# GSL-S: Gulf of St. Lawrence South (Canada) survey data processing summary

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

September, 2023

## Contents



### General info

This document presents the cleaning code and summary of the Gulf of St. Lawrence South (Canada) bottom trawl survey provided by Department of Fisheries and Oceans Canada. It contains data from 1970 and up to 2019.

### Data cleaning in R

```
##############################################################################################
#### R code to clean trawl survey for Gulf of St. Lawrence South
#### Public data Ocean Adapt
#### Contacts: Government of Canada; Fisheries and Oceans Canada
####gddaiss-dmsaisb.XLAU@dfo-mpo.gc.ca
#### Coding: Michelle Stuart, Dan Forrest, Zoë Kitchel December 2022
##############################################################################################
#NB: there are multiple events at similar locations on the same day because there
#is more than one vessel sampling, keep an eye on vessel name and haul_id

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

library(tidyverse)
library(lubridate)
library(googledrive)
library(taxize) # for getting correct species names
library(magrittr) # for names wrangling
library(readr)
library(dplyr)
library(PBSmapping)
library(readxl)

source("functions/clean_taxa.R")
source("functions/write_clean_data.R")
source("functions/apply_trimming_method1.R")
source("functions/apply_trimming_method2.R")
source("functions/flag_spp.R")
fishglob_data_columns <- read_excel("standard_formats/fishglob_data_columns.xlsx")

########Types of gear:
########"Western IIA trawl width 12.497m": 1987-2019; 0.041 km^2 in 30 minutes
#(avg trawl length)
########"Yankee #36 otter trawl width 10.668m" 1970-1986; 0.035 km^2 in 30 minutes
#(avg trawl length)
#Source: Page 11; https://waves-vagues.dfo-mpo.gc.ca/Library/115732.pdf

#Southern GSL -
#  Trawl Distance, 1.75 nau mi. 30 minute tow at 3.5 knots (via Daniel Ricard)

#We need to check to confirm that data we use are corrected for the gear change
#described above (Zoe, 22 Nov 2022)

#Data for the Gulf of St. Lawrence South can be accessed using the public
#Pinsky Lab OceanAdapt Git Hub Repository.
#Contact malin.pinsky@rutgers.edu for questions or help accessing


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

```r
GSLsouth <- read_csv(
  "https://github.com/pinskylab/OceanAdapt/raw/master/data_raw/GSLsouth.csv")

GSLsouth$haul_id <- paste(GSLsouth$year,GSLsouth$month,GSLsouth$day,
                          GSLsouth$start.hour,GSLsouth$start.minute, GSLsouth$longitude,
                          GSLsouth$latitude, sep="-")

GSLsouth <- GSLsouth %>%
  mutate(
    wgt = weight.caught,
    num = number.caught,
    sub_area = NA,
    depth = NA, #No depth data available - fill with NA
    station = NA,
    sst = NA,
    sbt = NA,
    season = NA_character_,
    haul_dur = 0.5, #hours
    area_swept = ifelse(gear.str == "Western IIA trawl", 0.041,
                        ifelse(
                          gear.str == "Yankee #36 otter trawl", 0.035, NA)),
    #average swept area values from document above
    gear = gear.str,
    country = "Canada",
    continent = "n_america",
    stat_rec = NA,
    num_h = num/haul_dur,
    wgt_h = wgt/haul_dur,
    num_cpue = num/area_swept,
    wgt_cpue = wgt/area_swept,
    verbatim_name = latin.name,
    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),
  )


GSLsouth <- GSLsouth %>%
  filter(
 #  remove unidentified spp and non-species
  verbatim_name != "" | !is.na(verbatim_name),
  !grepl("EGG", verbatim_name),
  !grepl("UNIDENTIFIED", verbatim_name)) %>%
  mutate(survey = "GSL-S",
         stratum = NA) %>%
  # add survey column
  select(survey, haul_id, 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 that the number of unique haul_ids * spp combinations is the same
```

```r
#as the number of rows in mar
nrow(GSLsouth) == nrow(unique(GSLsouth[,c("haul_id","verbatim_name")]))

#it's not, so let's see why we have extras
which(duplicated(GSLsouth[,c("haul_id","verbatim_name")]))

#Haul_ID "1994-9-22-8-27--61.6333333333333-46.4166666666667," Gadus morhua
#has two separate observations
#I will delete second observation because it's only 0.486111111 kg and no count info

GSLsouth <- GSLsouth[-31057,] #be sure to only do once!

#try again
which(duplicated(GSLsouth[,c("haul_id","verbatim_name")])) #success!

#correcting order of columns and adding final column
GSLsouth <- GSLsouth %>%
  mutate(verbatim_aphia_id = NA) %>%
  select(survey, haul_id, 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)

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

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

### Automatic cleaning
# Set Survey code
GSLsouth_survey_code <- "GSL-S"

GSLsouth <- GSLsouth %>%
  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(GSLsouth$taxa2),
                         input_survey = GSLsouth_survey_code, save = F,
                         output=NA, fishbase=T)
#takes 1.8 minutes

#This leaves out the following species, all of which are inverts or
#only ID to genus except for Cae cae
#Caelorinchus caelorinchus     (fish)
#Coelenterata
```

```r
#Nereidae
#Rhynchocoela
#Lithothamnium

cae_cae <- c("Caelorinchus caelorinchus", "398381", "1726",
             "Coelorinchus caelorhincus", "Animalia", "Chordata",
             "Actinopteri", "Gadiformes", "Macrouridae", "Coelorinchus", "Species",
             "GSL-S")

clean_auto_missing <- rbind(clean_auto, cae_cae)

#-----------------------------------------------------------------------------------------#
#### INTEGRATE CLEAN TAXA in GSL-South survey data ####
#-----------------------------------------------------------------------------------------#

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

clean_GSLsouth <- left_join(GSLsouth, 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,
         aphia_id = worms_id) %>%
  mutate(verbatim_aphia_id = NA,
         source = "DFO",
         timestamp = "2021",
         num_cpua = num_cpue,
         num_cpue = num_h,
         wgt_cpua = wgt_cpue,
         wgt_cpue = wgt_h,
         survey_unit = ifelse(survey %in% c("BITS","NS-IBTS","SWC-IBTS"),
                              paste0(survey,"-",quarter),survey),
         survey_unit = ifelse(survey %in% c("NEUS","SEUS","SCS","GMEX"),
                              paste0(survey,"-",season),survey_unit)) %>%
  select(fishglob_data_columns$`Column name fishglob`)




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

# Just run this routine should be good for all
write_clean_data(data = clean_GSLsouth, survey = "GSL-S", overwrite = T)



# -----------------------------------------------------------------------------------------#
#### FAGS ####
# -----------------------------------------------------------------------------------------#
```

```r
#install required packages that are not already installed
required_packages <- c("data.table",
                       "devtools",
                       "dggridR",
                       "dplyr",
                       "fields",
                       "forcats",
                       "ggplot2",
                       "here",
                       "magrittr",
                       "maps",
                       "maptools",
                       "raster",
                       "rcompendium",
                       "readr",
                       "remotes",
                       "rrtools",
                       "sf",
                       "sp",
                       "tidyr",
                       "usethis")

not_installed <- required_packages[!(required_packages %in% installed.packages()[ , "Package"])]
if(length(not_installed)) install.packages(not_installed)


#load pipe operator
library(magrittr)

######### Apply taxonomic flagging per region
#get vector of regions (here the survey column)
regions <- levels(as.factor(clean_GSLsouth$survey))

#run flag_spp function in a loop
for (r in regions) {
  flag_spp(clean_GSLsouth, r)
}

######### Apply trimming per survey_unit method 1
#apply trimming for hex size 7
dat_new_method1_hex7 <- apply_trimming_per_survey_unit_method1(clean_GSLsouth, 7)

#apply trimming for hex size 8
dat_new_method1_hex8 <- apply_trimming_per_survey_unit_method1(clean_GSLsouth, 8)

######### Apply trimming per survey_unit method 2
dat_new_method2 <- apply_trimming_per_survey_unit_method2(clean_GSLsouth)


#-----------------------------------------------------------------------------------------#
#### ADD STRANDARDIZATION FLAGS ####
#-----------------------------------------------------------------------------------------#
surveys <- sort(unique(clean_GSLsouth$survey))
```

```r
survey_units <- sort(unique(clean_GSLsouth$survey_unit))
survey_std <- clean_GSLsouth %>%
  mutate(flag_taxa = NA_character_,
         flag_trimming_hex7_0 = NA_character_,
         flag_trimming_hex7_2 = NA_character_,
         flag_trimming_hex8_0 = NA_character_,
         flag_trimming_hex8_2 = NA_character_,
         flag_trimming_2 = NA_character_)

# integrate taxonomic flags
for(i in 1:length(surveys)){
  if(!surveys[i] %in% c("FALK","GSL-N","MRT","NZ-CHAT","SCS", "SWC-IBTS")){
    xx <- data.frame(read_delim(paste0("outputs/Flags/taxonomic_flagging/",
                                       surveys[i],"_flagspp.txt"),
                                delim=";", escape_double = FALSE, col_names = FALSE,
                                trim_ws = TRUE))
    xx <- as.vector(unlist(xx[1,]))

    survey_std <- survey_std %>%
      mutate(flag_taxa = ifelse(survey == surveys[i] & accepted_name %in% xx,
                                "TRUE",flag_taxa))

    rm(xx)
  }
}

# integrate spatio-temporal flags
for(i in 1:length(survey_units)){

  if(!survey_units[i] %in% c("DFO-SOG","IS-TAU","SCS-FALL","WBLS")){

    hex_res7_0 <- read.csv(paste0("outputs/Flags/trimming_method1/hex_res7/",
                                  survey_units[i], "_hex_res_7_trimming_0_hauls_removed.csv"),
                           sep = ";")
    hex_res7_0 <- as.vector(hex_res7_0[,1])

    hex_res7_2 <- read.csv(paste0("outputs/Flags/trimming_method1/hex_res7/",
                                  survey_units[i], "_hex_res_7_trimming_02_hauls_removed.csv"),
                           sep = ";")
    hex_res7_2 <- as.vector(hex_res7_2[,1])

    hex_res8_0 <- read.csv(paste0("outputs/Flags/trimming_method1/hex_res8/",
                                  survey_units[i], "_hex_res_8_trimming_0_hauls_removed.csv"),
                           sep= ";")
    hex_res8_0 <- as.vector(hex_res8_0[,1])

    hex_res8_2 <- read.csv(paste0("outputs/Flags/trimming_method1/hex_res8/",
                                  survey_units[i], "_hex_res_8_trimming_02_hauls_removed.csv"),
                           sep = ";")
    hex_res8_2 <- as.vector(hex_res8_2[,1])

    trim_2 <- read.csv(paste0("outputs/Flags/trimming_method2/",
                              survey_units[i],"_hauls_removed.csv"))
```

```
    trim_2 <- as.vector(trim_2[,1])

    survey_std <- survey_std %>%
      mutate(flag_trimming_hex7_0 = ifelse(survey_unit == survey_units[i] & haul_id %in% hex_res7_0,
                                           "TRUE",flag_trimming_hex7_0),
             flag_trimming_hex7_2 = ifelse(survey_unit == survey_units[i] & haul_id %in% hex_res7_2,
                                           "TRUE",flag_trimming_hex7_2),
             flag_trimming_hex8_0 = ifelse(survey_unit == survey_units[i] & haul_id %in% hex_res8_0,
                                           "TRUE",flag_trimming_hex8_0),
             flag_trimming_hex8_2 = ifelse(survey_unit == survey_units[i] & haul_id %in% hex_res8_2,
                                           "TRUE",flag_trimming_hex8_2),
             flag_trimming_2 = ifelse(survey_unit == survey_units[i] & haul_id %in% trim_2,
                                      "TRUE", flag_trimming_2)
      )
    rm(hex_res7_0, hex_res7_2, hex_res8_0, hex_res8_2, trim_2)
  }
}


# Just run this routine should be good for all
write_clean_data(data = survey_std, survey = "GSL-S_std",
                 overwrite = T, rdata=TRUE)
```

## 1. Overview of the survey data table

| survey | source | timestamp | haul_id | country | sub_area |
|---|---|---|---|---|---|
| GSL-S | DFO | 2021 | 1970-9-15-12-25–64.7166666666667-47.9833333333333 | Canada | NA |
| GSL-S | DFO | 2021 | 1970-9-15-12-25–64.7166666666667-47.9833333333333 | Canada | NA |
| GSL-S | DFO | 2021 | 1970-9-15-12-25–64.7166666666667-47.9833333333333 | Canada | NA |
| GSL-S | DFO | 2021 | 1970-9-15-12-25–64.7166666666667-47.9833333333333 | Canada | NA |
| GSL-S | DFO | 2021 | 1970-9-15-12-25–64.7166666666667-47.9833333333333 | Canada | NA |

| continent | stat_rec | station | stratum | year | month | day | quarter | season |
|---|---|---|---|---|---|---|---|---|
| n_america | NA | NA | NA | 1970 | 9 | 15 | 3 | NA |
| n_america | NA | NA | NA | 1970 | 9 | 15 | 3 | NA |
| n_america | NA | NA | NA | 1970 | 9 | 15 | 3 | NA |
| n_america | NA | NA | NA | 1970 | 9 | 15 | 3 | NA |
| n_america | NA | NA | NA | 1970 | 9 | 15 | 3 | NA |

| latitude | longitude | haul_dur | area_swept | gear | depth | sbt | sst |
|---|---|---|---|---|---|---|---|
| 47.98333 | -64.71667 | 0.5 | 0.035 | Yankee #36 otter trawl | NA | NA | NA |
| 47.98333 | -64.71667 | 0.5 | 0.035 | Yankee #36 otter trawl | NA | NA | NA |
| 47.98333 | -64.71667 | 0.5 | 0.035 | Yankee #36 otter trawl | NA | NA | NA |
| 47.98333 | -64.71667 | 0.5 | 0.035 | Yankee #36 otter trawl | NA | NA | NA |
| 47.98333 | -64.71667 | 0.5 | 0.035 | Yankee #36 otter trawl | NA | NA | NA |

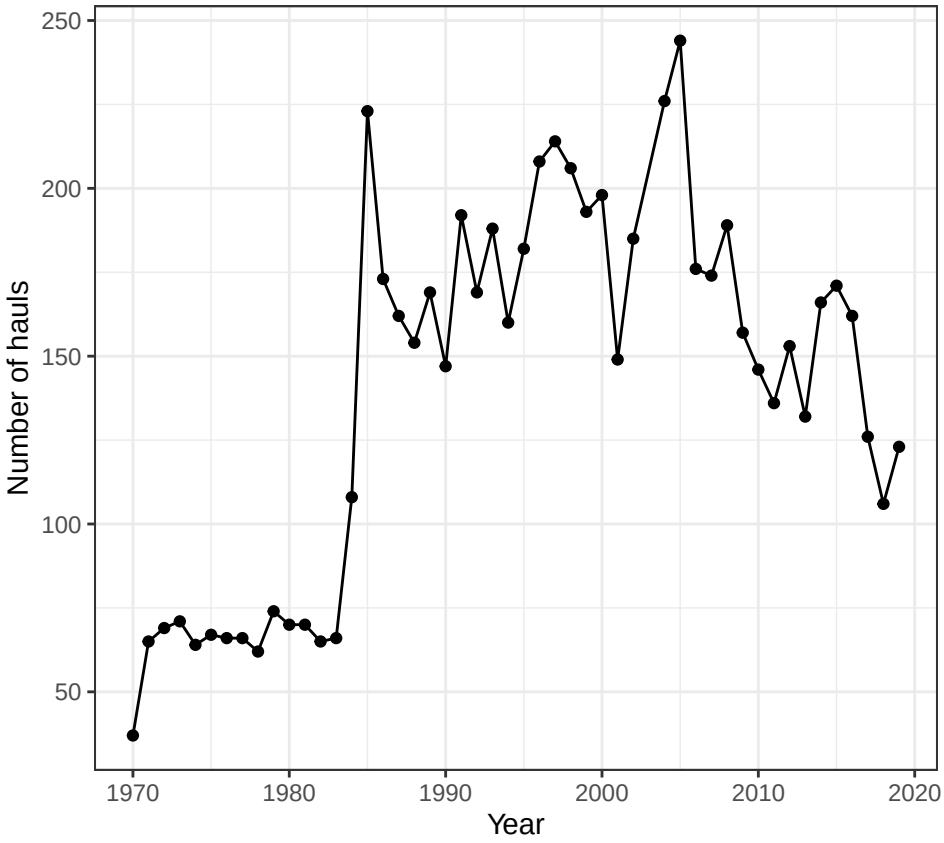
| num | num_cpue | num_cpua | wgt | wgt_cpue | wgt_cpua | verbatim_name |
|---|---|---|---|---|---|---|
| 60.328947 | 120.65789 | 1723.6842 | 80.840789 | 161.681579 | 2309.7368 | GADUS MORHUA |
| 316.363158 | 632.72632 | 9038.9474 | 60.531579 | 121.063158 | 1729.4737 | HIPPOGLOSSOIDES PLATESSOIDES |
| 4.282895 | 8.56579 | 122.3684 | 4.282895 | 8.565789 | 122.3684 | GLYPTOCEPHALUS CYNOGLOSSUS |
| 10.948620 | 21.89724 | 312.8177 | 3.515468 | 7.030936 | 100.4419 | LIMANDA FERRUGINEA |
| 59.989612 | 119.97922 | 1713.9889 | 15.148892 | 30.297784 | 432.8255 | CLUPEA HARENGUS |

| verbatim_aphia_id | accepted_name | aphia_id | SpecCode | kingdom |
|---|---|---|---|---|
| NA | Gadus morhua | 126436 | 69 | Animalia |
| NA | Hippoglossoides platessoides | 127137 | 4239 | Animalia |
| NA | Glyptocephalus cynoglossus | 127136 | 26 | Animalia |
| NA | Limanda ferruginea | 158879 | 521 | Animalia |
| NA | Clupea harengus | 126417 | 24 | Animalia |

| phylum | class | order | family | genus | rank | survey_unit |
|---|---|---|---|---|---|---|
| Chordata | Teleostei | Gadiformes | Gadidae | Gadus | Species | GSL-S |
| Chordata | Teleostei | Pleuronectiformes | Pleuronectidae | Hippoglossoides | Species | GSL-S |
| Chordata | Teleostei | Pleuronectiformes | Pleuronectidae | Glyptocephalus | Species | GSL-S |
| Chordata | Teleostei | Pleuronectiformes | Pleuronectidae | Limanda | Species | GSL-S |
| Chordata | Teleostei | Clupeiformes | Clupeidae | Clupea | Species | GSL-S |

## 2. Summary of sampling intensity

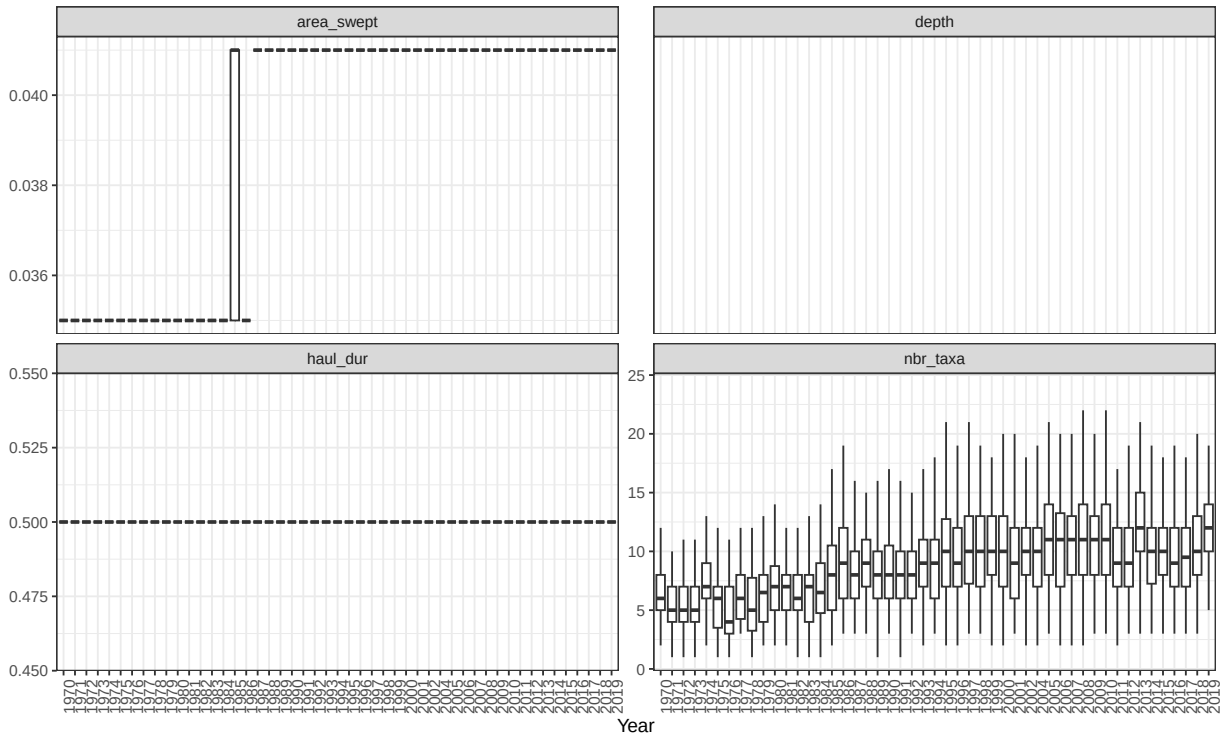
Number of hauls per year performed during the survey after data processing.



## 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:

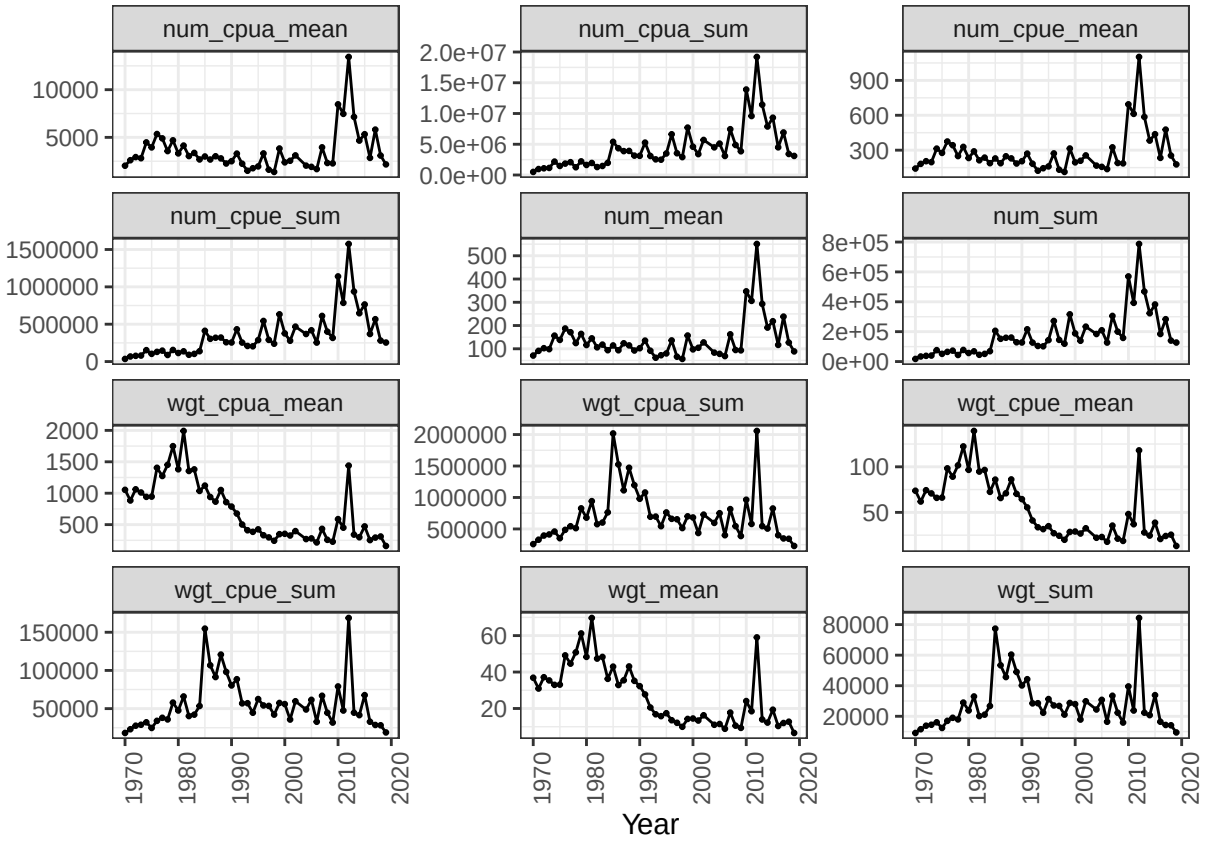
- *area_swept*, swept area by the bottom trawl gear $km^2$
- *depth*, sampling depth in $m$
- *haul_dur*, haul sampling duration *hour*
- *number of marine fish taxa*, taxa were cleaned following the last version of taxonomy from the World Register of Marine Species (https://www.marinespecies.org/, October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



## 4. Summary of biological variables

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

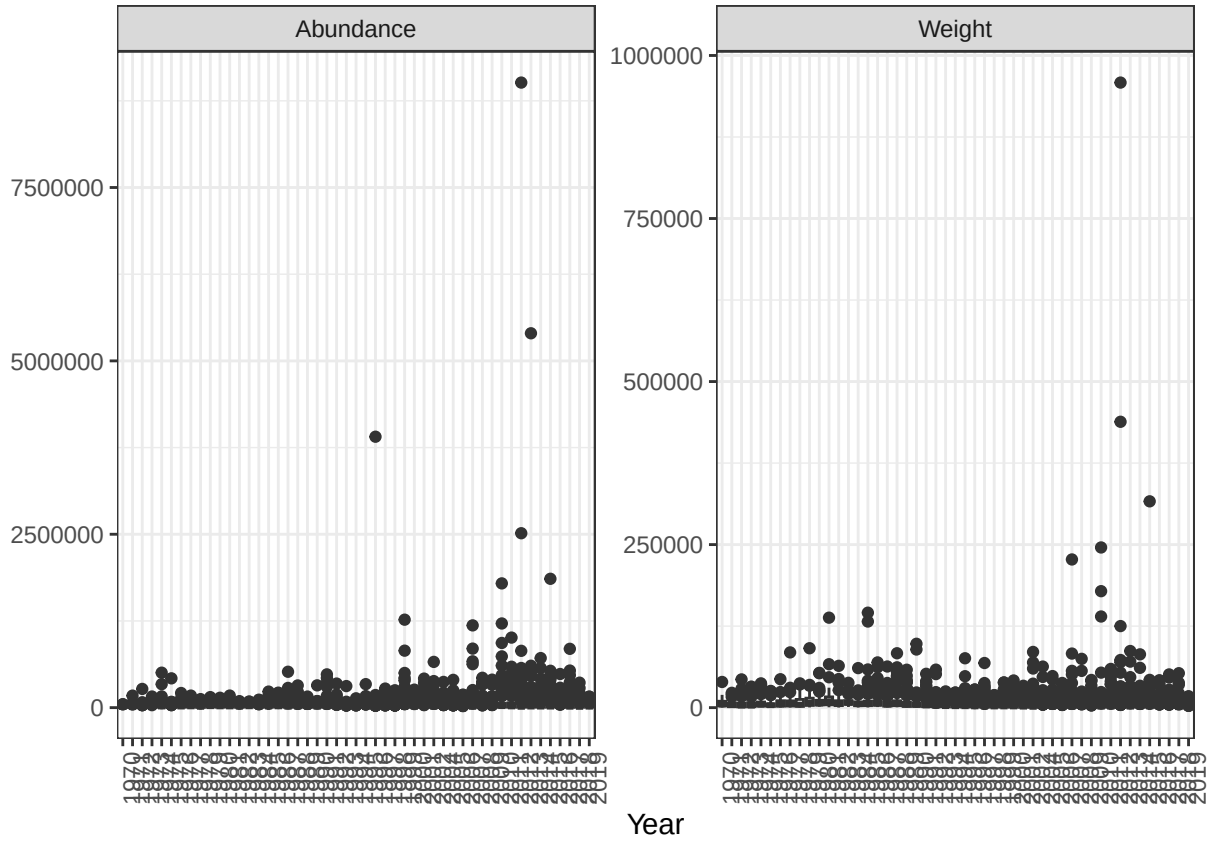
- *num_cpua*, number of individuals (abundance) in $\frac{individuals}{km^2}$
- *num_cpue*, number of individuals (abundance) in $\frac{individuals}{h}$
- *num*, number of individuals (abundance)
- *wgt_cpua*, weight in $\frac{kg}{km^2}$
- *wgt_cpue*, weight in $\frac{kg}{h}$
- *wgt*, weight in $kg$



## 5. Extreme values

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

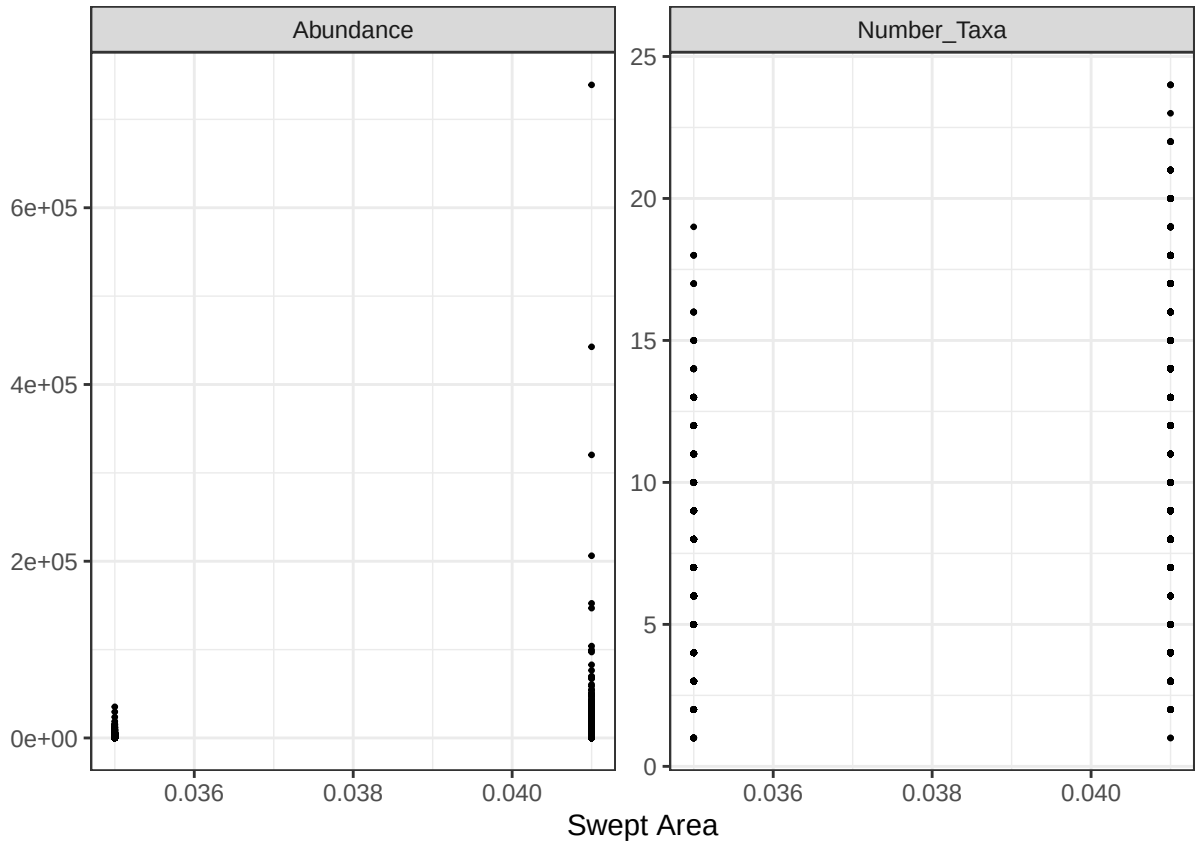
- *num_cpue*, number of individuals (abundance) in $\frac{individuals}{km^2}$
- *wgt_cpue*, weight in $\frac{kg}{km^2}$



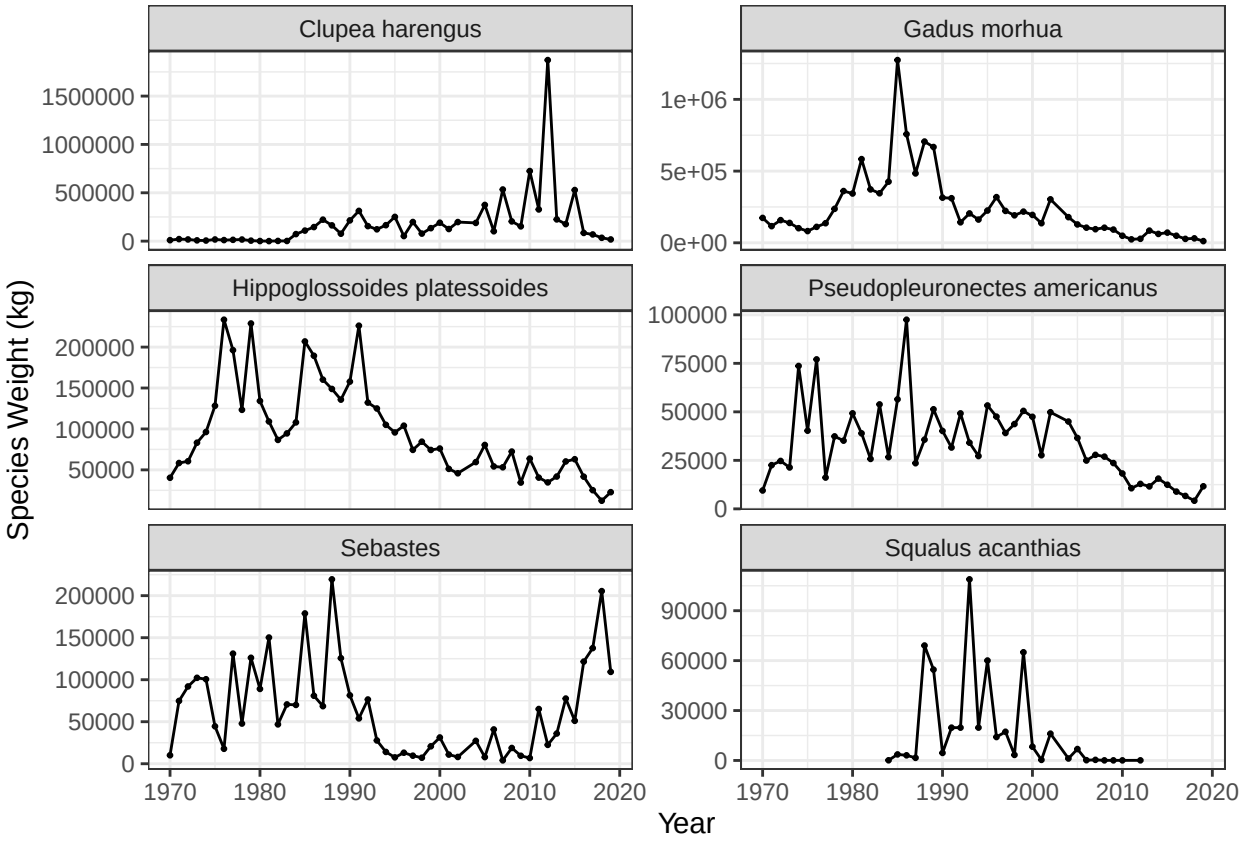
## 6. Summary of variables against swept area

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_cpua*, number of individuals (abundance) in $\frac{individuals}{km^2}$
- *wgt_cpua*, weight in $\frac{kg}{km^2}$

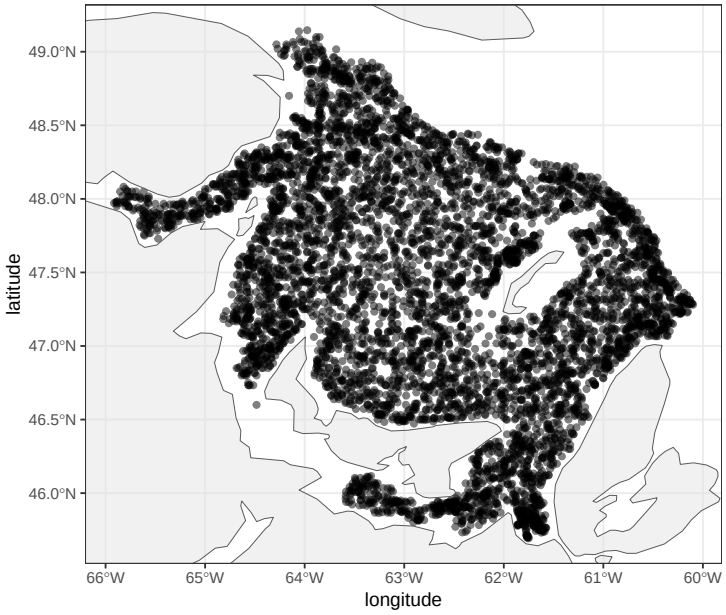


## 7. Abundance or Weight trends of the six most abundant species



## 8. Distribution mapping

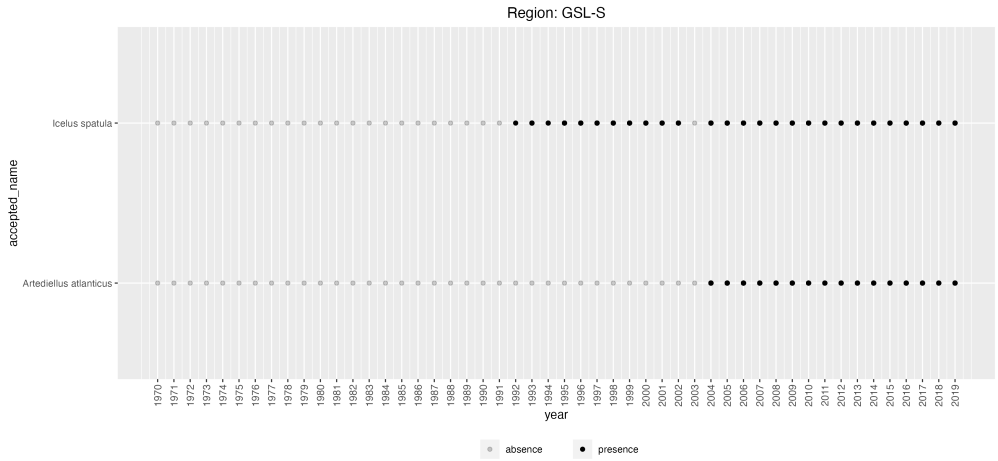
Map of the sampling distribution in space. Note that we only show one year per coordinate.



## 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

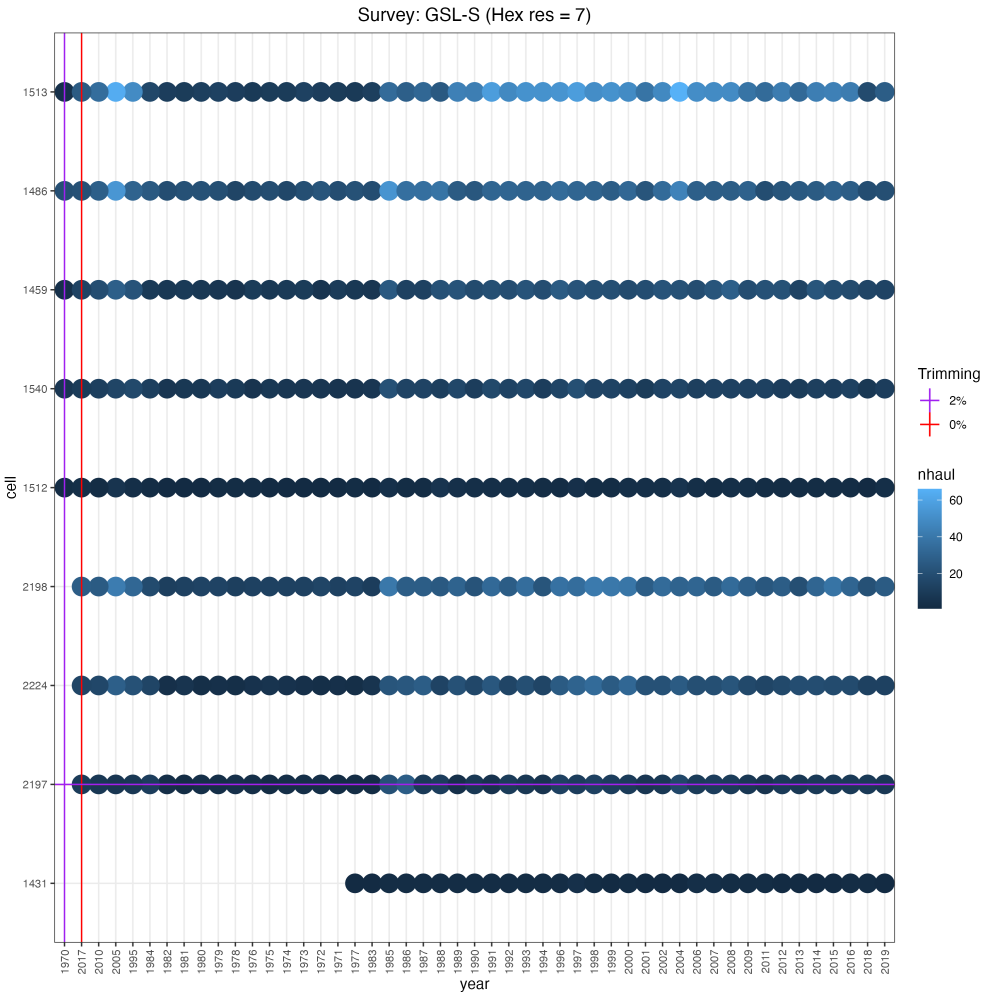
| | |
|---|---|
| Total number of species | 134.0 |
| Percentage of species flagged | 1.5 |

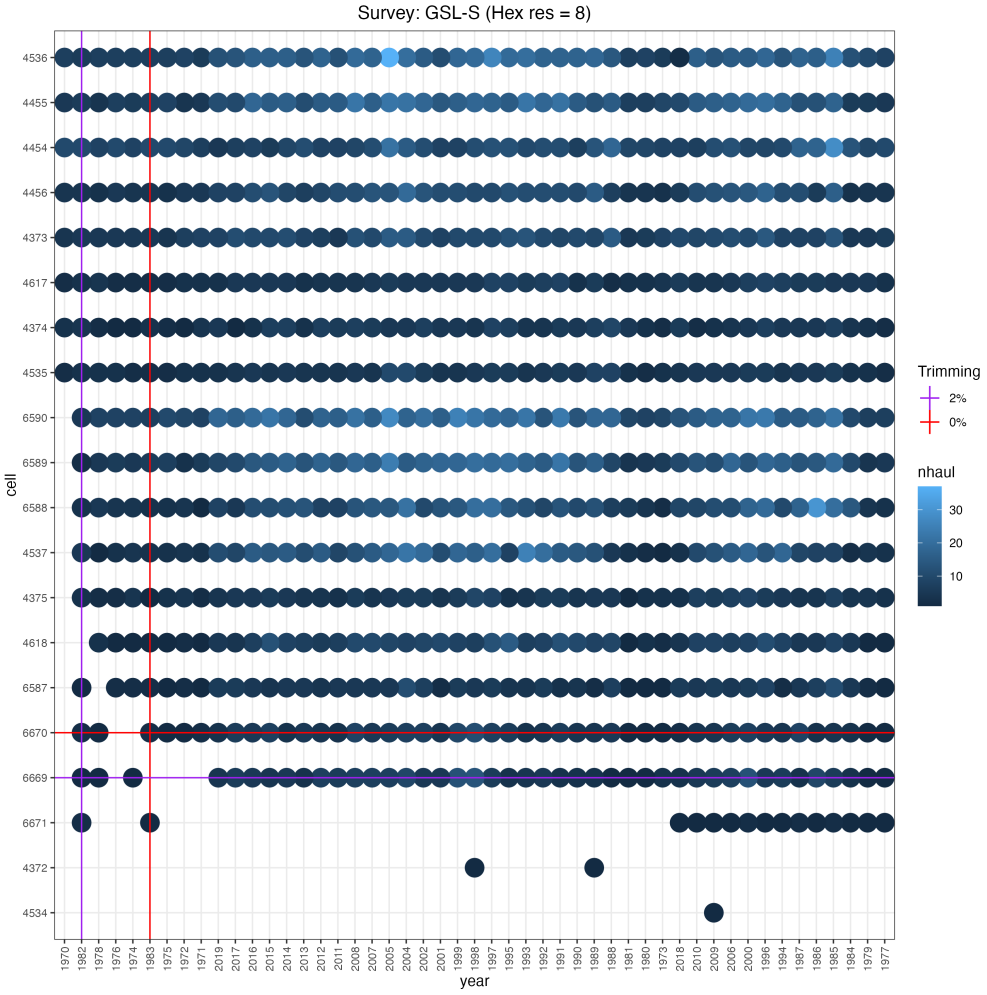
## 10. Spatio-temporal standardization

### a. Standardization method 1

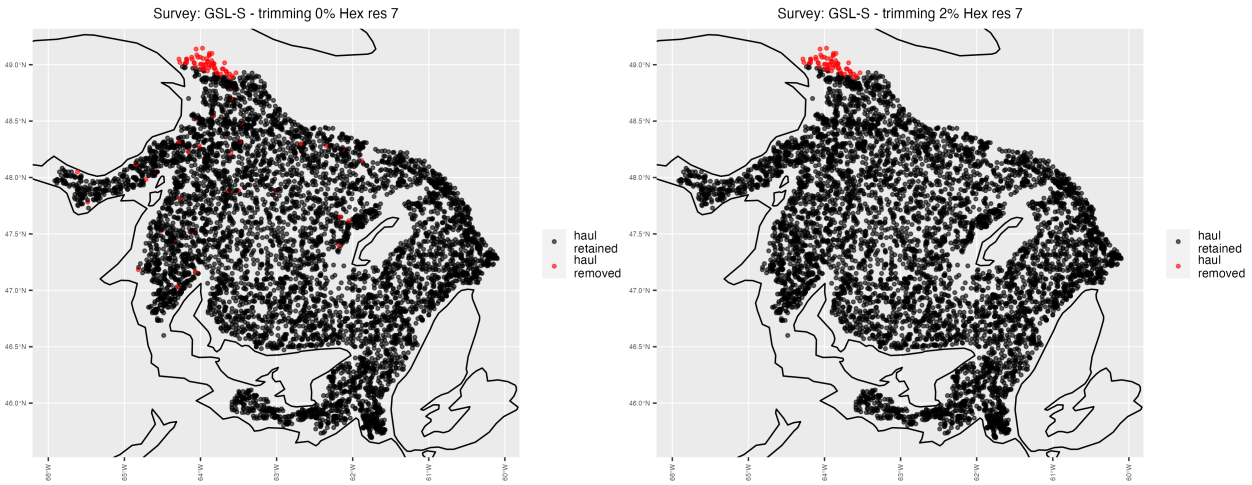
This standardization method was adapted from https://github.com/zoekitchel/trawl_spatial_turnover/blob/master/data_prep_code/species/explore_NorthSea_trimming.Rmd
It was run for hex resolution 7 and 8.

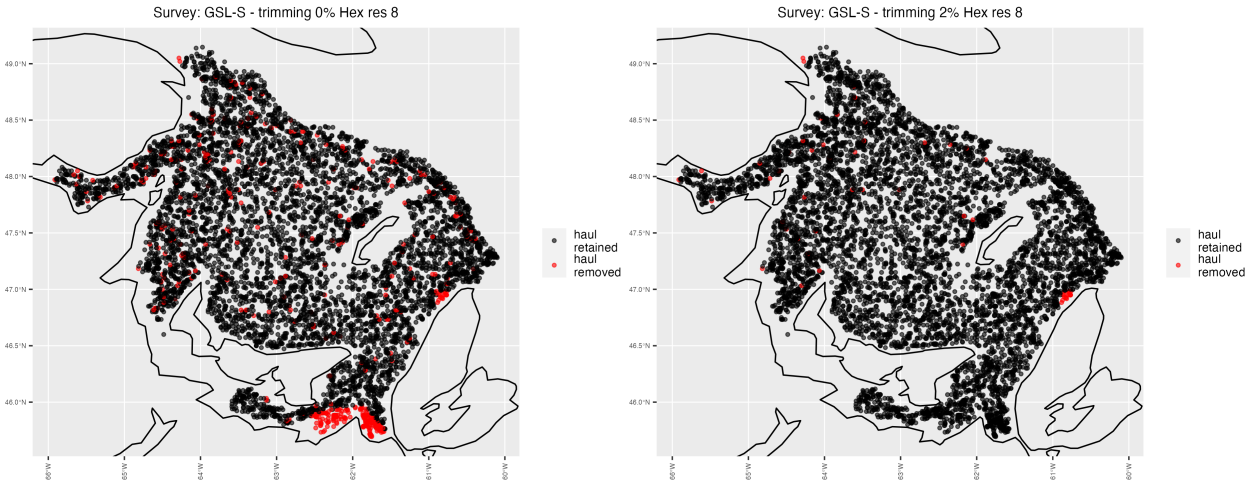
Plot of number of cells x years with overlaid flagging options



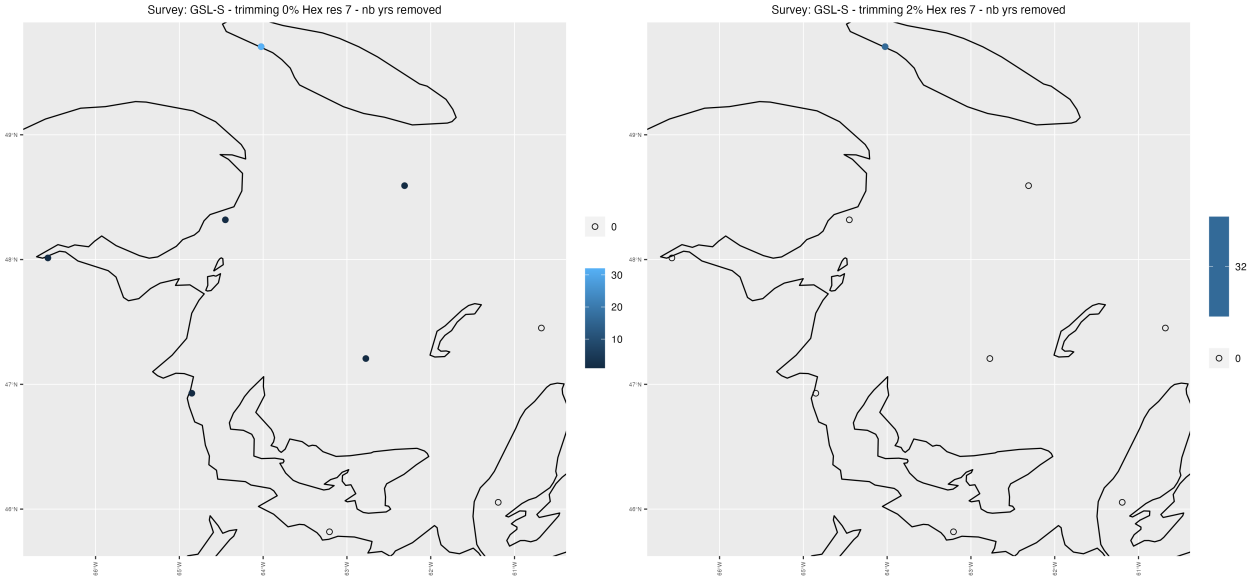


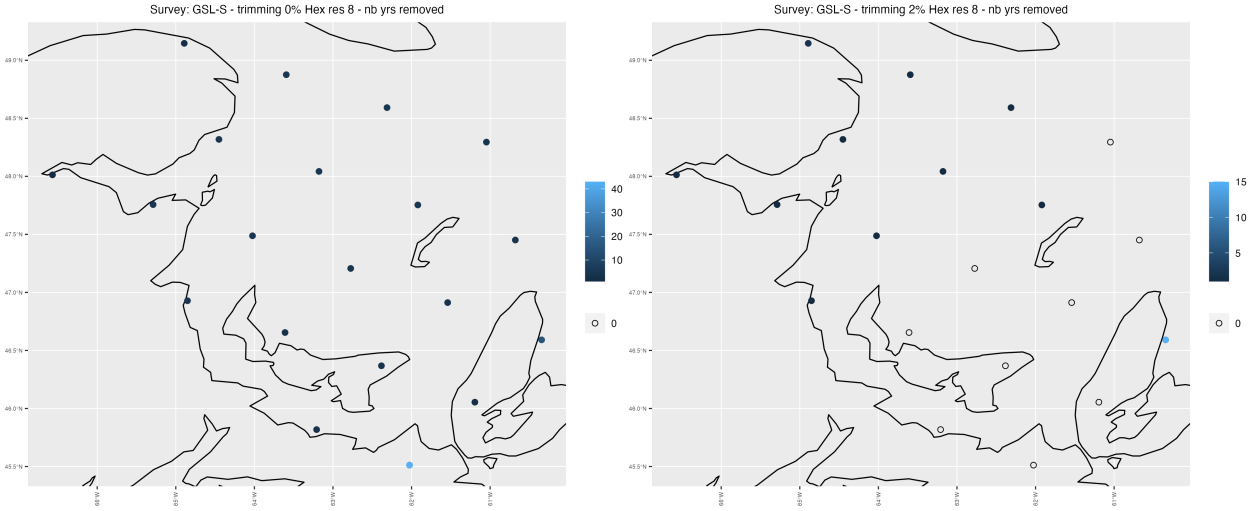

Map of hauls retained and removed per flagging method and threshold






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

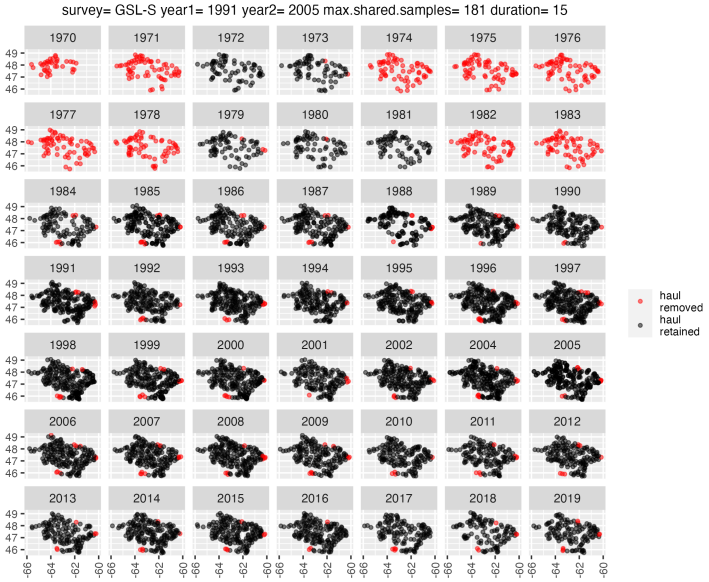





## b. Standardization method 2

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 | 94.0 | 57.0 | 519.0 | 59.0 | 5973.0 |
| percentage of hauls removed | 1.4 | 0.8 | 7.5 | 0.9 | 9.4 |