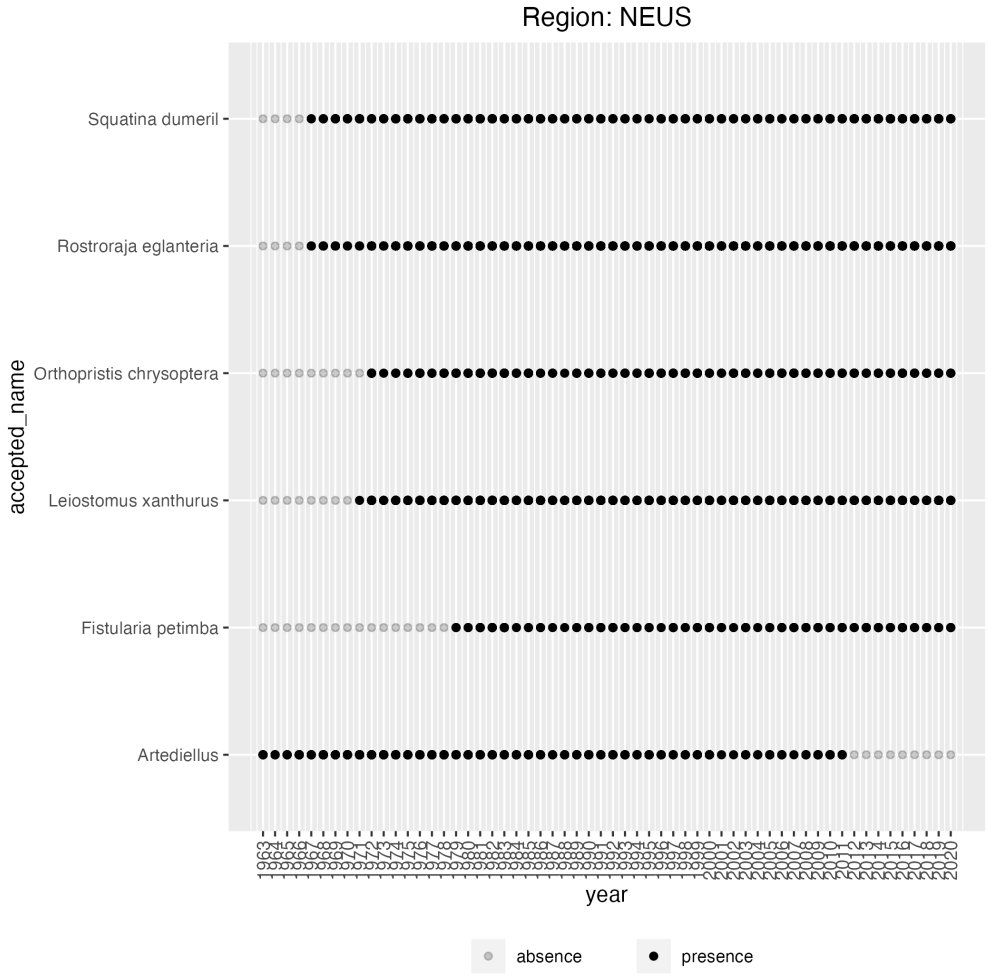
Map of the sampling distribution in space. Note that we only show one year per coordinate.



9. Taxonomic flagging

This species flagging method was adapted from https://github.com/pinskylab/OceanAdapt/blob/master/R/add-spp-to-taxonomy.Rmd#L33

Visualization of flagged taxa



Statistics related to the taxonomic flagging outputs

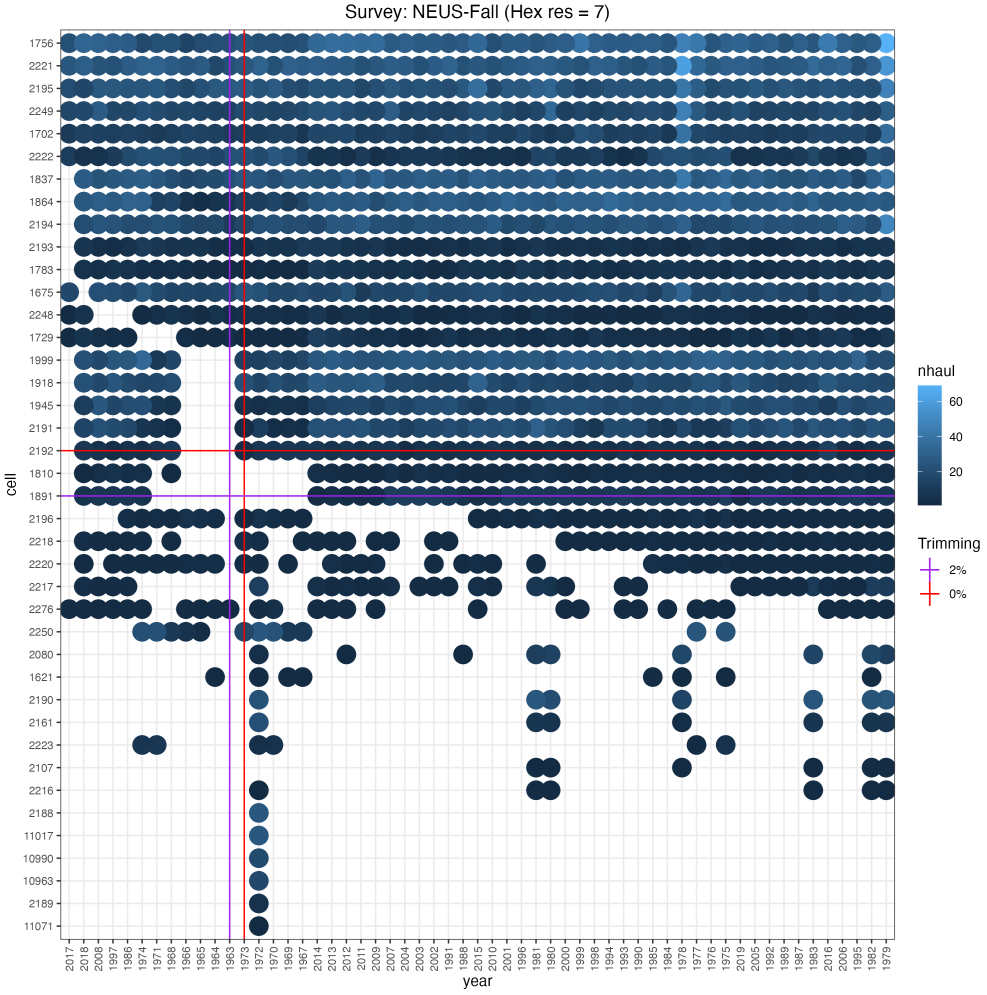
Total number of species	560.0
Percentage of species flagged	1.1

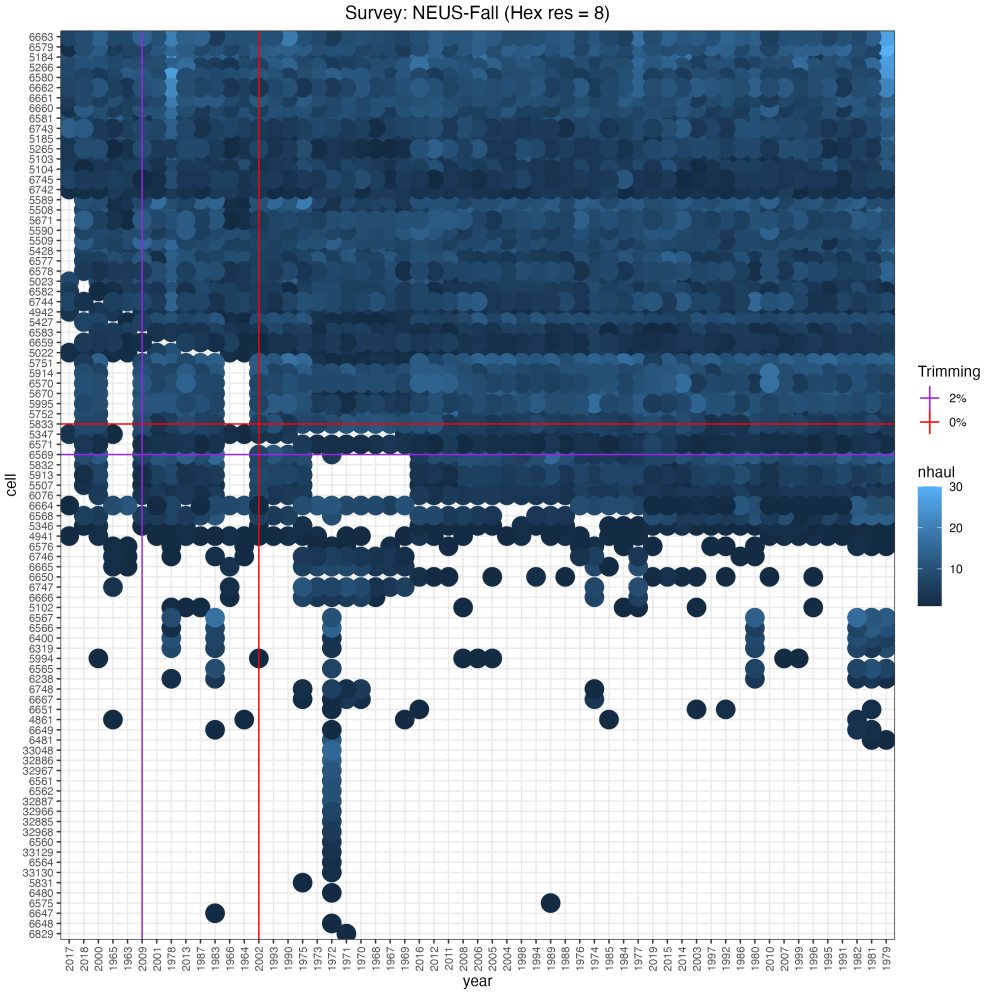
10. Spatio-temporal standardization: NEUS-Fall

a. Standardization method 1

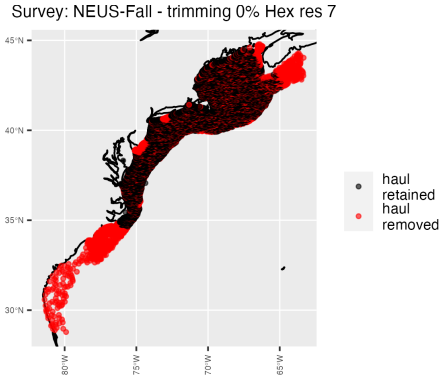
This standardization method was adapted from https://github.com/zoekitchel/trawl_spatial_turnover/blob/master/data_prep_code/species/explore_NorthSea_trimming.Rmd
It was run for hex resolution 7 and 8.

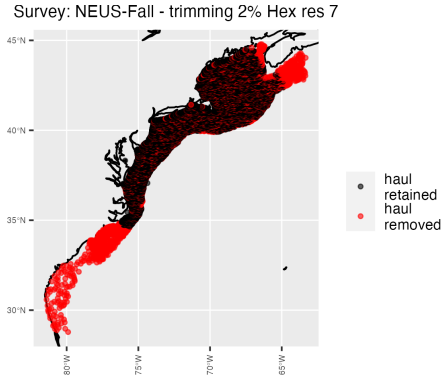
Plot of number of cells x years with overlaid flagging options

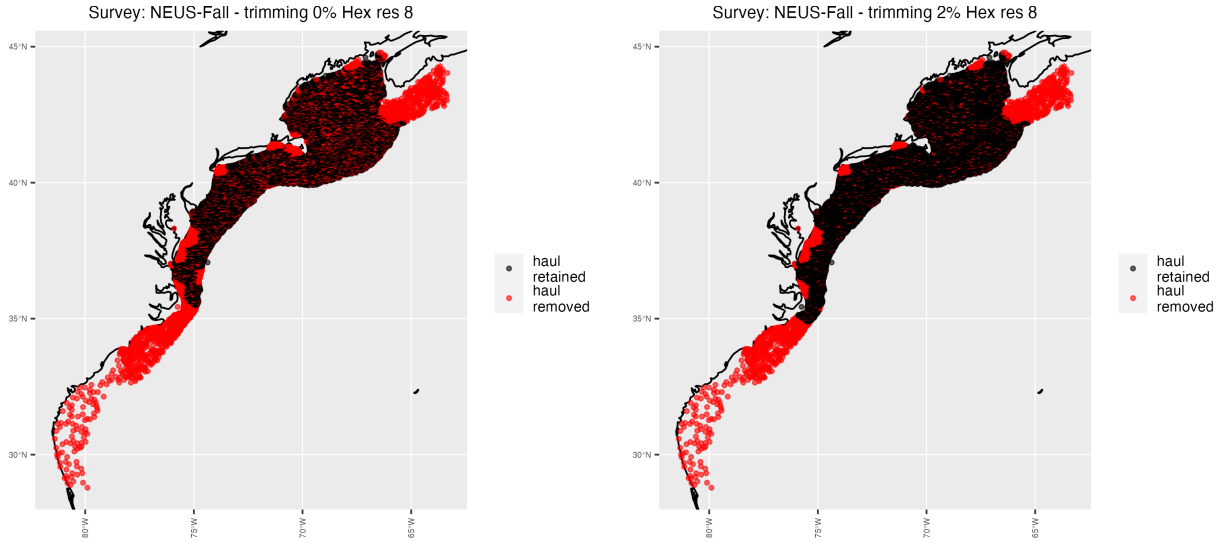




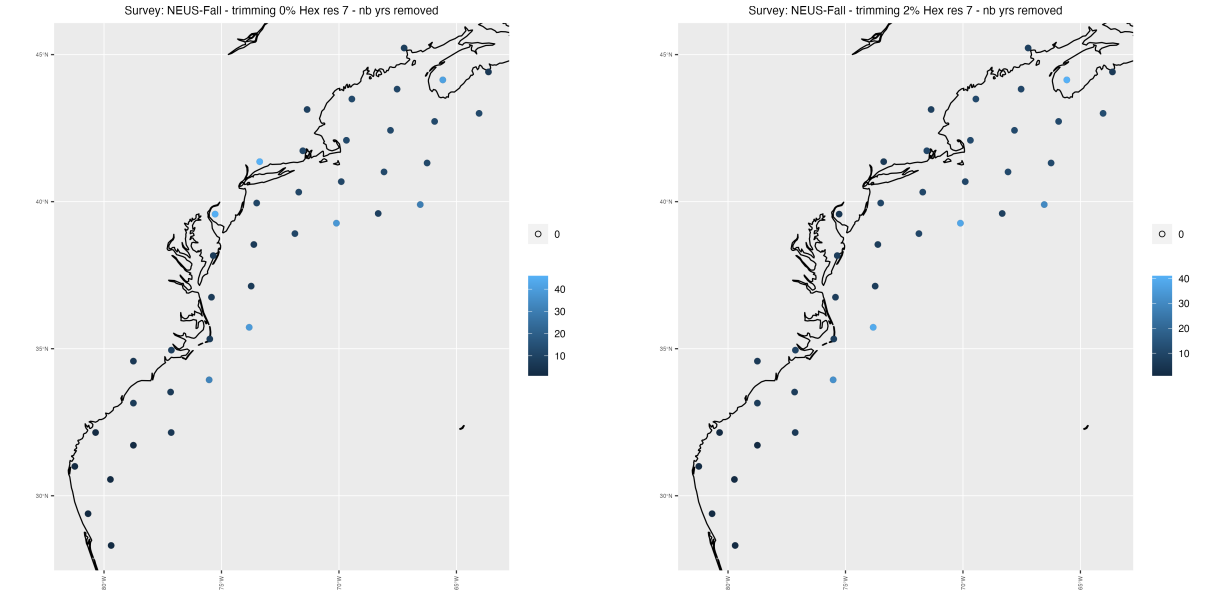
Map of hauls retained and removed per flagging method and threshold

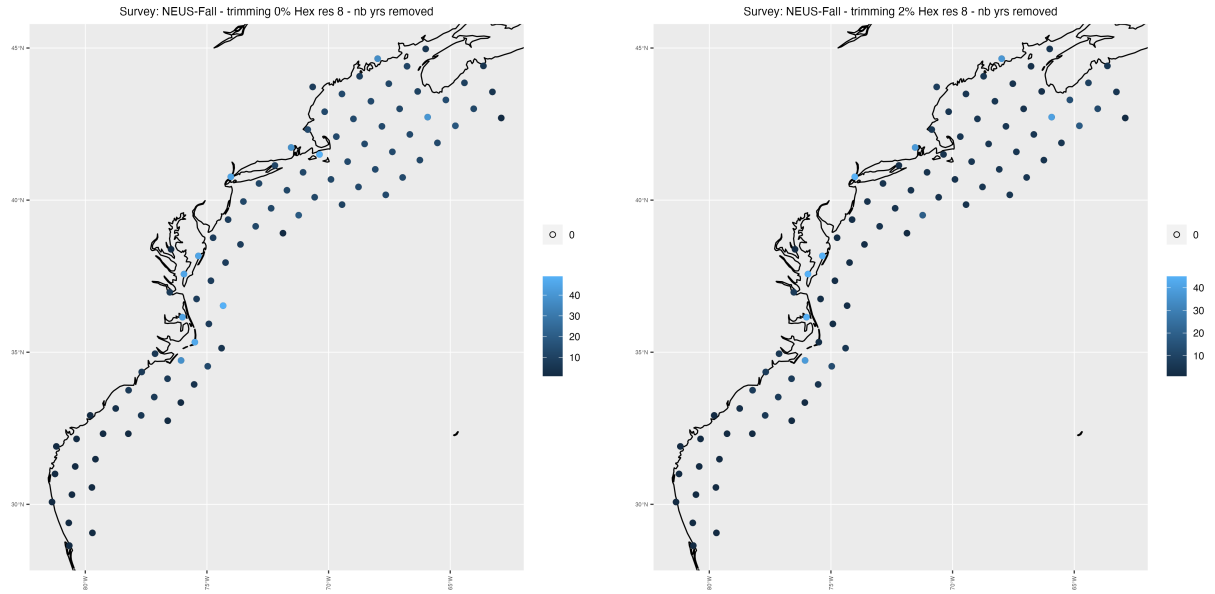






Map of numbers of years removed per grid cell and flagging method/threshold

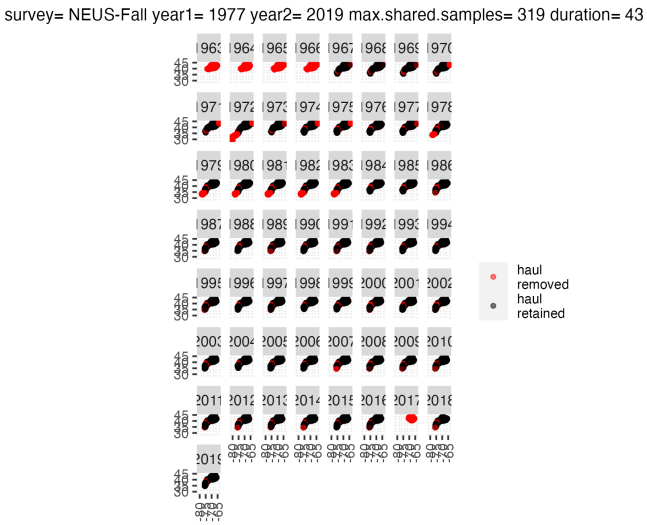




b. Standardization method 2

This standardization method was adapted from BioTIME code from https://github.com/Wubing-Xu/Range_size_winners_losers

Map of hauls retained and removed



c. Standardization summary

Statistics of hauls removed for each standardization method

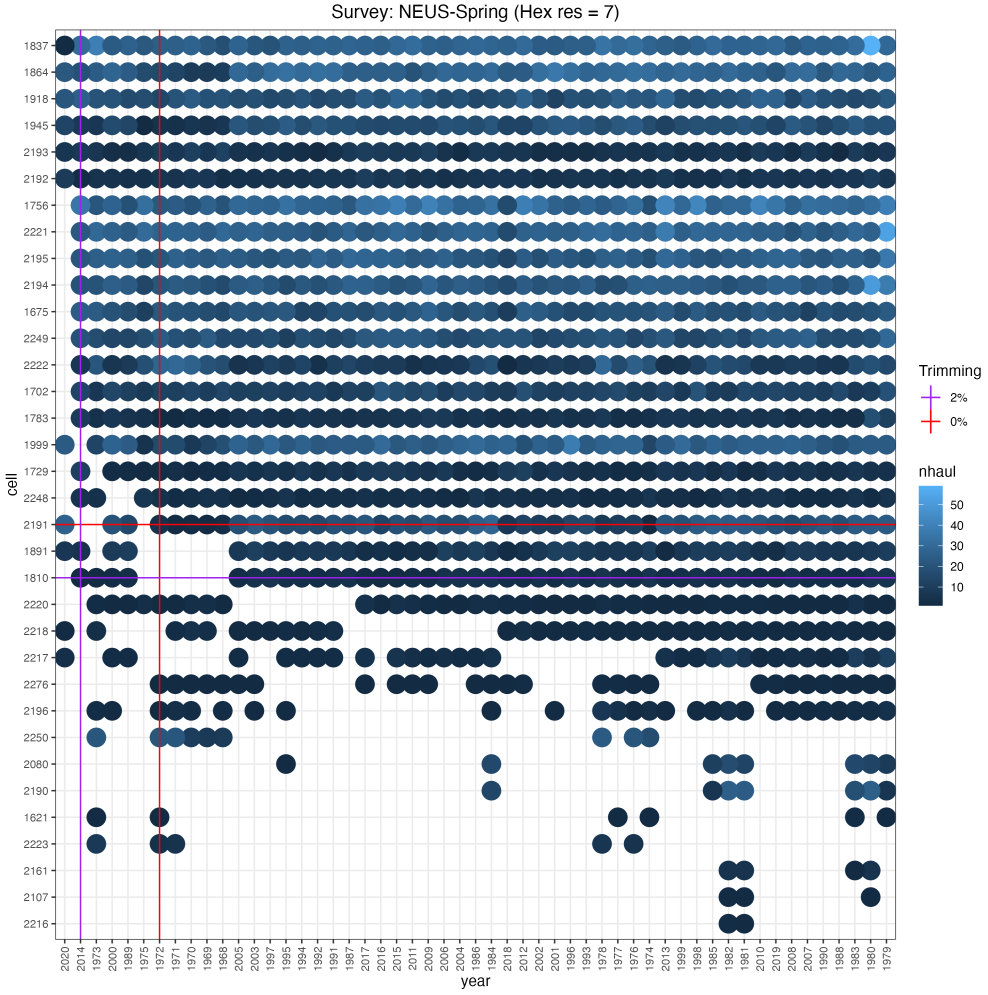
summary	grid cell 7, 0% threshold	grid cell 7, 2% threshold	grid cell 8, 0% threshold	grid cell 8, 2% threshold	method 2 (biotime)
number of hauls removed	4466.0	3854.0	6293.0	3341.0	30151.0
percentage of hauls removed	23.5	20.3	33.1	17.6	13.8

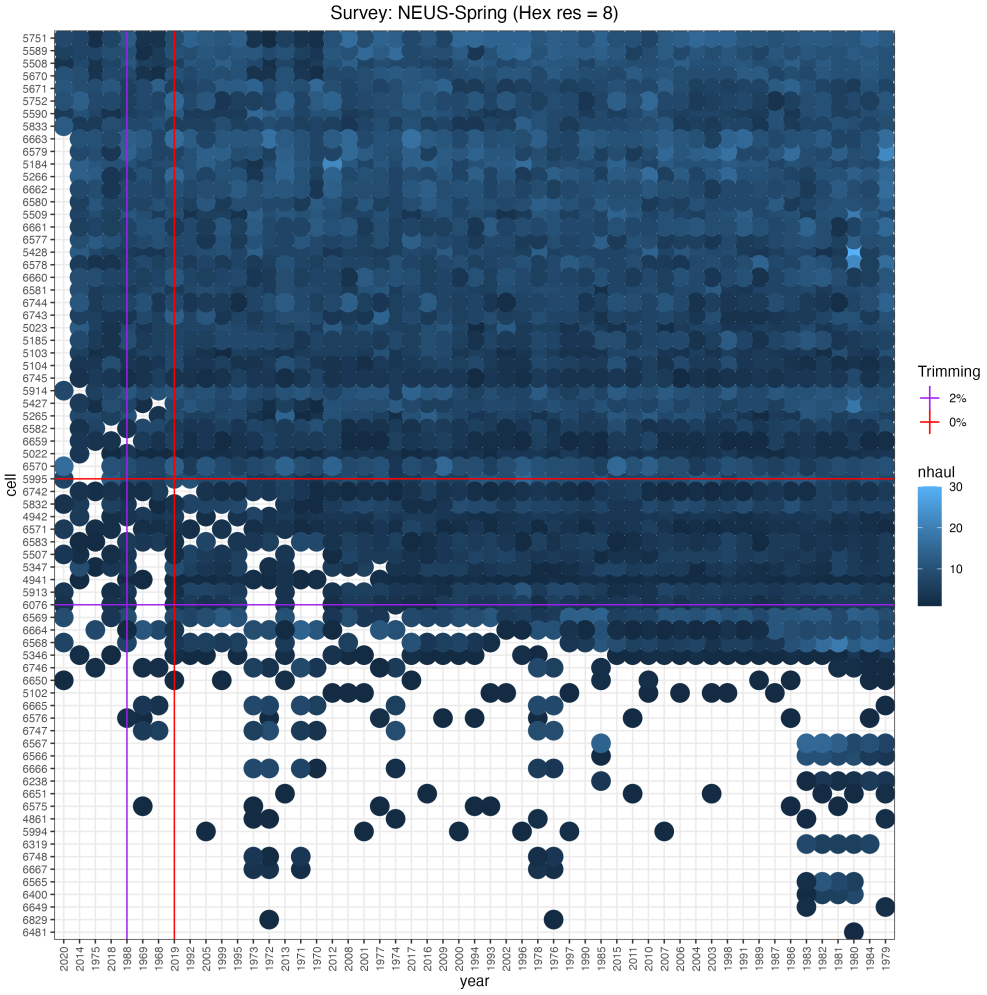
11. Spatio-temporal standardization: NEUS-Spring

a. Standardization method 1

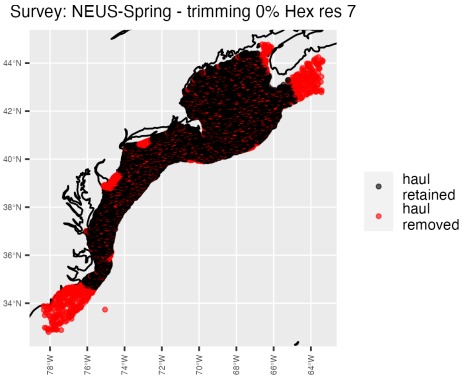
This standardization method was adapted from https://github.com/zoekitchel/trawl_spatial_turnover/blob/master/data_prep_code/species/explore_NorthSea_trimming.Rmd
It was run for hex resolution 7 and 8.

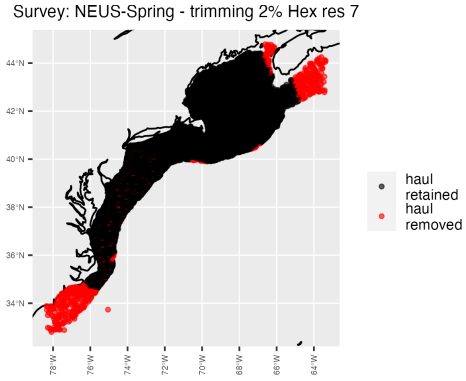
Plot of number of cells x years with overlaid flagging options

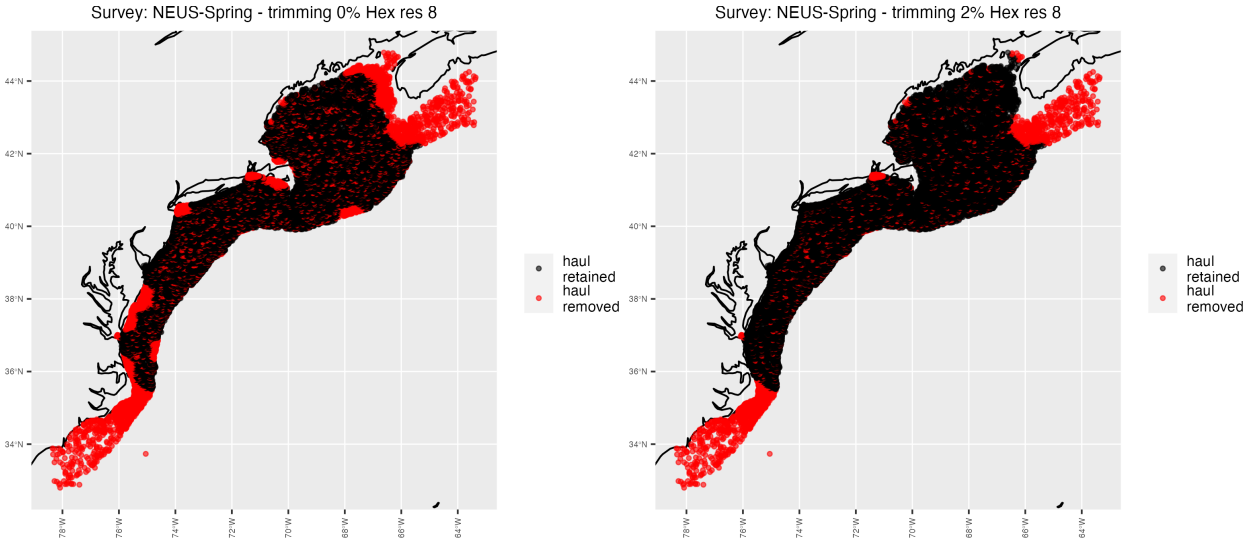




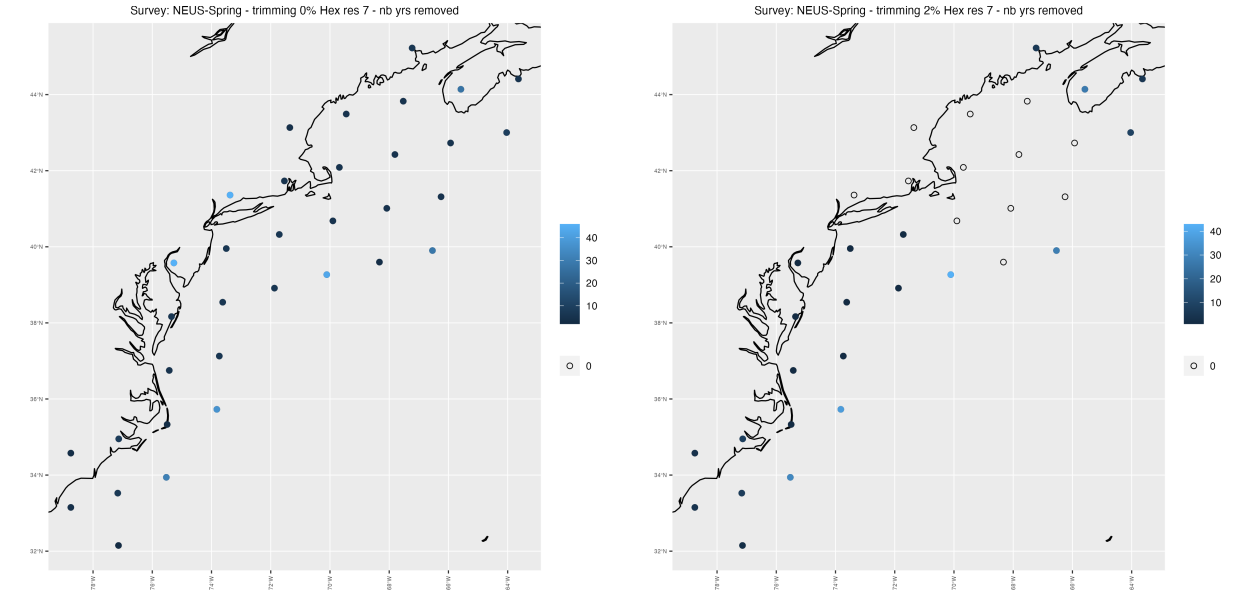
Map of hauls retained and removed per flagging method and threshold

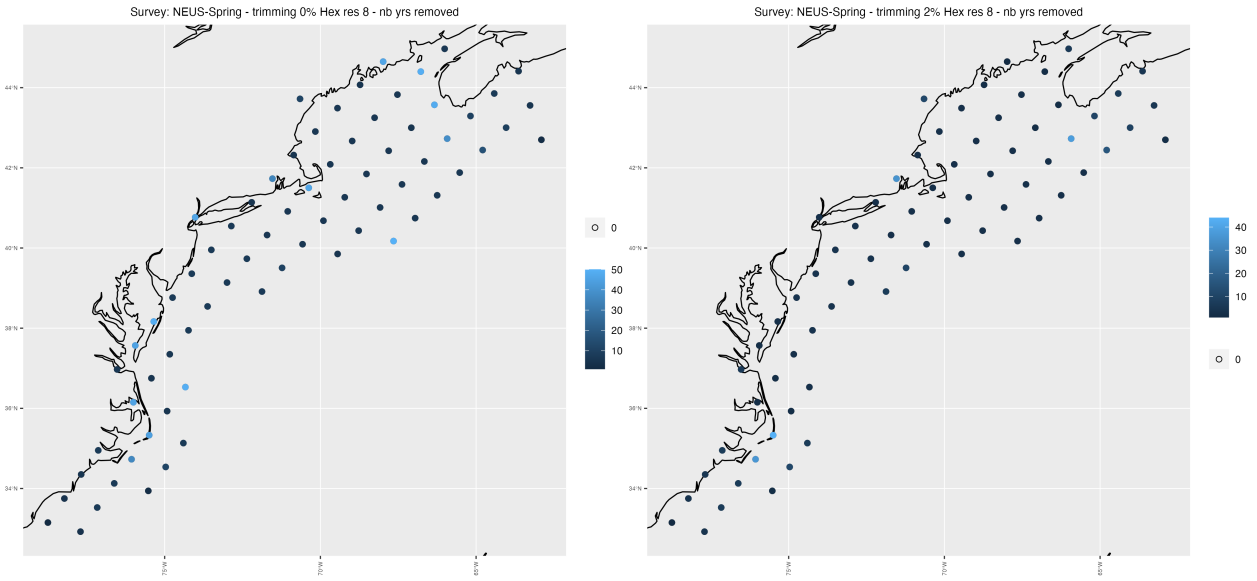






Map of numbers of years removed per grid cell and flagging method/threshold

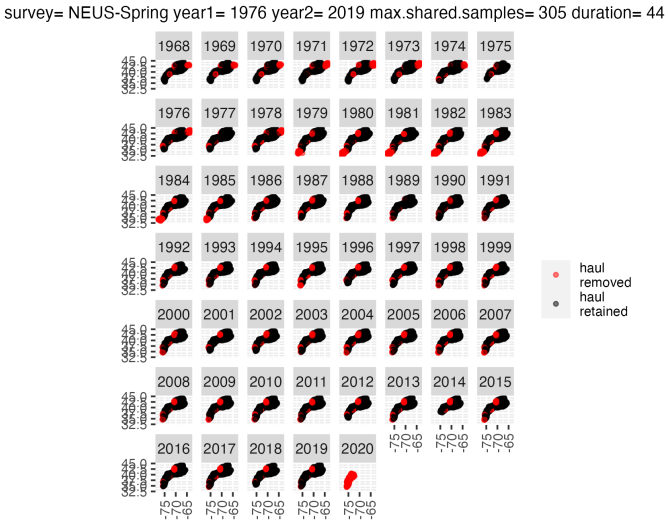




b. Standardization method 2

This standardization method was adapted from BioTIME code from https://github.com/Wubing-Xu/Range_size_winners_losers

Map of hauls retained and removed



c. Standardization summary

Statistics of hauls removed for each standardization method

summary	grid cell 7, 0% threshold	grid cell 7, 2% threshold	grid cell 8, 0% threshold	grid cell 8, 2% threshold	method 2 (biotime)
number of hauls removed	2706.0	876.0	4835.0	2344.0	21518.0
percentage of hauls removed	15.2	4.9	27.2	13.2	11.1