GSL-N: Gulf of St. Lawrence North survey data processing summary

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

January, 2023

Contents

General info

This document presents the cleaning code and summary of the Gulf of St. Lawrence North (Canada) bottom trawl survey provided by Fisheries and Oceans Canada. It contains data from 1980 and up to 2019.

Data cleaning in R

```
##############################################################################
#### R code to clean trawl survey for Gulf of St. Lawrence North
#### 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 November 2021
##############################################################################
#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")

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

#Note that there have been gear changes and required calibrations
#for GSL-N
#and described well in here:
#Bourdages, H., Brassard, C., Desgagnés, M., Galbraith, P., Gauthier, J., Lambert, J., Légaré,
#B., Parent, E. and Schwab P. 2015. Preliminary results from the groundfish and shrimp
#multidisciplinary survey in August 2014 in the Estuary and northern Gulf of St. Lawrence.
#DFO Can. Sci. Advis. Sec. Res. Doc. 2014/115. v + 96 p.
#The analysis of 2014 abundance and biomass data were integrated into the combined
#annual summer survey series initiated in 1990. This combined series was developed
#following a comparative study between the two vessel-gear tandems (1990-2005: CCGS
#Alfred Needler - URI 81'/114' trawl; 2004-2012: CCGS Teleost - Campelen 1800 trawl) to
#establish specific correction factors for about twenty species caught (Bourdages et al.
#2007). This resulted in adjustment of Needler catches into Teleost equivalent catches.
#Note that the distinction between the two redfish species, Sebastes fasciatus and S.
#mentella, is based on the analysis of the soft anal fin rays count and the depth of capture
#of individuals (H. Bourdages, DFO Mont-Joli, pers. comm.).

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

#GSL North Sentinel

GSLnor_sent <- read.csv(
  "https://github.com/pinskylab/OceanAdapt/raw/master/data_raw/GSLnorth_sentinel.csv")

#GSL North Gadus

GSLnor_gad <- read.csv(
  "https://github.com/pinskylab/OceanAdapt/raw/master/data_raw/GSLnorth_gadus.csv")

#GSL North Hammond

GSLnor_ham <- read.csv(
  "https://github.com/pinskylab/OceanAdapt/raw/master/data_raw/GSLnorth_hammond.csv")
```

```r
#GSL North Needler

GSLnor_need <- read.csv(
  "https://github.com/pinskylab/OceanAdapt/raw/master/data_raw/GSLnorth_needler.csv")

#GSL North Teleost

GSLnor_tel <- read.csv(
  "https://github.com/pinskylab/OceanAdapt/raw/master/data_raw/GSLnorth_teleost.csv")

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

#Bind all datasets

GSLnor <- plyr::rbind.fill(GSLnor_sent, GSLnor_gad, GSLnor_ham, GSLnor_need, GSLnor_tel)
GSLnor$lat <-as.numeric(as.character(GSLnor$Latit_Deb))
GSLnor$lon <-as.numeric(as.character(GSLnor$Longit_Deb))
GSLnor$depth <-as.numeric(as.character(GSLnor$Prof_Max))
GSLnor$Dist_Chalute_Position <-as.numeric(as.character(GSLnor$Dist_Chalute_Position))
GSLnor$Pds_Capture <- as.double(GSLnor$Pds_Capture)
GSLnor$Nb_Ind_Capture <- as.numeric(as.character(GSLnor$Nb_Ind_Capture))
GSLnor$Date <-as.Date(GSLnor$Date_Deb_Trait)
GSLnor$year <- as.integer(year(GSLnor$Date))
GSLnor$verbatim_name <- trimws(as.character(GSLnor$Nom_Scient_Esp), which = "right")



GSLnor <- GSLnor[!is.na(GSLnor$lat),] #only keep rows with latitude
GSLnor <- GSLnor[!is.na(GSLnor$depth),] #only keep rows with depth

GSLnor <- GSLnor %>%
  # Create a unique haul_id
  mutate(
    haul_id = paste(GSLnor$Nom_Navire, GSLnor$No_Releve,GSLnor$Trait,
                    GSLnor$Date_Deb_Trait,GSLnor$Hre_Deb, sep="-"),
    #area in km^2 =
#Dist_Chalute_Position (nautical miles) * 1852 m/1 nautical mile *
#                                                trawl width *(1km^2/1000000m^2)
    area_swept = Dist_Chalute_Position * 1852 * 12.497 *(1/1000000),
    wgt = Pds_Capture, #in kg
    num = Nb_Ind_Capture, #in pieces
    # (via Daniel Ricard) trawl width, 12.497 m. Hurlbut and Clay (1990)
    # catch weight (kg.) per tow    /km^2,
    wgt_cpue = (Pds_Capture)/area_swept,
    #weight in kg/time in minutes*60minutes/1hour
    wgt_h = (Pds_Capture)/Duree*60,
    #abundance in number/km^2
    num_cpue = Nb_Ind_Capture/area_swept,
    #abundance in number/hour
    num_h = Nb_Ind_Capture/Duree*60,
  )
```

```r
GSLnor <- GSLnor %>%
  filter(
    # remove unidentified spp and non-species
    verbatim_name != "" | !is.na(verbatim_name),
    !grepl("EGG", verbatim_name),
    !grepl("UNIDENTIFIED", verbatim_name)) %>%
  # add survey column
  mutate(survey = "GSL-N")

#check that the number of unique haul_ids *
#                      spp combinations is the same as the number of rows in mar
nrow(GSLnor) == nrow(unique(GSLnor[,c("haul_id","verbatim_name")]))

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

GSLnor <- GSLnor %>%
  # Adding extra columns and setting proper format
  mutate(
    country = "Canada",
    sub_area = NA,
    continent = "n_america",
    stat_rec = NA,
    station = NA,
    stratum = NA,
    season = NA,
    latitude = lat,
    longitude = lon,
    month = month(Date),
    day = day(Date),
    haul_dur = ifelse(Duree > 0, Duree/60, NA),
    #get rid of negative duration values and code them as NA
    gear = Engin,
    sbt = NA,
    sst = NA,
    quarter = ifelse(month %in% c(1,2,3),1,
                     ifelse(month %in% c(4,5,6),2,
                            ifelse(month %in% c(7,8,9),3,
                                   4
                            )
                     )
    ),
    aphia_id = NA,
    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 ####
#-------------------------------------------------------------------------------#
```

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

### Automatic cleaning
# Set Survey code
GSLnor_survey_code <- "GSL-N"

GSLnor <- GSLnor %>%
  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(GSLnor$taxa2),
                         input_survey = GSLnor_survey_code, save = F, output=NA,
                         fishbase=T)

#This leaves out the following species, all of which are inverts
#Eualus gaimardii belcheri (invert)

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

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

clean_GSLnor <- left_join(GSLnor, 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`)
```

```r
# ------------------------------------------------------------------------------------------#
#### SAVE DATABASE IN GOOGLE DRIVE ####
# ------------------------------------------------------------------------------------------#

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


# ------------------------------------------------------------------------------------------#
#### FAGS ####
# ------------------------------------------------------------------------------------------#
#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_GSLnor$survey))

#run flag_spp function in a loop
for (r in regions) {
  flag_spp(clean_GSLnor, 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_GSLnor, 7)
```

```r
#apply trimming for hex size 8
dat_new_method1_hex8 <- apply_trimming_per_survey_unit_method1(clean_GSLnor, 8)

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


#-----------------------------------------------------------------------------------------------#
#### ADD STRANDARDIZATION FLAGS ####
#-----------------------------------------------------------------------------------------------#
surveys <- sort(unique(clean_GSLnor$survey))
survey_units <- sort(unique(clean_GSLnor$survey_unit))
survey_std <- clean_GSLnor %>%
  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-N_std",
                 overwrite = T, rdata=TRUE)
```

1. Overview of the survey data table

survey	source	timestamp	haul_id	country	sub_area
GSL-N	DFO	2021	Annie-Annick -3-1-1995-08-07 -05:40:00	Canada	NA
GSL-N	DFO	2021	Annie-Annick -3-1-1995-08-07 -05:40:00	Canada	NA
GSL-N	DFO	2021	Annie-Annick -3-1-1995-08-07 -05:40:00	Canada	NA
GSL-N	DFO	2021	Annie-Annick -3-2-1995-08-07 -07:45:00	Canada	NA
GSL-N	DFO	2021	Annie-Annick -3-2-1995-08-07 -07:45:00	Canada	NA

continent	stat_rec	station	stratum	year	month	day	quarter	season
n_america	NA	NA	NA	1995	8	7	3	NA
n_america	NA	NA	NA	1995	8	7	3	NA
n_america	NA	NA	NA	1995	8	7	3	NA
n_america	NA	NA	NA	1995	8	7	3	NA
n_america	NA	NA	NA	1995	8	7	3	NA

latitude	longitude	haul_dur	area_swept	gear	depth	sbt	sst
49.85500	-62.530	0.5	0.0284677	Chalut de fond	182	NA	NA
49.85500	-62.530	0.5	0.0284677	Chalut de fond	182	NA	NA
49.85500	-62.530	0.5	0.0284677	Chalut de fond	182	NA	NA
49.70167	-62.465	0.5	0.0261532	Chalut de fond	229	NA	NA
49.70167	-62.465	0.5	0.0261532	Chalut de fond	229	NA	NA

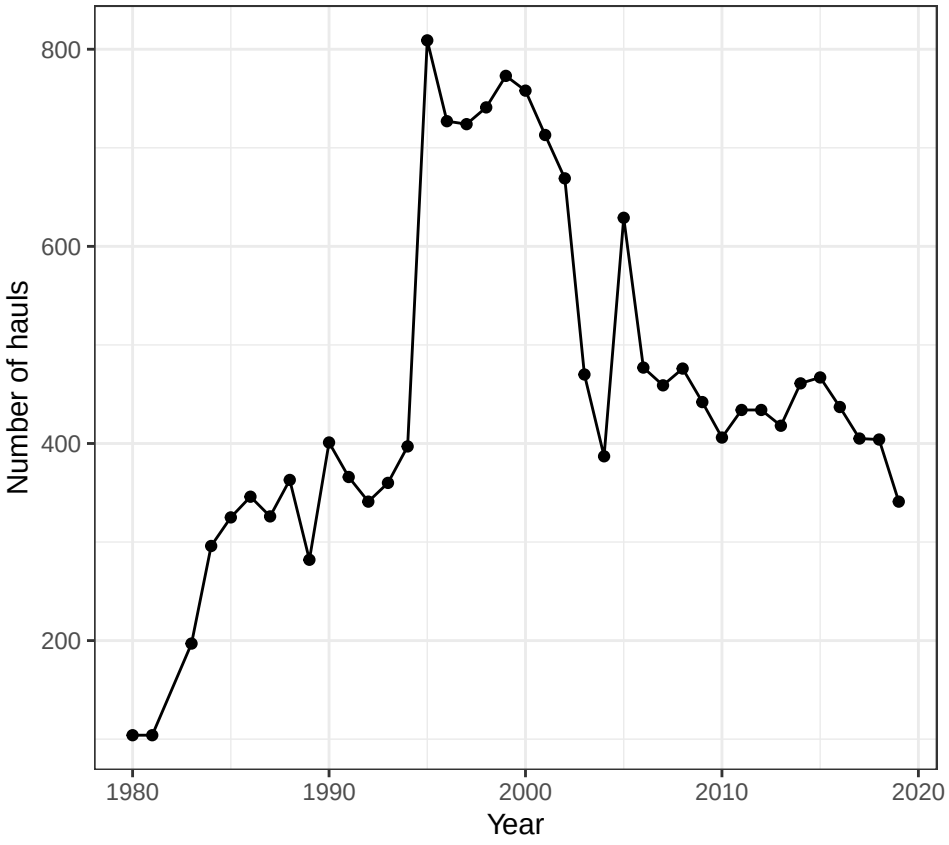
num	num_cpue	num_cpua	wgt	wgt_cpue	wgt_cpua	verbatim_name
16	32	562.04116	7.00	14.00	245.893006	Gadus morhua
2	4	70.25514	0.06	0.12	2.107654	Sebastes sp.
34	68	1194.33746	3.50	7.00	122.946503	Reinhardtius hippoglossoides
2	4	76.47241	0.11	0.22	4.205983	Sebastes sp.
17	34	650.01552	9.00	18.00	344.125863	Reinhardtius hippoglossoides

verbatim_aphia_id	accepted_name	aphia_id	SpecCode	kingdom
NA	Gadus morhua	126436	69	Animalia
NA	Sebastes	126175	NA	Animalia
NA	Reinhardtius hippoglossoides	127144	516	Animalia
NA	Sebastes	126175	NA	Animalia
NA	Reinhardtius hippoglossoides	127144	516	Animalia

phylum	class	order	family	genus	rank	survey_unit
Chordata	Teleostei	Gadiformes	Gadidae	Gadus	Species	GSL-N
Chordata	Teleostei	Perciformes	Sebastidae	Sebastes	Genus	GSL-N
Chordata	Teleostei	Pleuronectiformes	Pleuronectidae	Reinhardtius	Species	GSL-N
Chordata	Teleostei	Perciformes	Sebastidae	Sebastes	Genus	GSL-N
Chordata	Teleostei	Pleuronectiformes	Pleuronectidae	Reinhardtius	Species	GSL-N

2. Summary of sampling intensity

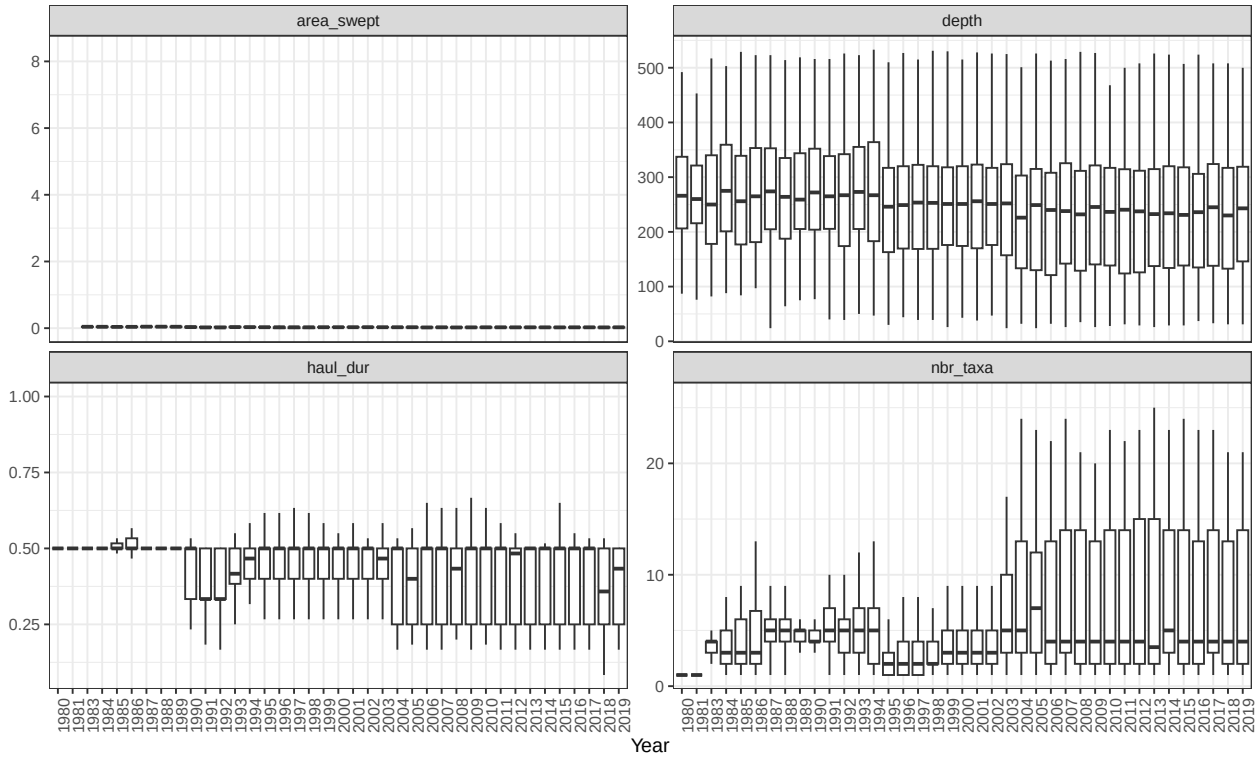
Number of hauls per year performed during the survey after data processing.



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:

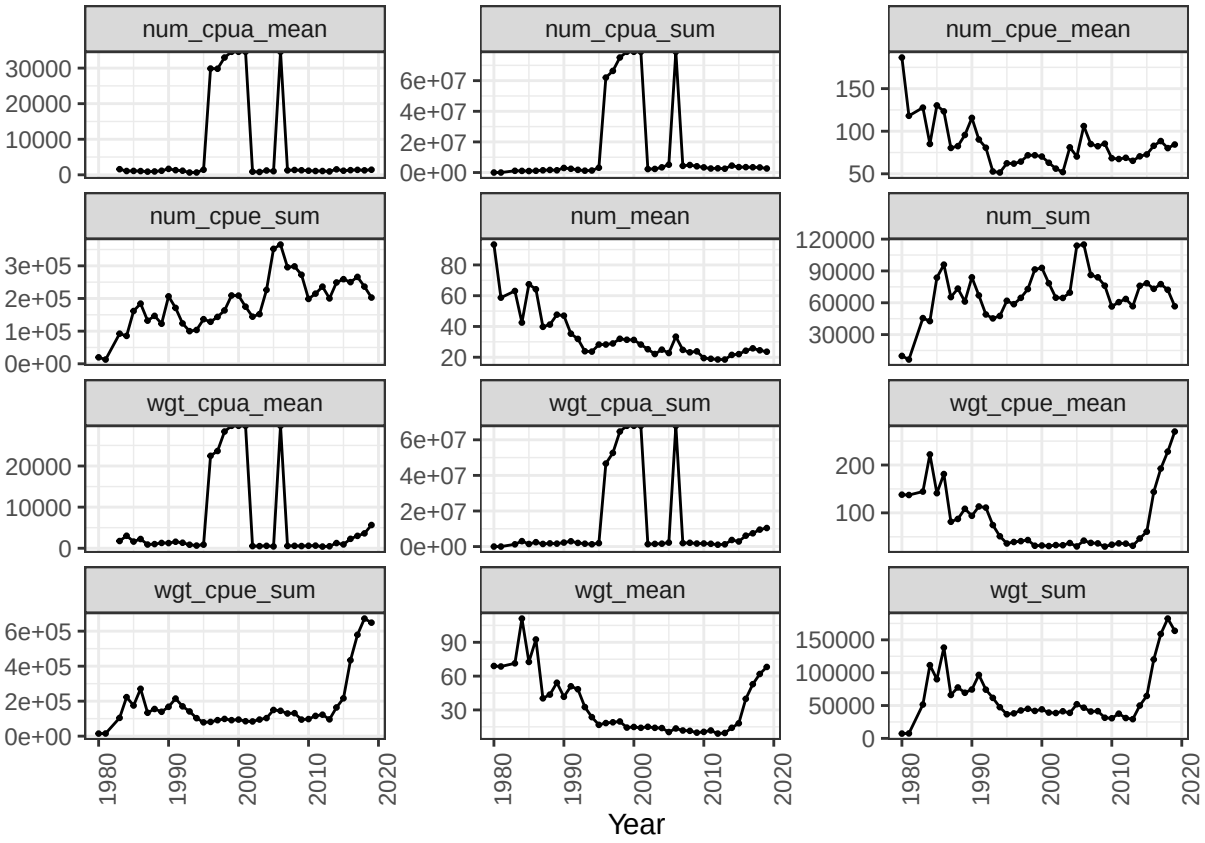
- *area_swept*, swept area by the bottom trawl gear km^2
- *depth*, sampling depth in m
- *haul_dur*, haul sampling duration *hour*
- *number of marine fish taxa*, taxa were cleaned following the last version of taxonomy from the World Register of Marine Species (https://www.marinespecies.org/, October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



4. Summary of biological variables

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

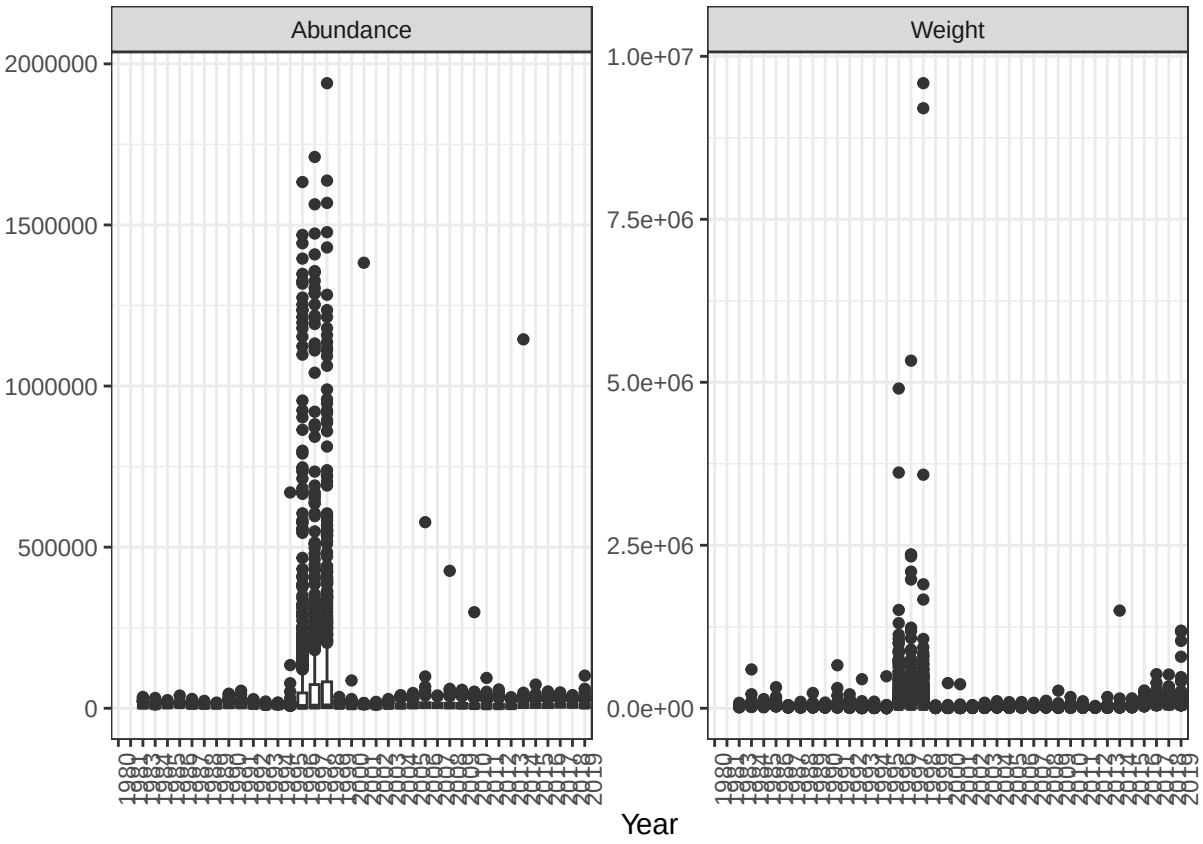
- *num_cpua*, number of individuals (abundance) in $\frac{individuals}{km^2}$
- *num_cpue*, number of individuals (abundance) in $\frac{individuals}{h}$
- *num*, number of individuals (abundance)
- *wgt_cpua*, weight in $\frac{kg}{km^2}$
- *wgt_cpue*, weight in $\frac{kg}{h}$
- *wgt*, weight in kg



5. Extreme values

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

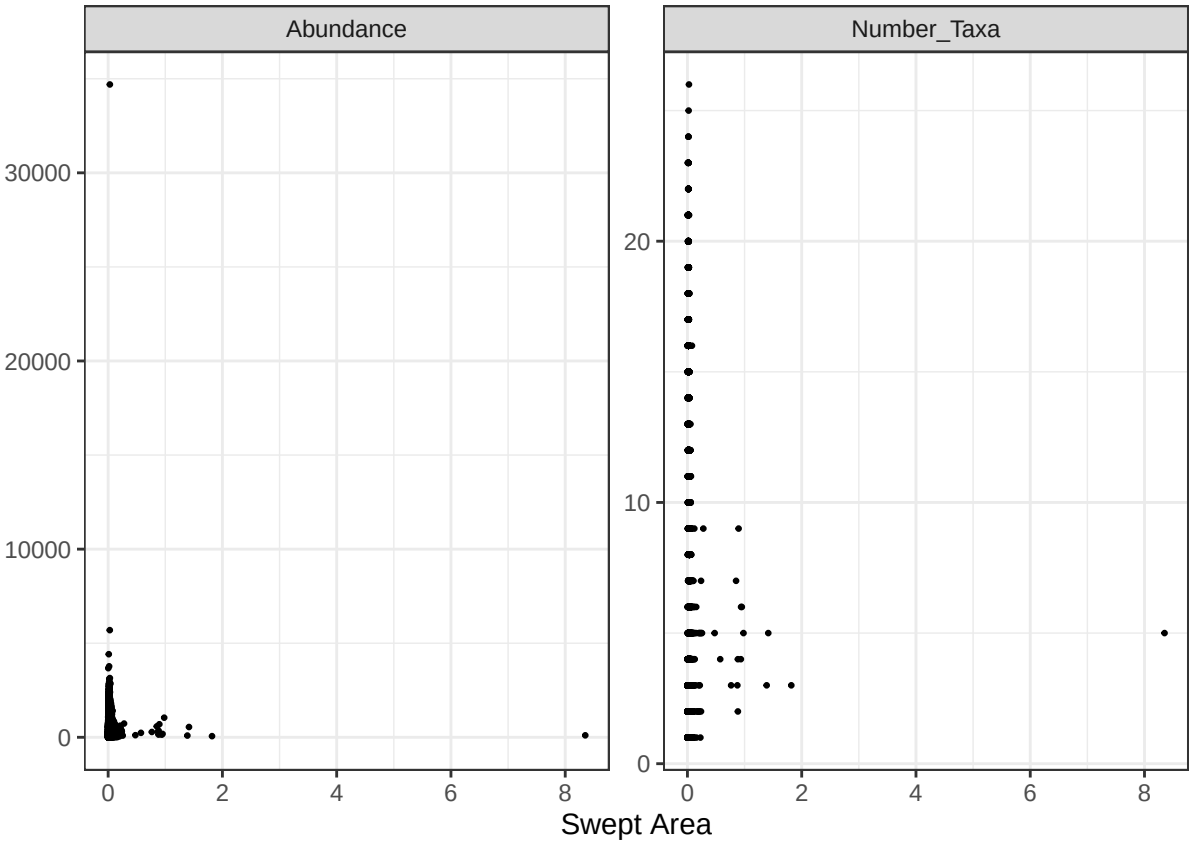
- *num_cpue*, number of individuals (abundance) in $\frac{individuals}{km^2}$
- *wgt_cpue*, weight in $\frac{kg}{km^2}$



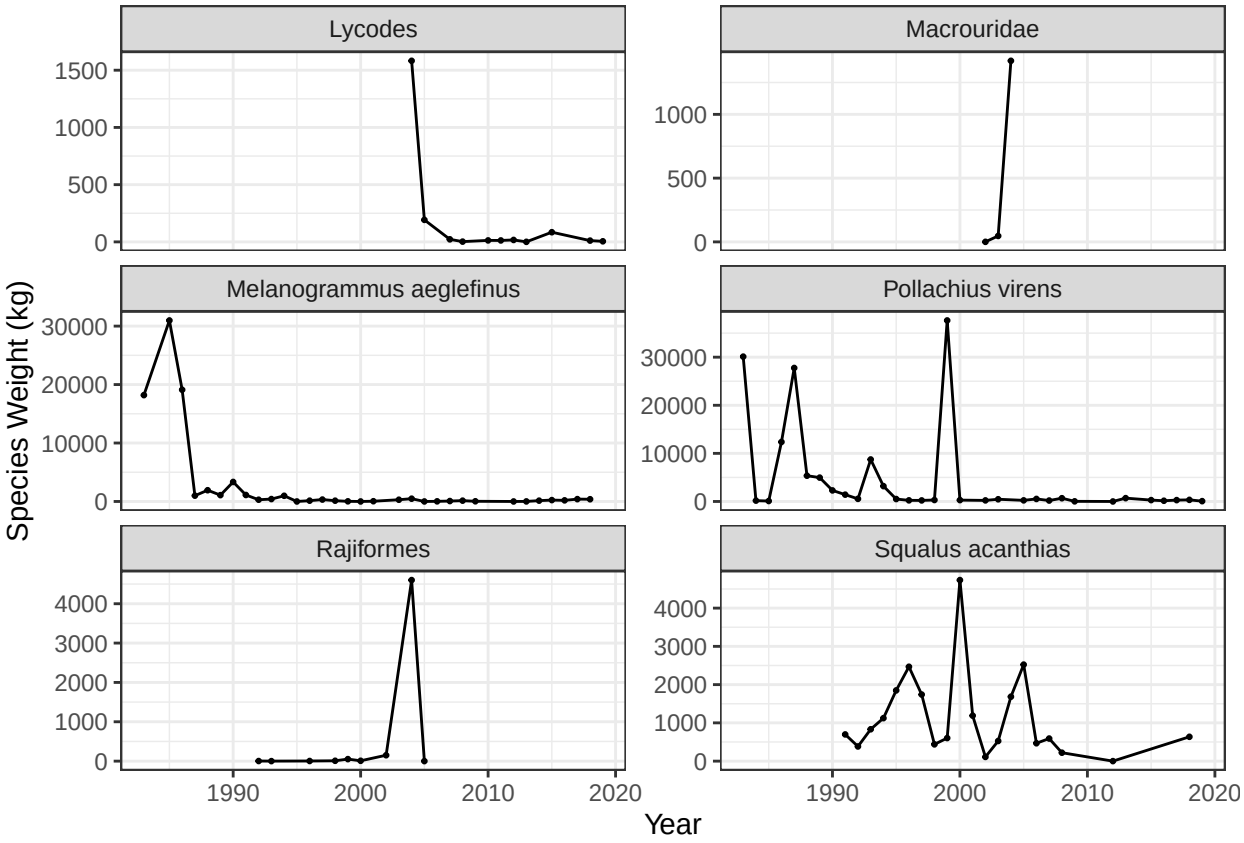
6. Summary of variables against swept area

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 $\frac{kg}{km^2}$

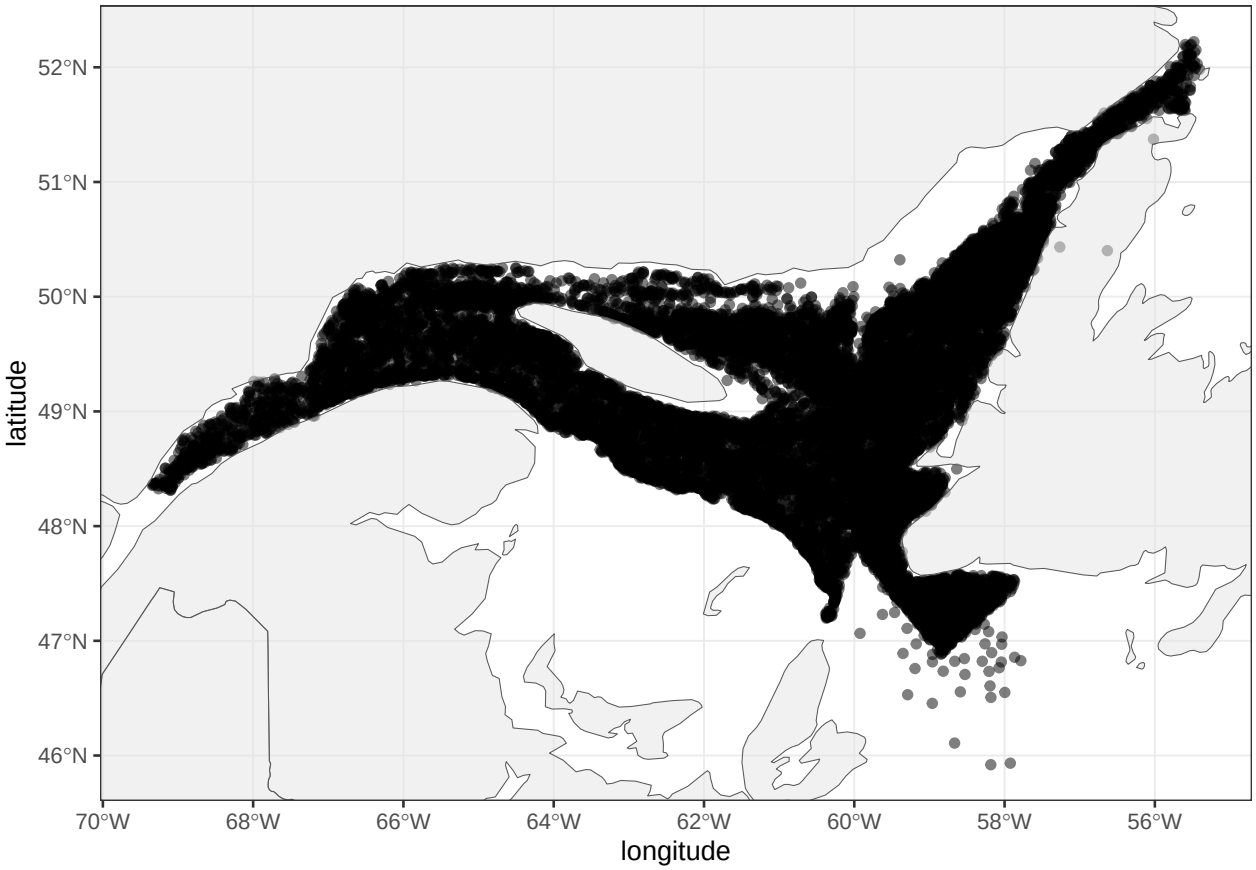


7. Abundance or Weight trends of the six most abundant species



8. Distribution mapping

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



9. Taxonomic flagging

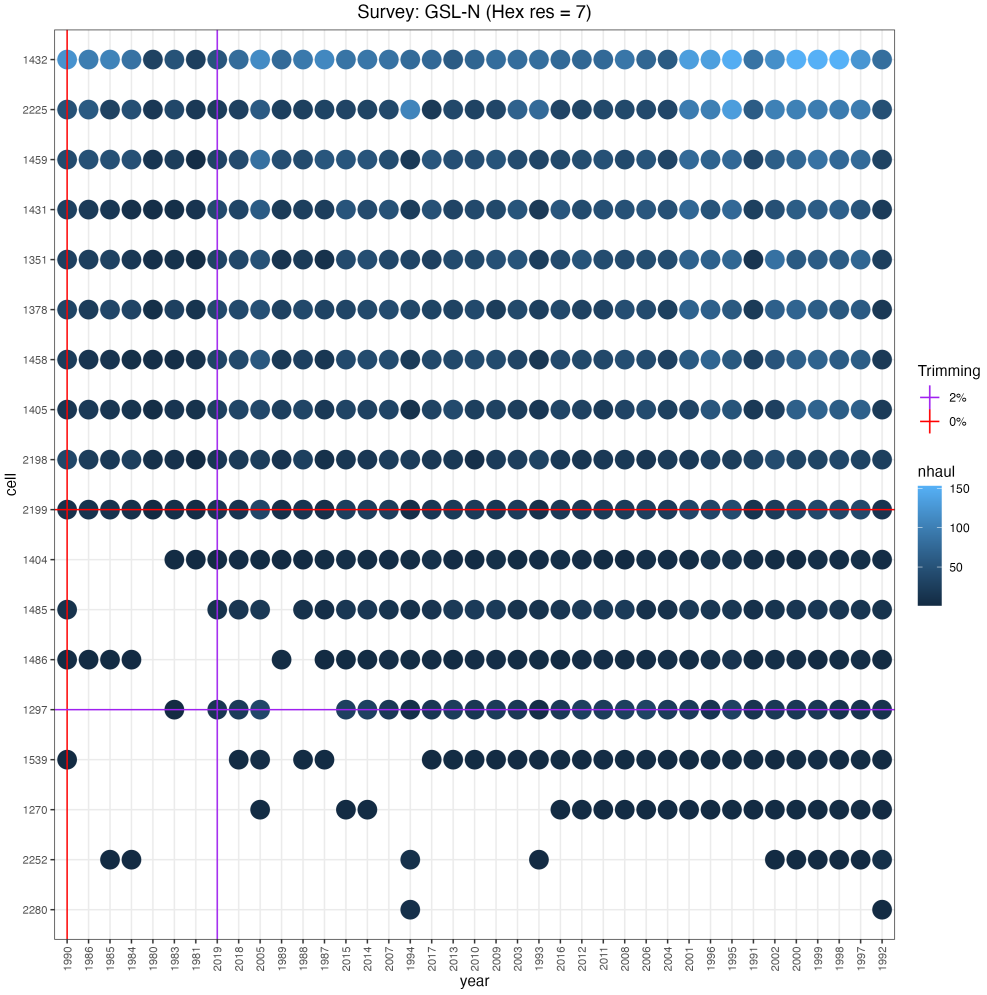
No flags

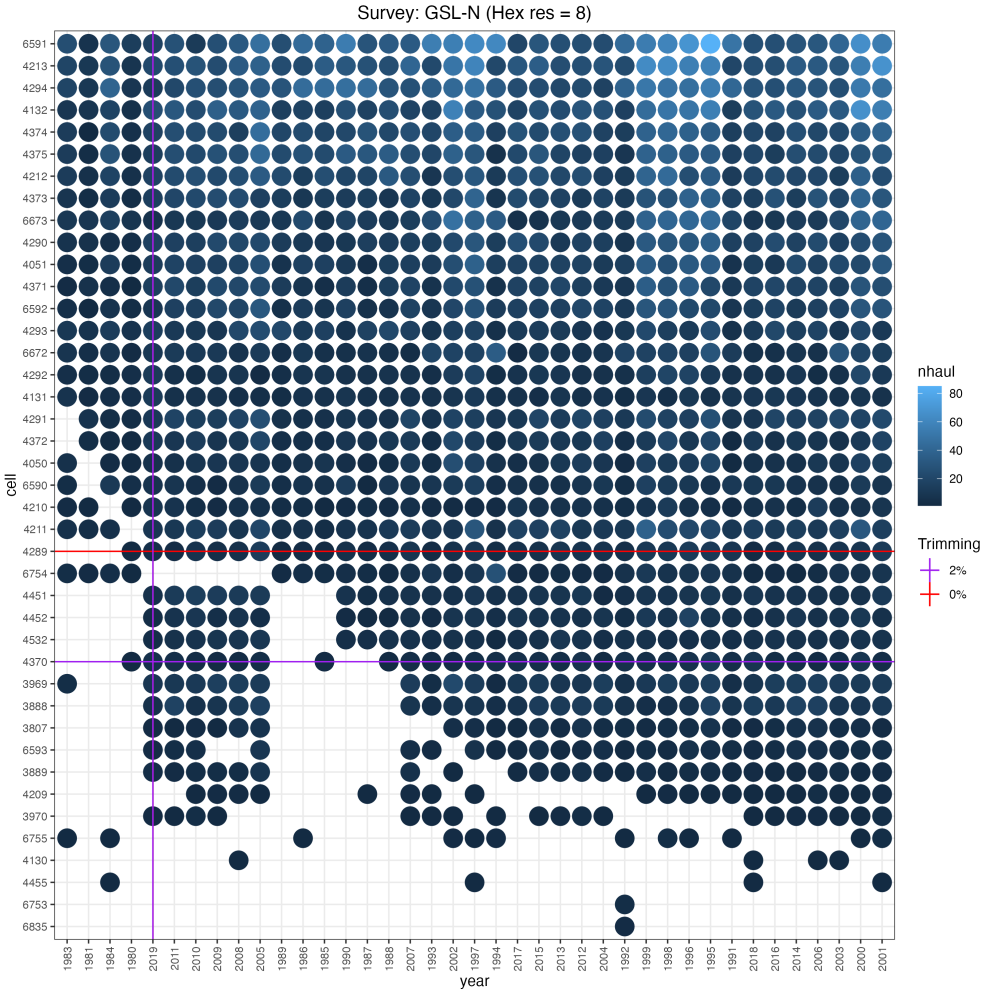
10. Spatio-temporal standardization

a. Standardization method 1

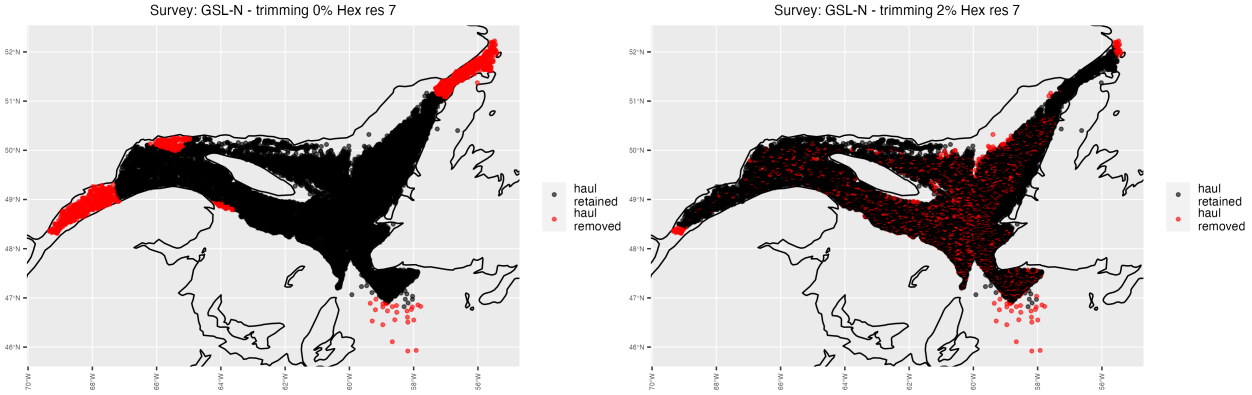
This standardization method was adapted from https://github.com/zoekitchel/trawl_spatial_turnover/blob/master/data_prep_code/species/explore_NorthSea_trimming.Rmd
It was run for hex resolution 7 and 8.

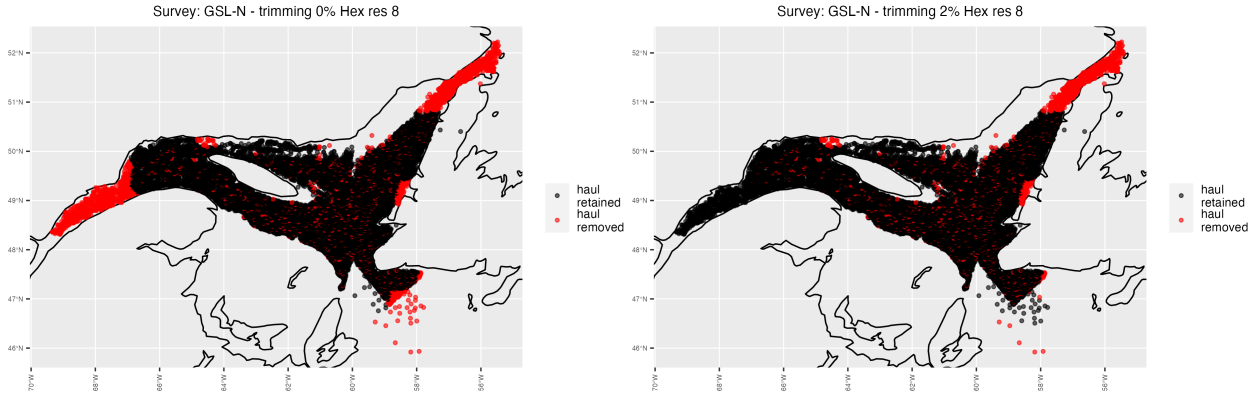
Plot of number of cells x years with overlaid flagging options



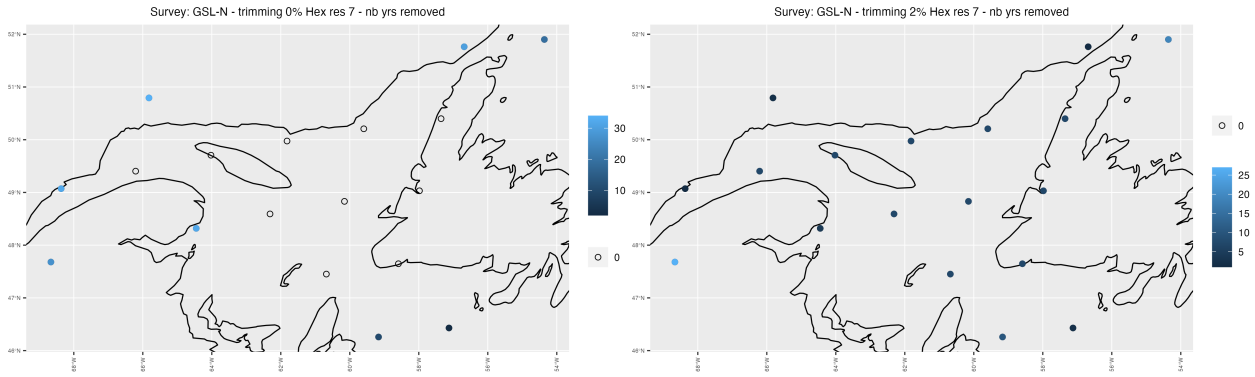


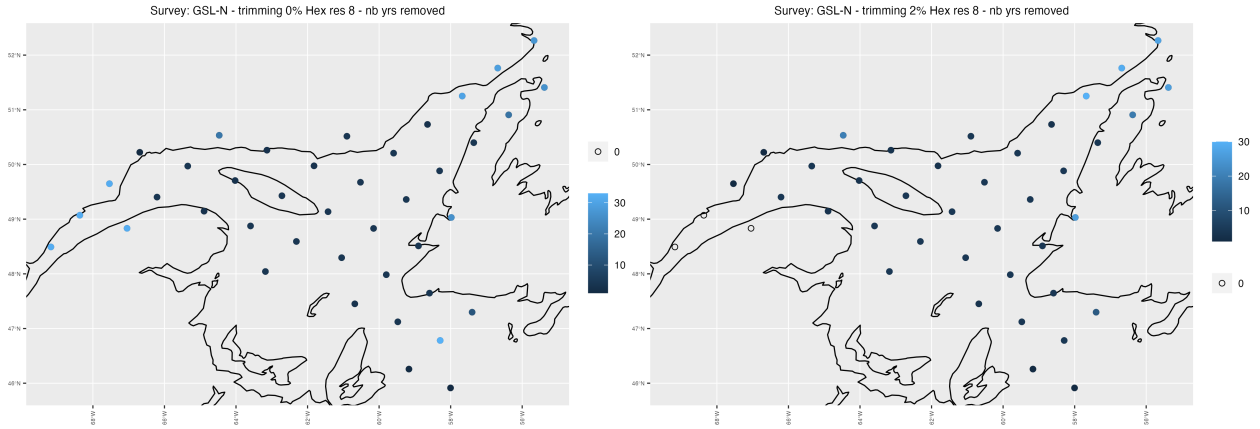
Map of hauls retained and removed per flagging method and threshold





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

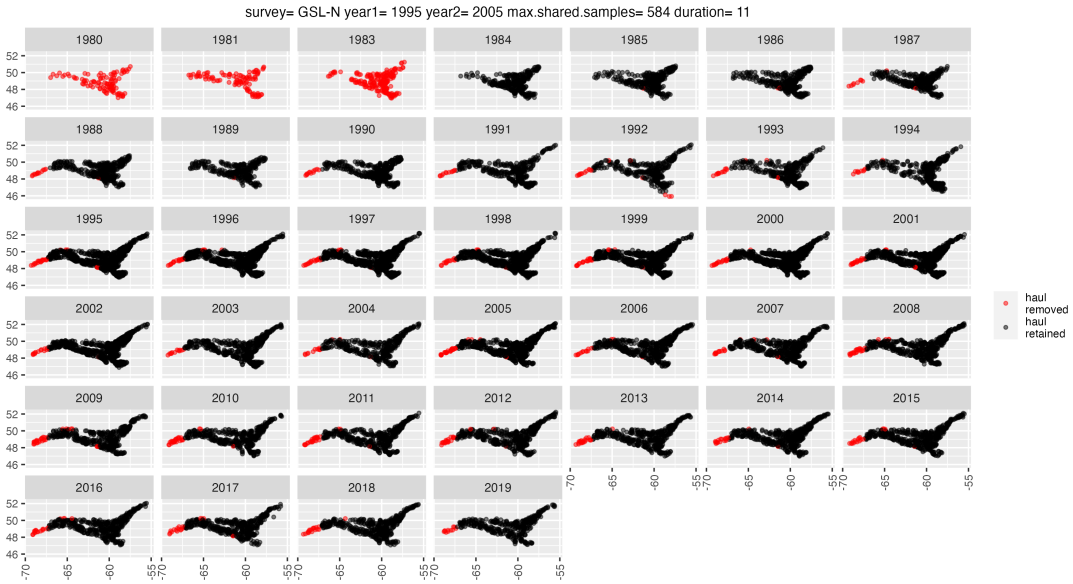




b. Standardization method 2

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	1418	1889.0	2533.0	1643.0	7281.0
percentage of hauls removed	8	10.7	14.3	9.3	7.7