SCS: Canadian Maritimes survey data processing summary

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

December, 2022

Contents

General info

This document presents the cleaning code and summary of the Canadian Maritimes bottom trawl survey provided by Mike McMahon, Don Clark, and Brian Bower. It contains data from 1970 and up to 2020.

Data cleaning in R

```
##############################################################################
#### R code to clean trawl survey for Canadian Maritimes
#### Public data Ocean Adapt
#### Contacts: Mike McMahon mike.mcmahon@dfo-mpo.gc.ca  Aquatic Science Biologist
####                 Population Ecology Division, DFO Canada
####            Don Clark   don.clark@dfo-mpo.gc.ca Biologist, DFO Canada
####            Brian Bower brian.bower@dfo-mpo.gc.ca
####                 GIS Analyst/ Physical Scientist at Fisheries and Oceans Canada
#### Coding: Michelle Stuart, Dan Forrest, Zoë Kitchel November 2021
```

```
##############################################################################################

#-------------------------------------------------------------------------------------------#
#### 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)

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

#"CPUE generally represents catch (numbers or weight) per standard tow length or per
#unit area. In the NAFO area, the primary sampling unit is the area swept by the trawl
#(AS) and is generally estimated by the product of the tow distance (t) and wing
#spread (WS). The true estimate of swept area is probably best represented by
#trawl door spread (DS), instead of wing spread (see Fig. 2) and will be discussed later."

#Therefore, at the suggestion of Capt. Baker, then
#Master of "Lady Hammond," the Atlantic Western IIA
#trawl was adopted as the standard groundfish survey
#trawl for Scotia-Fundy. This trawl was already
#highly successful in the regional, commercial
#fishing fleet and could be handled easily on "Lady
#Hammond." Being a box trawl, it fishes with a good
#headline height (about 15 ft (4.6 m)) and it has a
#similar wing spread (about 35 ft (10.7 m)) to the
#Yankee 36 trawl which had been the standard
#Scotia-Fundy groundfish survey trawl for years.
#Door spread: Door spread 110 ft (33.6 m)
#https://waves-vagues.dfo-mpo.gc.ca/Library/108919.pdf

#Data for the Canadian Maritimes can be best accessed using the Pinsky Lab
#Ocean Adapt Git Hub Repository.
#Contact malin.pinsky@rutgers.edu for questions or help accessing

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

spp_files <- list(
"https://raw.githubusercontent.com/pinskylab/OceanAdapt/master/data_raw/MAR_FALL_SPP.csv",
"https://raw.githubusercontent.com/pinskylab/OceanAdapt/master/data_raw/MAR_SPRING__SPP.csv",
"https://raw.githubusercontent.com/pinskylab/OceanAdapt/master/data_raw/MAR_SUMMER_SPP.csv")

#spp_files <- as.list(dir(pattern = "_SPP", path = "data_raw", full.names = T))
mar_spp <- spp_files %>% #this pulls in species from all three surveys, so there are
```

```
  #some repeats which I remove below
  map_dfr(~ read_csv(.x, col_types = cols(
    SPEC = col_character()
  )))

mar_spp <- mar_spp %>%
  rename(spp = SPEC,
         SPEC = CODE) %>%
  distinct()


mission_files <- list(
"https://raw.githubusercontent.com/pinskylab/OceanAdapt/master/data_raw/MAR_FALL_MISSION.csv",
"https://raw.githubusercontent.com/pinskylab/OceanAdapt/master/data_raw/MAR_SPRING_MISSION.csv",
"https://raw.githubusercontent.com/pinskylab/OceanAdapt/master/data_raw/MAR_SUMMER_MISSION.csv")
#mission_files <- as.list(dir(pattern = "_MISSION", path = "data_raw", full.names = T))
mar_missions <- mission_files %>%
  map_dfr(~ read_csv(.x, col_types = cols(
    .default = col_double(),
    MISSION = col_character(),
    VESEL = col_character(),
    SEASON = col_character()
  )))

info_files <- list(
"https://raw.githubusercontent.com/pinskylab/OceanAdapt/master/data_raw/MAR_FALL_INF.csv",
"https://raw.githubusercontent.com/pinskylab/OceanAdapt/master/data_raw/MAR_SPRING__INF.csv",
"https://raw.githubusercontent.com/pinskylab/OceanAdapt/master/data_raw/MAR_SUMMER_INF.csv")
#info_files <- as.list(dir(pattern = "_INF", path = "data_raw", full.names = T))
mar_info <- info_files %>%
  map_dfr(~ read_csv(.x, col_types = cols(
    .default = col_double(),
    MISSION = col_character(),
    SDATE = col_character(),
    GEARDESC = col_character(),
    STRAT = col_character()
  )))

catch_files <- list(
"https://raw.githubusercontent.com/pinskylab/OceanAdapt/master/data_raw/MAR_FALL_CATCH.csv",
"https://raw.githubusercontent.com/pinskylab/OceanAdapt/master/data_raw/MAR_SPRING__CATCH.csv",
"https://raw.githubusercontent.com/pinskylab/OceanAdapt/master/data_raw/MAR_SUMMER_CATCH.csv")
#catch_files <- as.list(dir(pattern = "_CATCH", path = "data_raw", full.names = T))
mar_catch <- catch_files %>%
  map_dfr(~ read_csv(.x, col_types = cols(
    .default = col_double(),
    MISSION = col_character()
  )))

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

```r
mar <- left_join(mar_catch, mar_missions, by = "MISSION")

mar <- mar %>%
  # Create a unique haul_id
  mutate(
    haul_id = paste(formatC(MISSION, width=3, flag=0),
                    formatC(SETNO, width=3, flag=0), sep = "_"))

mar_info <- mar_info %>%
  # Create a unique haul_id
  mutate(
    haul_id = paste(formatC(MISSION, width=3, flag=0),
                    formatC(SETNO, width=3, flag=0), sep = "_"))

mar <- left_join(mar, mar_info, by = c("haul_id", "MISSION","SETNO")) #206202 rows
mar <- left_join(mar, mar_spp, by = "SPEC")
mar$survey <- "SCS"

names(mar) <- tolower(names(mar))


mar <- mar %>%
  # convert mission to haul_id
  rename(wgt = totwgt,
         num = totno,
         latitude = slat,
         longitude = slong,
         stratum = strat,
         gear = geardesc,
         sbt = bott_temp,
         sst = surf_temp,
         verbatim_name = spp,
         year = year,
         depth = depth) %>%
 # area swept by net in km^2 = 33.6 m door spread *
  #DIST in nautical miles * 1852 m/1 nautical mile * 1 km^2/1000000 m^2
  mutate(area_swept = 33.6 * dist * 1852 *(1/1000000),
         month = month(as.Date(sdate)),
         day = day(as.Date(sdate)),
         haul_dur = dur/60) #minutes to hours


# Does the spp column contain any eggs or non-organism notes?
#As of 2021, only "UNIDENTIFIED" to be  removed
test <- mar %>%
  select(verbatim_name) %>%
  filter(!is.na(verbatim_name)) %>%
  distinct() %>%
  filter((grepl("egg", verbatim_name) & grepl("", verbatim_name)) |
           grepl("UNIDENTIFIED", verbatim_name)) #does it contain egg or unidentified?
stopifnot(nrow(test)==0)

#delete any rows with any of these
```

```
mar <- mar %>% #206202 to 205205 rows
  filter(!grepl("UNIDENTIFIED",verbatim_name))

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

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

# combine the wtcpue for each species by haul
mar <- mar %>%
  mutate(
    wgt_cpue = wgt/area_swept,
    wgt_h = wgt/haul_dur, #may need to change this unit, currently in minutes
    num_cpue = num/area_swept,
    num_h = num/haul_dur

  )

mar <- mar %>%
# Adding extra columns and setting proper format
mutate(
  country = "Canada",
  source = "DFO",
  timestamp = mdy("02/08/2021"),
  sub_area = NA,
  continent = "n_america",
  stat_rec = NA,
  station = 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
                          )
                   )
  ),
  verbatim_aphia_id = 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, verbatim_name, num, num_h, num_cpue,
         wgt, wgt_h, wgt_cpue, verbatim_name, verbatim_aphia_id)

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

#none!

#which ones are duplicated?
```

```r
unique_name_match <- count_mar %>%
  group_by(verbatim_name) %>%
  filter(count>1) %>%
  distinct(verbatim_name)

unique_name_match
#empty (fixed earlier in ~178)

#-------------------------------------------------------------------------------------------#
#### 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
scs_survey_code <- "SCS"

scs <- mar %>%
  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(scs$taxa2), input_survey = scs_survey_code,
                         save = F, output=NA)
#takes 3.9 minutes

#This leaves out the following species, of which 2 are fish that need to be added back
#Caelorinchus caelorinchus    #fish
#Porania pulvillus
#Poraniomorpha borealis
#Notoscopelus elongatus kroyeri #fish, different fishbase record for Noto elon and
#                                      Noto kroy
#                                      Noto elon is endemic to Mediterranean, so we
#                                      will move forward as if this is Notoscopelus kroyeri
#Spirontocaris fabricii
#Nereidae
#Coelenterata

cae_cae <- c("Caelorinchus caelorinchus", "398381", "1726", "Coelorinchus caelorhincus",
             "Animalia", "Chordata", "Actinopteri", "Gadiformes", "Macrouridae",
             "Coelorinchus", "Species",
             "SCS")
not_elo <- c("Notoscopelus elogatus kroyeri", "272728", "27753", "Notoscopelus kroyeri",
             "Animalia", "Chordata", "Actinopteri", "Myctophiformes", "Myctophidae",
             "Notoscopelus", "Species",
             "SCS")
```

```r
clean_auto_missing <- rbind(clean_auto, cae_cae, not_elo)

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

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

clean_scs <- left_join(scs, 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) %>%
  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, verbatim_aphia_id, accepted_name, aphia_id, SpecCode,
         kingdom, phylum, class, order, family, genus, rank)


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

#none!

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

unique_name_match
#empty


# -----------------------------------------------------------------------------------------#
#### SAVE DATABASE IN GOOGLE DRIVE ####
# -----------------------------------------------------------------------------------------#
# Just run this routine should be good for all
write_clean_data(data = clean_scs, survey = "SCS", overwrite = T)
```

1. Overview of the survey data table

survey	haul_id	source	timestamp	country	sub_area	continent
SCS	HAM1982085_054	DFO	2021-02-08	Canada	NA	n_america
SCS	HAM1982084_011	DFO	2021-02-08	Canada	NA	n_america
SCS	HAM1982084_019	DFO	2021-02-08	Canada	NA	n_america
SCS	HAM1981064_025	DFO	2021-02-08	Canada	NA	n_america
SCS	HAM1982084_054	DFO	2021-02-08	Canada	NA	n_america

stat_rec	station	stratum	year	month	day	quarter	season
NA	NA	481	1982	10	20	4	FALL
NA	NA	454	1982	9	29	3	FALL
NA	NA	455	1982	9	29	3	FALL
NA	NA	464	1981	10	3	4	FALL
NA	NA	443	1982	10	4	4	FALL

latitude	longitude	haul_dur	area_swept	gear	depth
42.48333	-65.40000	0.5	0.1057862	Western IIA trawl	98.76
43.26667	-61.26667	0.5	0.1057862	Western IIA trawl	118.87
43.86667	-60.33333	0.5	0.1120090	Western IIA trawl	31.09
43.73333	-61.38333	0.5	0.1057862	Western IIA trawl	51.21
45.35000	-59.26667	0.5	0.1120090	Western IIA trawl	91.44

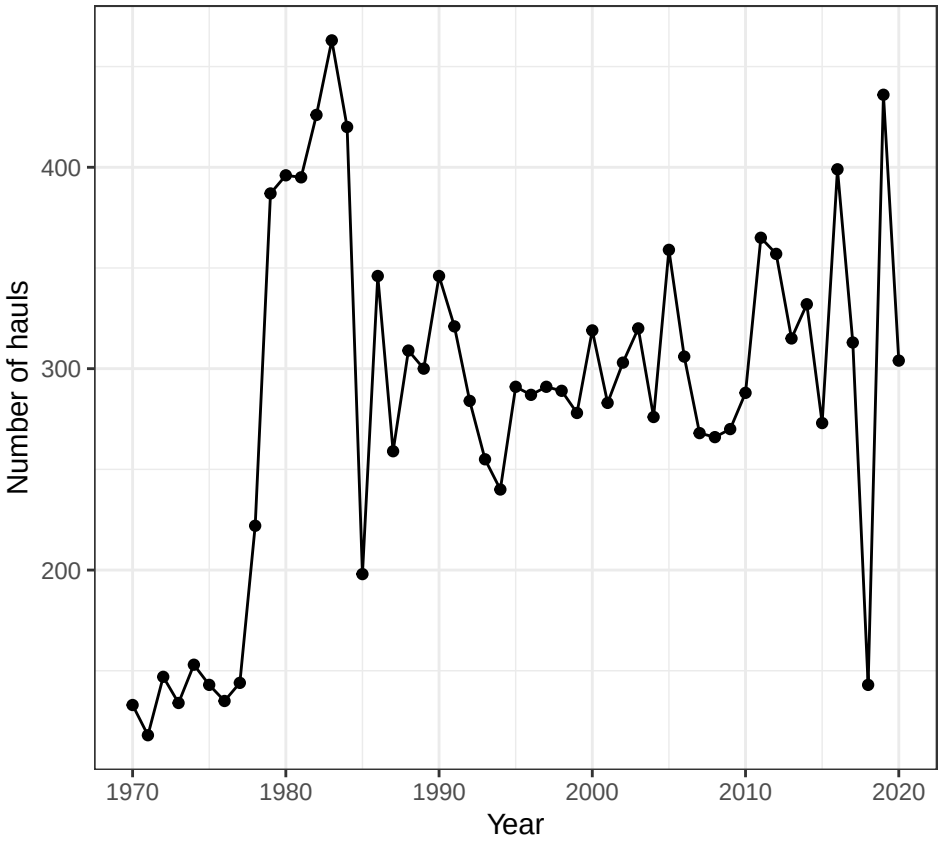
sbt	sst	num	num_h	num_cpue	wgt
7.20	12.8	1	2	9.453025	0
2.56	17.9	1	2	9.453025	0
15.42	16.2	1	2	8.927857	0
12.45	15.5	1	2	9.453025	0
1.46	13.5	1	2	8.927857	0

wgt_h	wgt_cpue	verbatim_name	verbatim_aphia_id	accepted_name
0	0	BALISTES CAPRISCUS	NA	Balistes capriscus
0	0	STEPHANOLEPIS HISPIDUS	NA	Stephanolepis hispida
0	0	STEPHANOLEPIS HISPIDUS	NA	Stephanolepis hispida
0	0	STEPHANOLEPIS HISPIDUS	NA	Stephanolepis hispida
0	0	STEPHANOLEPIS HISPIDUS	NA	Stephanolepis hispida

aphia_id	SpecCode	kingdom	phylum	class	order	family
154721	7327	Animalia	Chordata	Actinopteri	Tetraodontiformes	Balistidae
307126	4281	Animalia	Chordata	Actinopteri	Tetraodontiformes	Monacanthidae
307126	4281	Animalia	Chordata	Actinopteri	Tetraodontiformes	Monacanthidae
307126	4281	Animalia	Chordata	Actinopteri	Tetraodontiformes	Monacanthidae
307126	4281	Animalia	Chordata	Actinopteri	Tetraodontiformes	Monacanthidae

2. Summary of sampling intensity

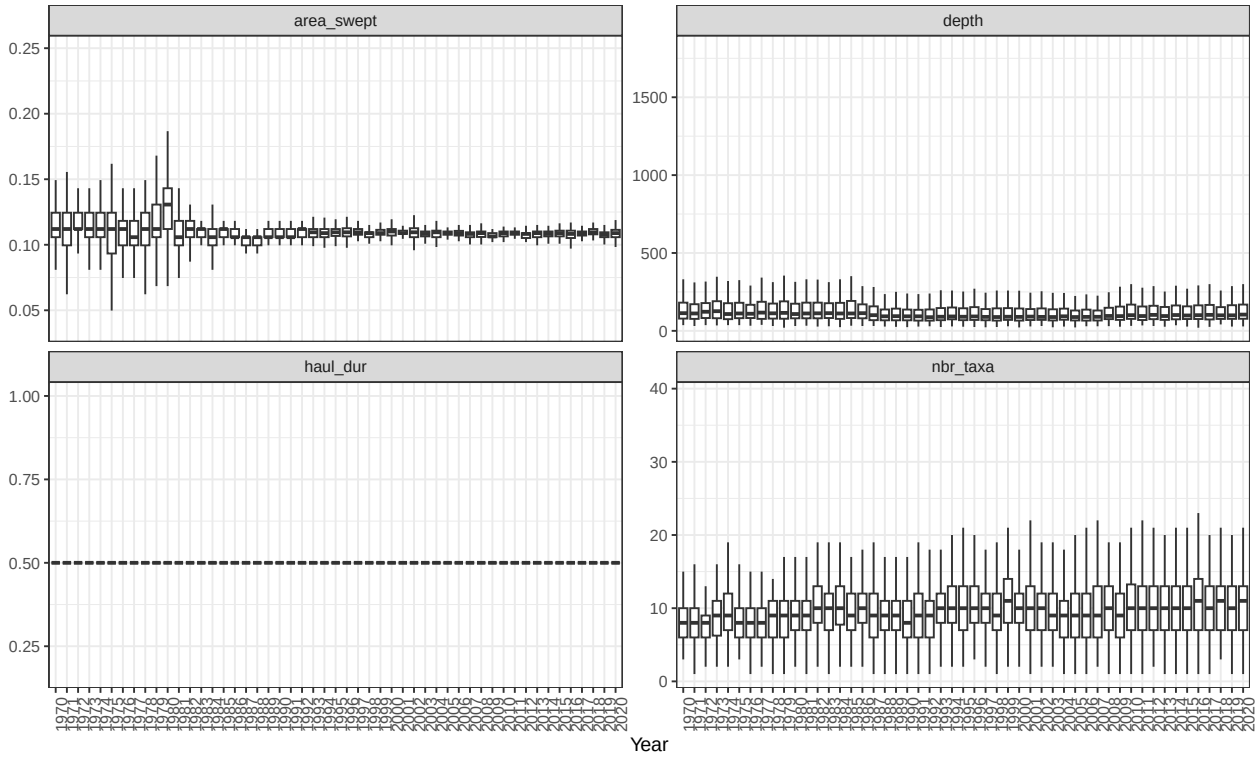
Number of hauls per year performed during the survey after data processing.



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:

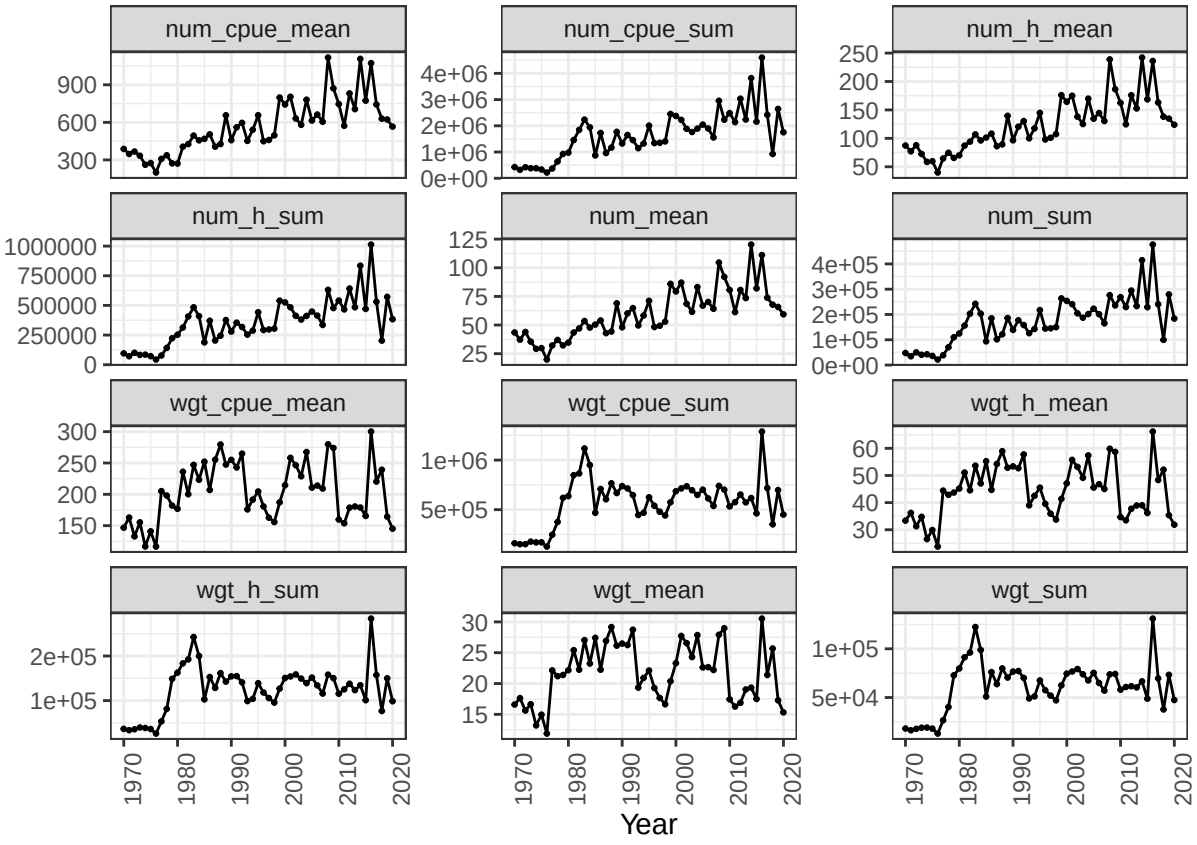
- *area_swept*, swept area by the bottom trawl gear km^2
- *depth*, sampling depth in m
- *haul_dur*, haul sampling duration *hour*
- *number of marine fish taxa*, taxa were cleaned following the last version of taxonomy from the World Register of Marine Species (https://www.marinespecies.org/, October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



4. Summary of biological variables

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

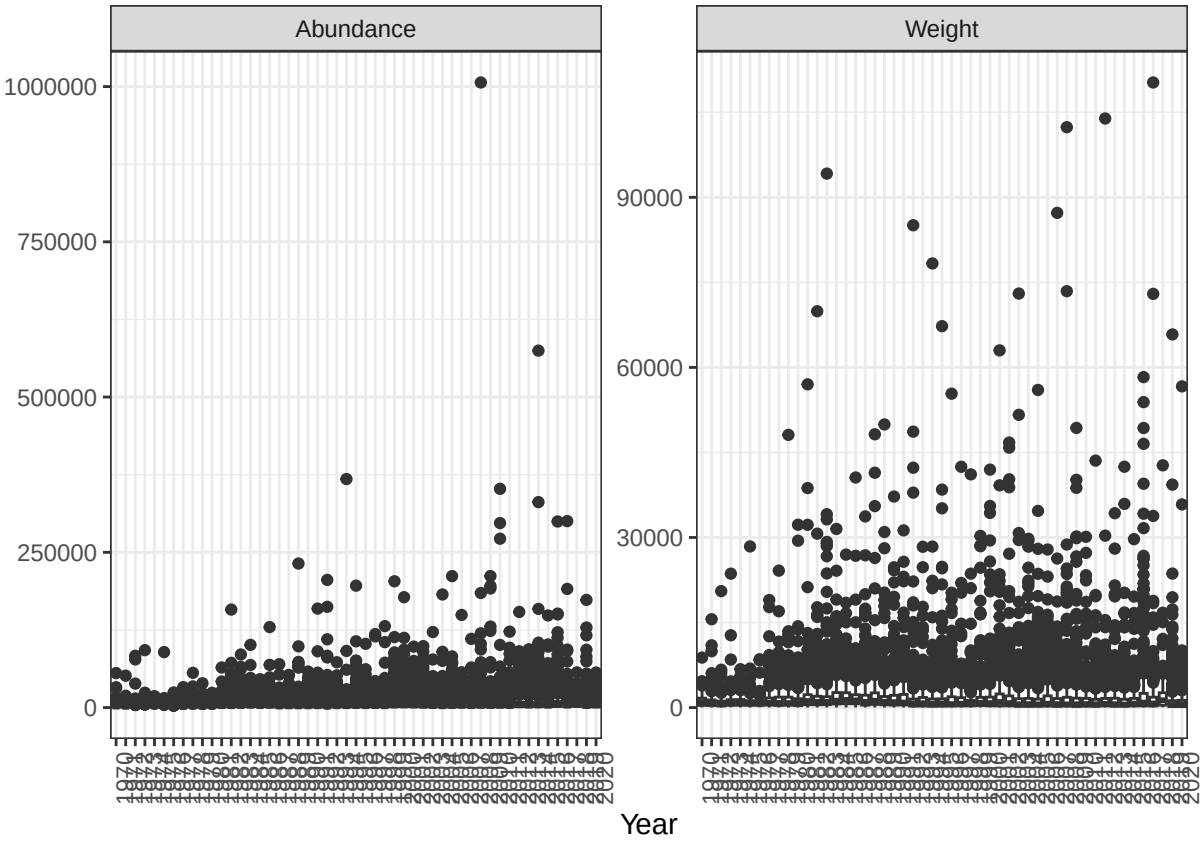
- *num_cpue*, number of individuals (abundance) in $\frac{individuals}{km^2}$
- *num_h*, number of individuals (abundance) in $\frac{individuals}{h}$
- *num*, number of individuals (abundance)
- *wgt_cpue*, weight in $\frac{kg}{km^2}$
- *wgt_h*, weight in $\frac{kg}{h}$
- *wgt*, weight in kg



5. Extreme values

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

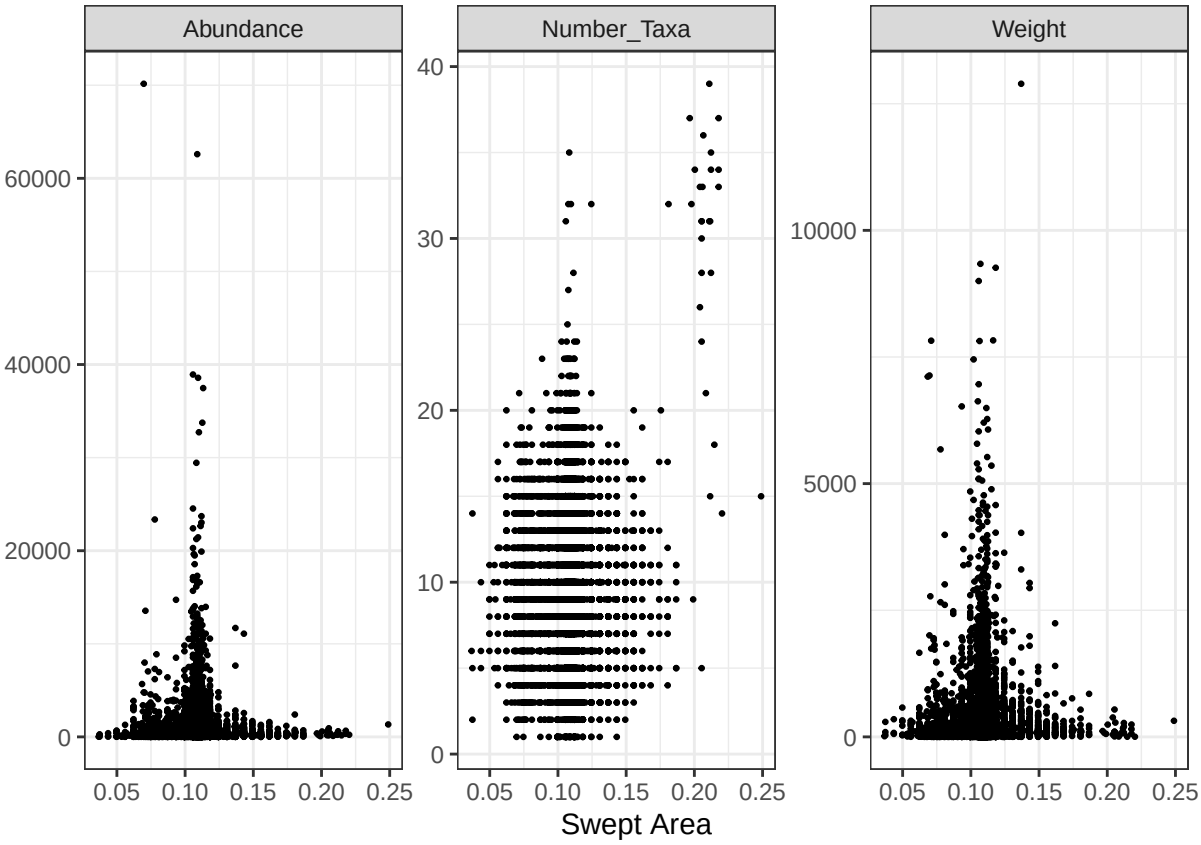
- *wgt*, total weight in *kg* per haul and year per haul and year, if available in the survey data
- *num*, total number of individuals, if available in the survey data



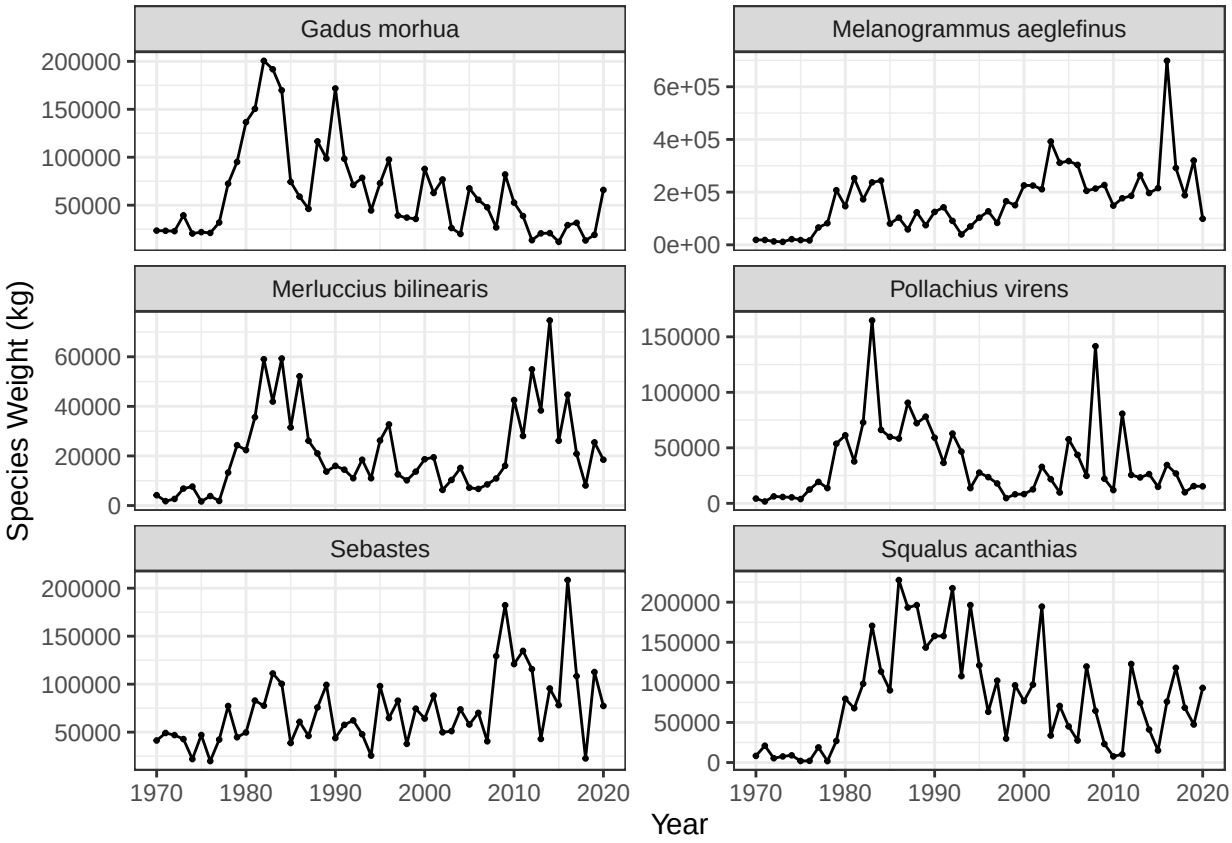
6. Summary of variables against swept area

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*, number of individuals, if available in the survey data
- *wgt*, weight in *kg*, if available in the survey data

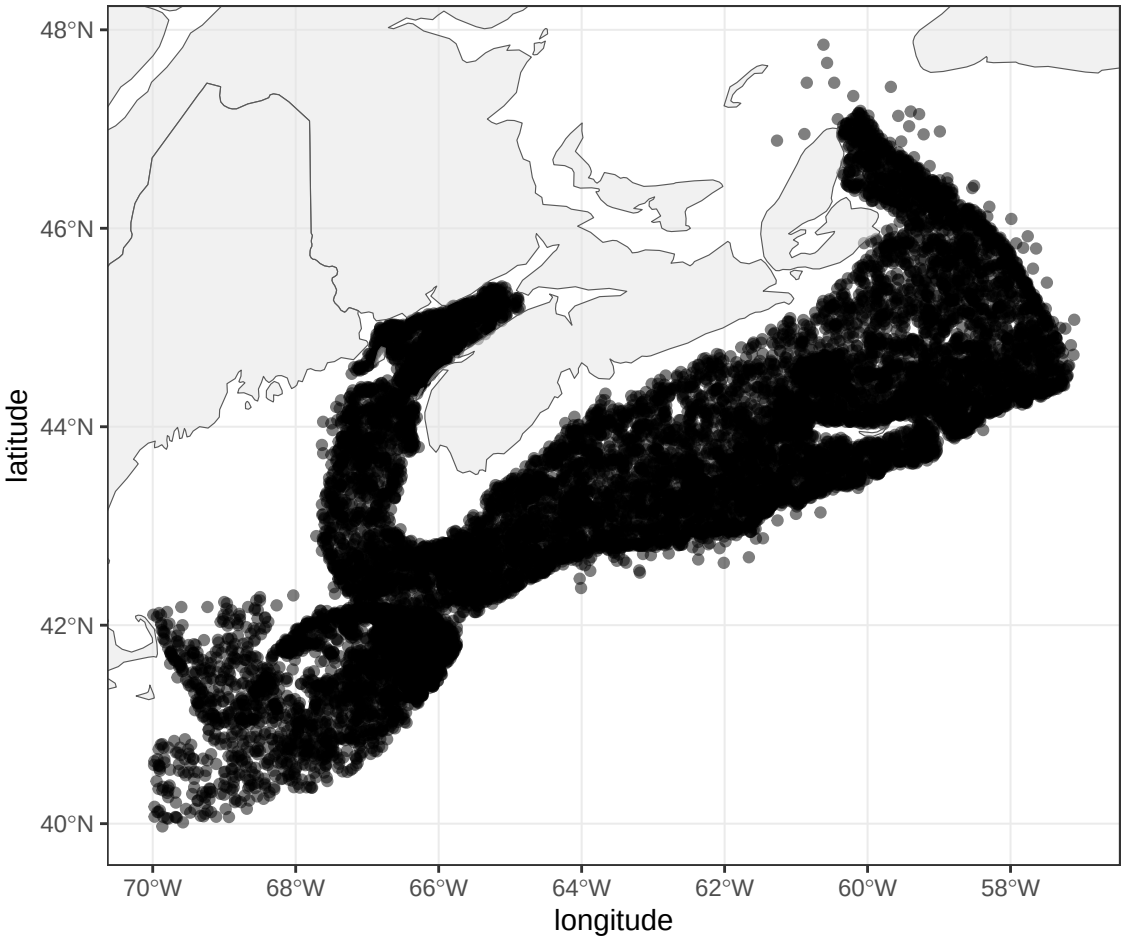


7. Abundance or Weight trends of the six most abundant species



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

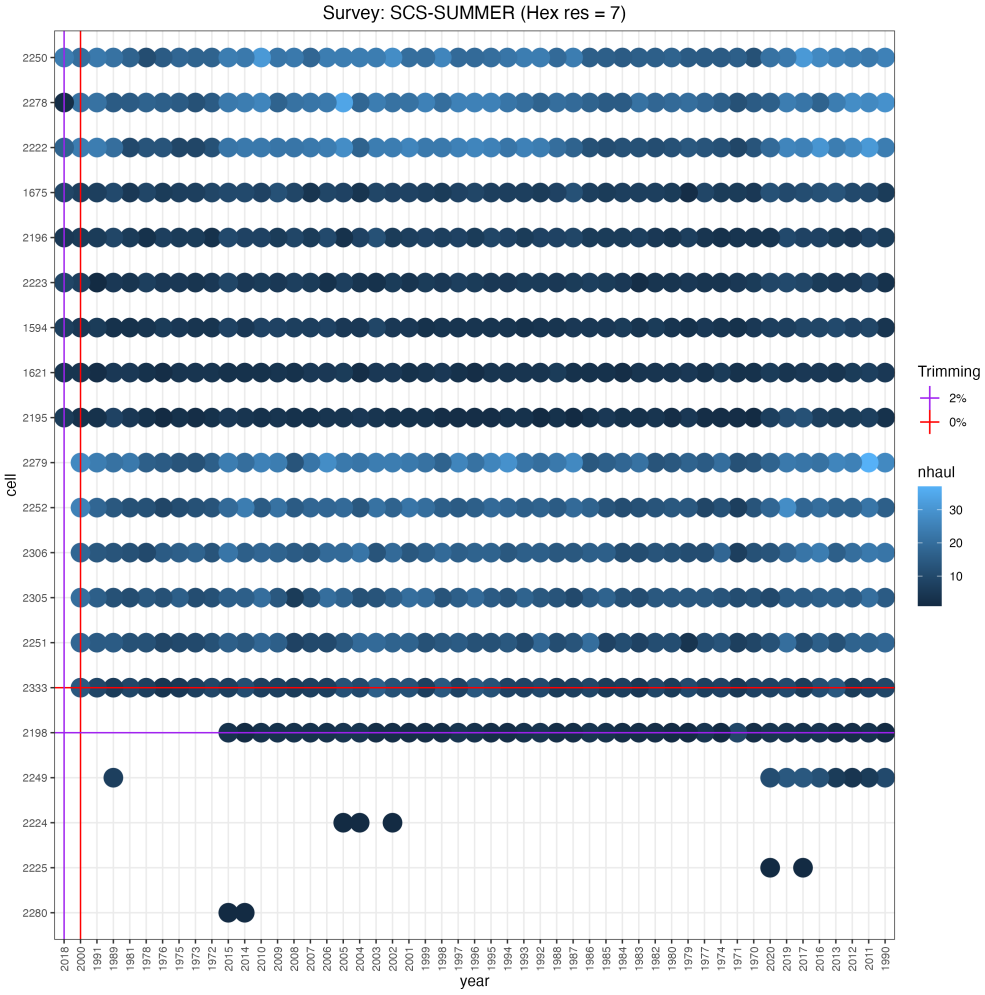
Statistics related to the taxonomic flagging outputs

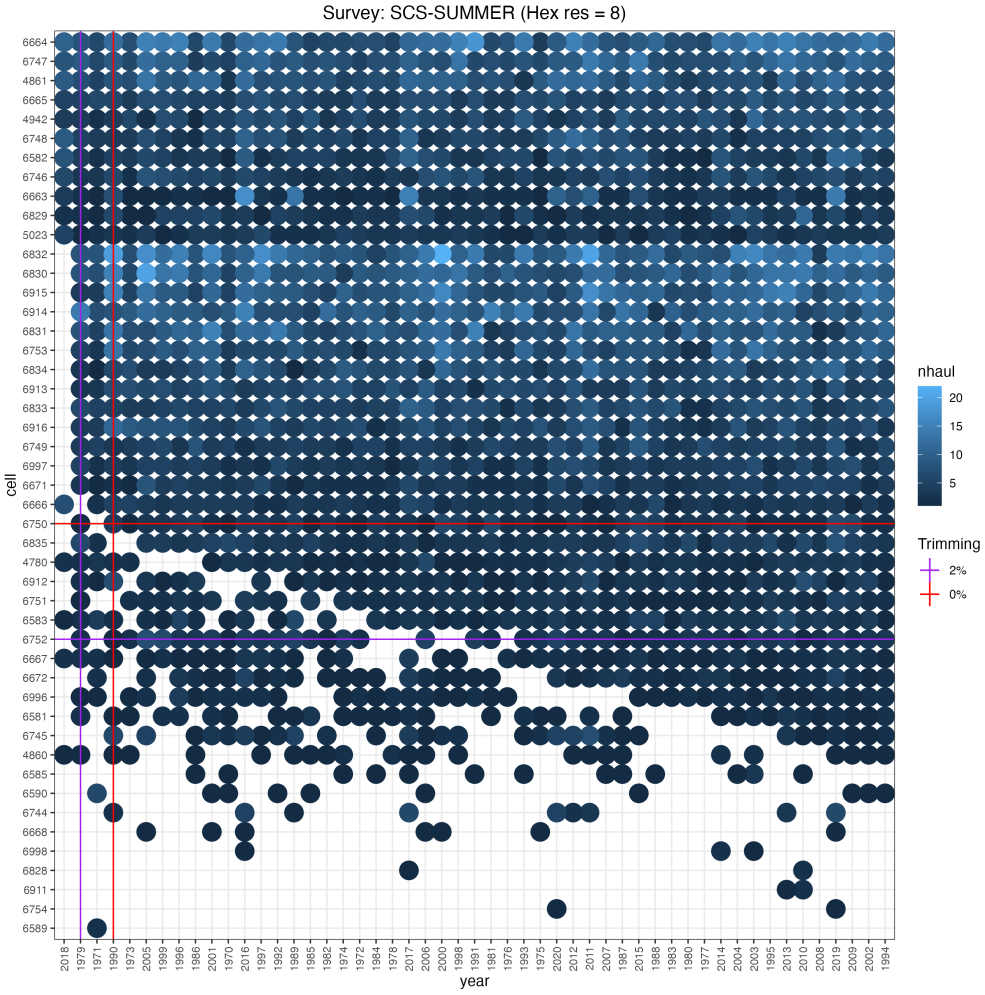
10. Spatio-temporal standardization: SCS-SUMMER

a. Standardization method 1

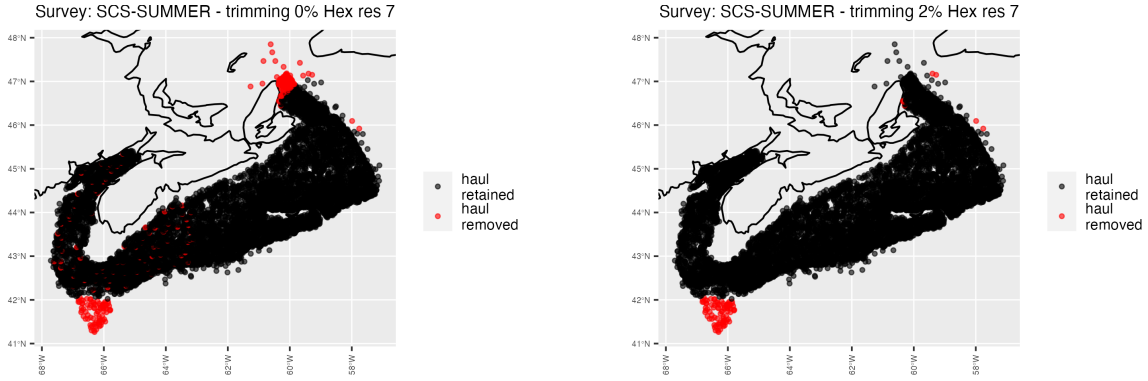
This standardization method was adapted from https://github.com/zoekitchel/trawl_spatial_turnover/blob/master/data_prep_code/species/explore_NorthSea_trimming.Rmd
It was run for hex resolution 7 and 8.

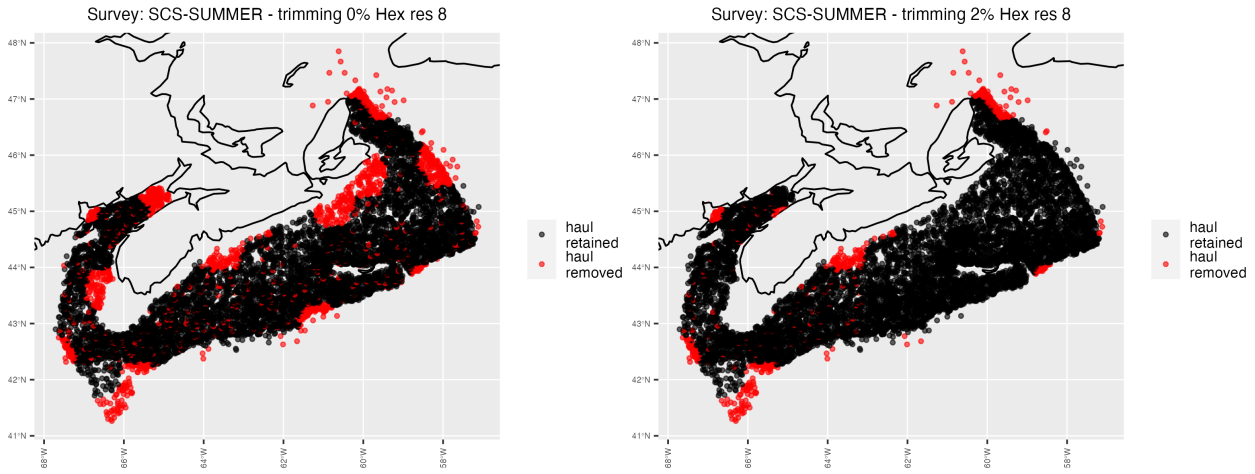
Plot of number of cells x years with overlaid flagging options


Survey: SCS-SUMMER (Hex res = 7)
cell
2250
2278
2222
1675
2196
2223
1594
1621
2195
2279
2252
2306
2305
2251
2333
2198
2249
2224
2225
2280
year
2018 2000 1991 1989 1981 1978 1976 1975 1973 1972 2015 2014 2010 2009 2008 2007 2006 2005 2004 2003 2002 2001 1999 1998 1997 1996 1995 1994 1993 1992 1988 1987 1986 1985 1984 1983 1982 1980 1979 1977 1974 1971 1970 2020 2019 2017 2016 2013 2012 2011 1990
Trimming
2%
0%
nhaul
30
20
10

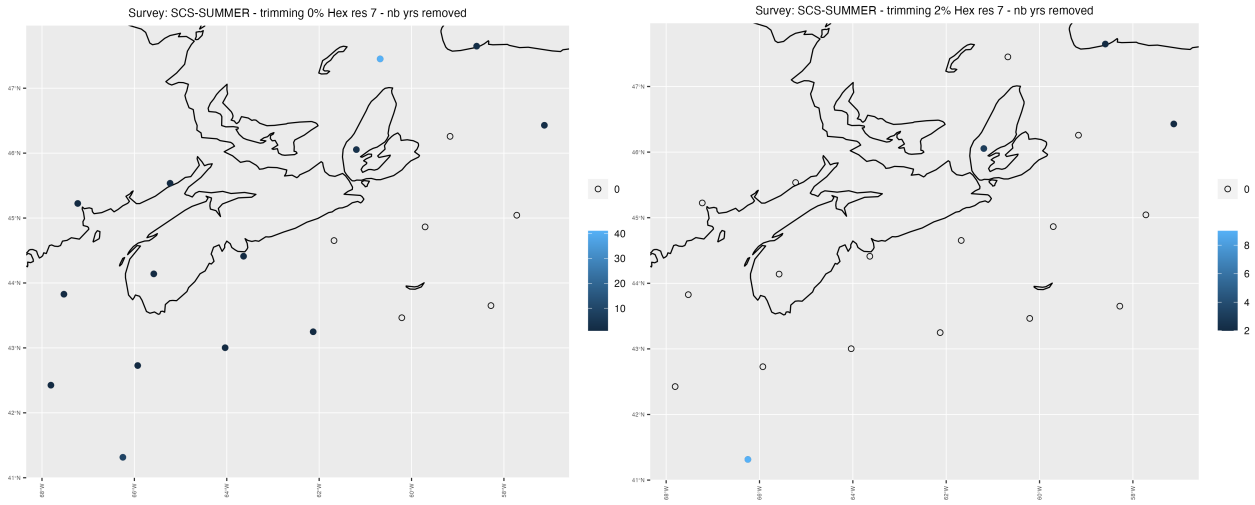


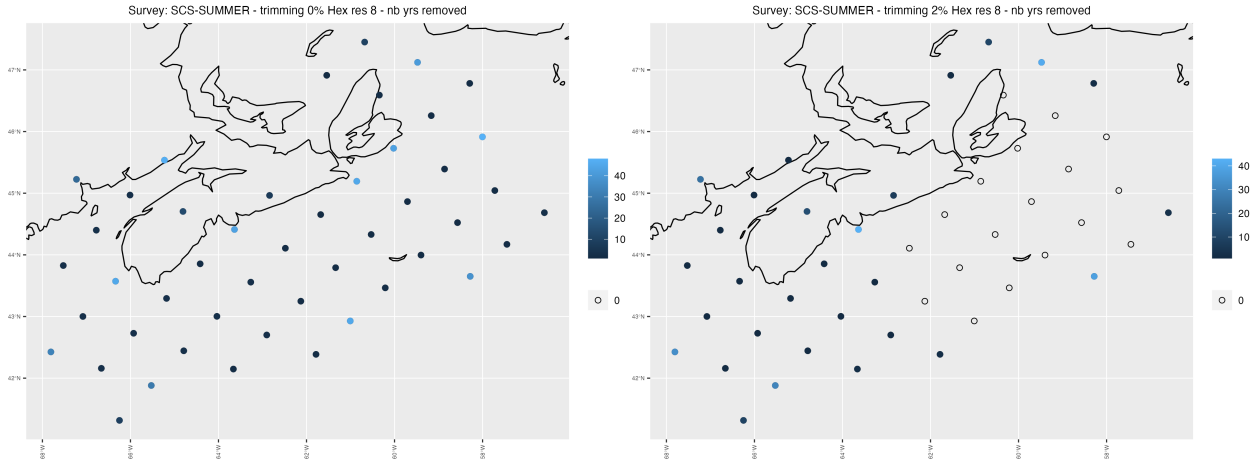
Map of hauls retained and removed per flagging method and threshold





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

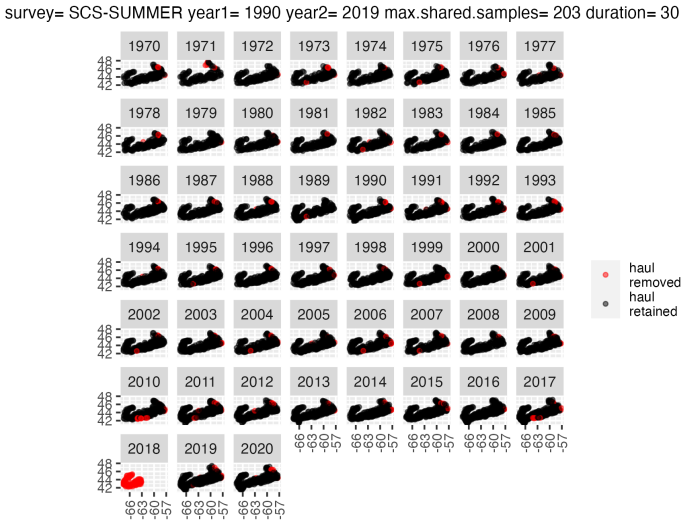




b. Standardization method 2

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

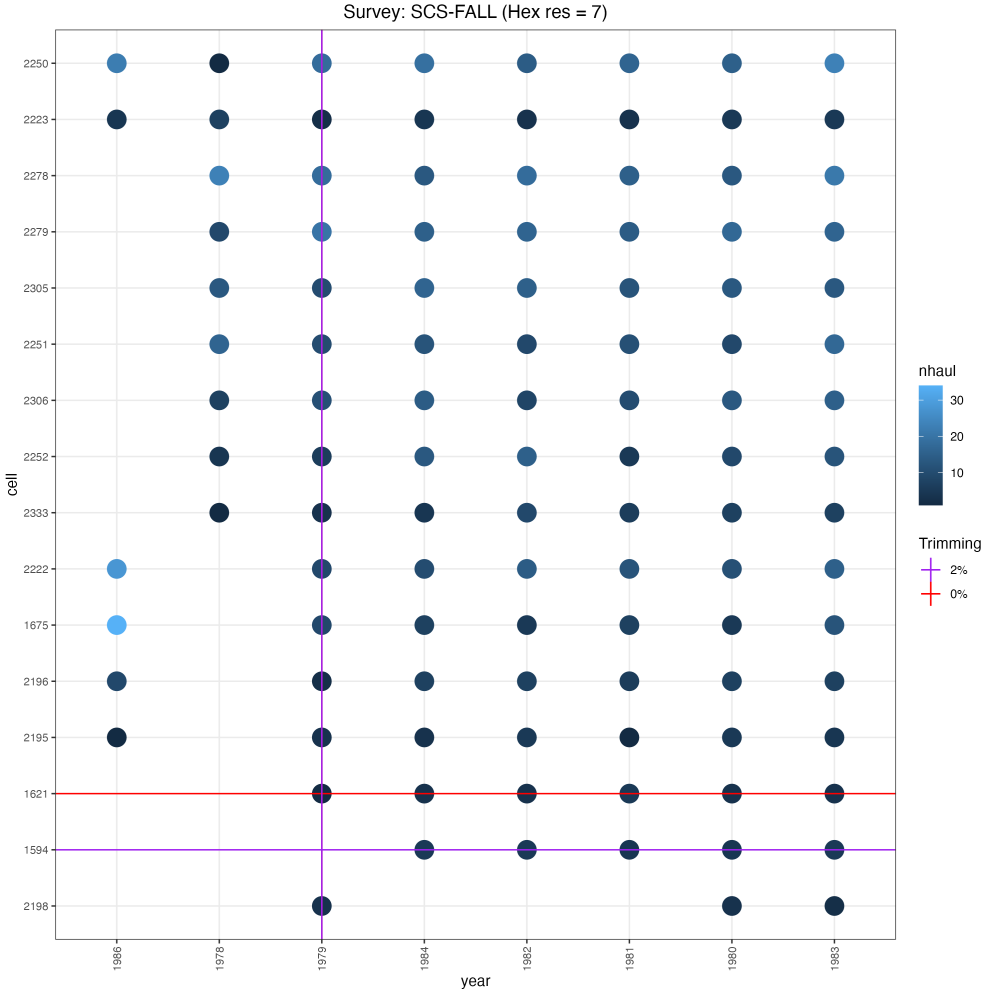
summary	grid cell 7, 0% threshold	grid cell 7, 2% threshold	grid cell 8, 0% threshold	grid cell 8, 2% threshold	method 2 (biotime)
number of hauls removed	287	91	1524.0	529.0	3309.0
percentage of hauls removed	3	1	16.2	5.6	3.6

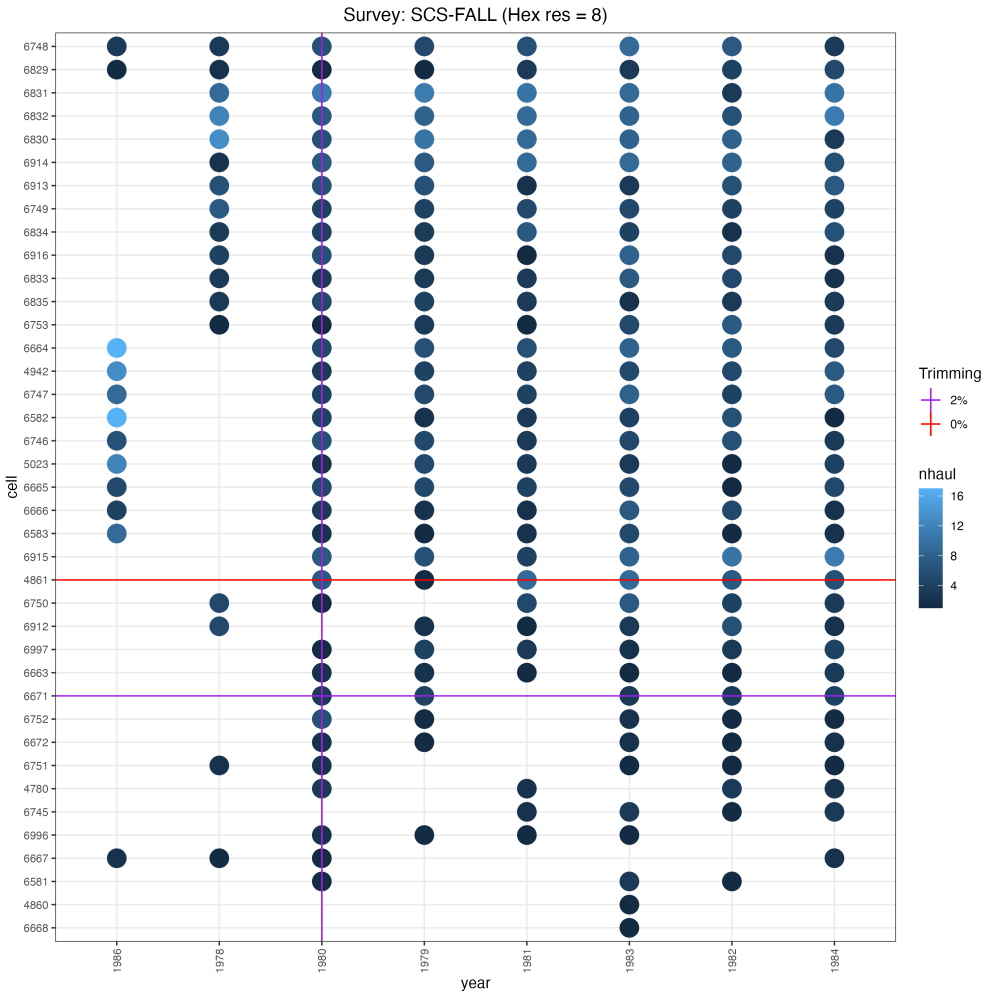
11. Spatio-temporal standardization: SCS-FALL

a. Standardization method 1

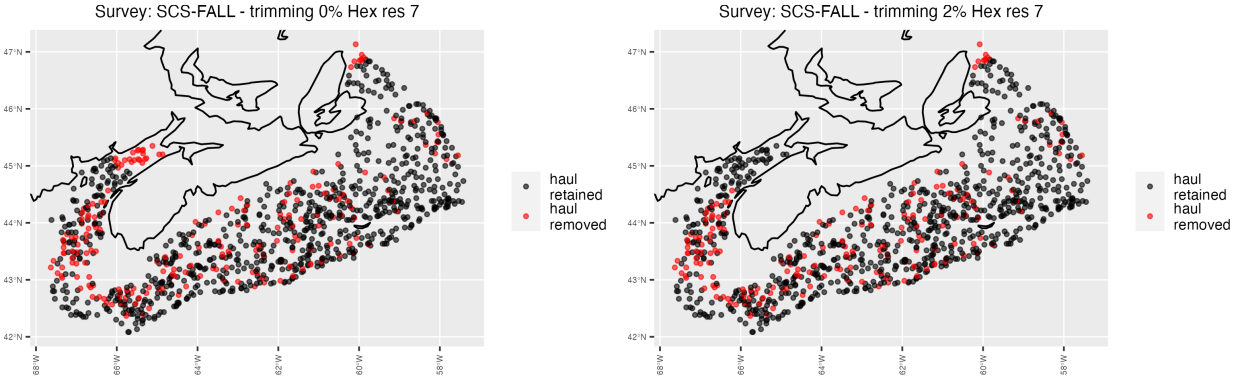
This standardization method was adapted from https://github.com/zoekitchel/trawl_spatial_turnover/blob/master/data_prep_code/species/explore_NorthSea_trimming.Rmd
It was run for hex resolution 7 and 8.

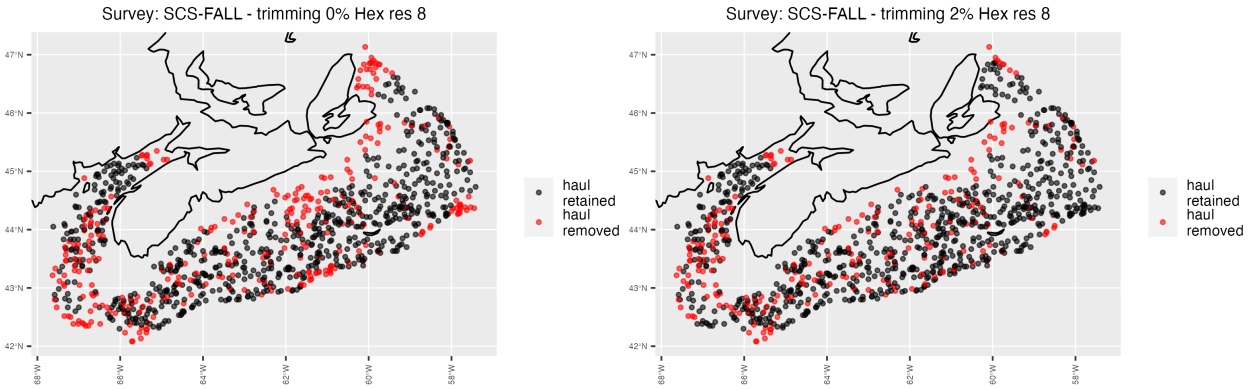
Plot of number of cells x years with overlaid flagging options



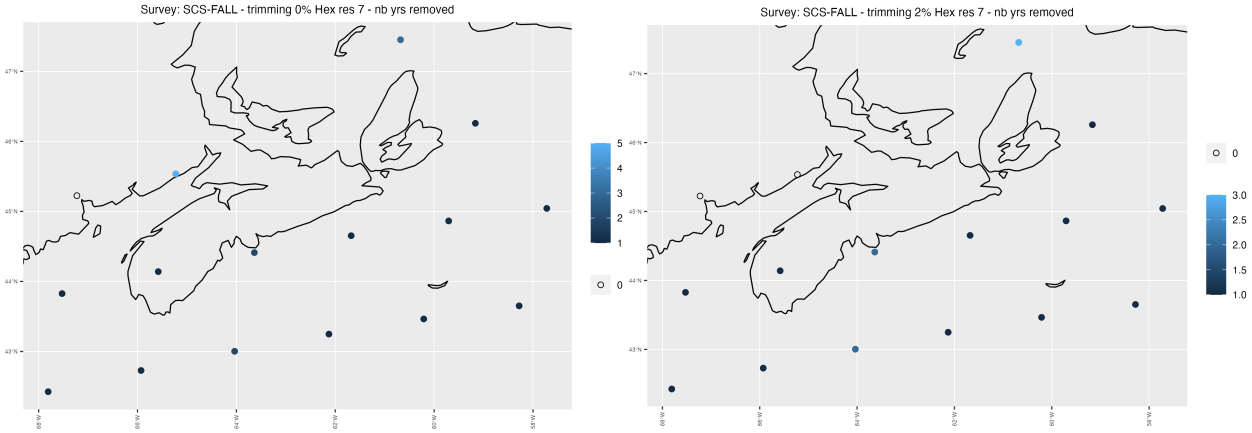


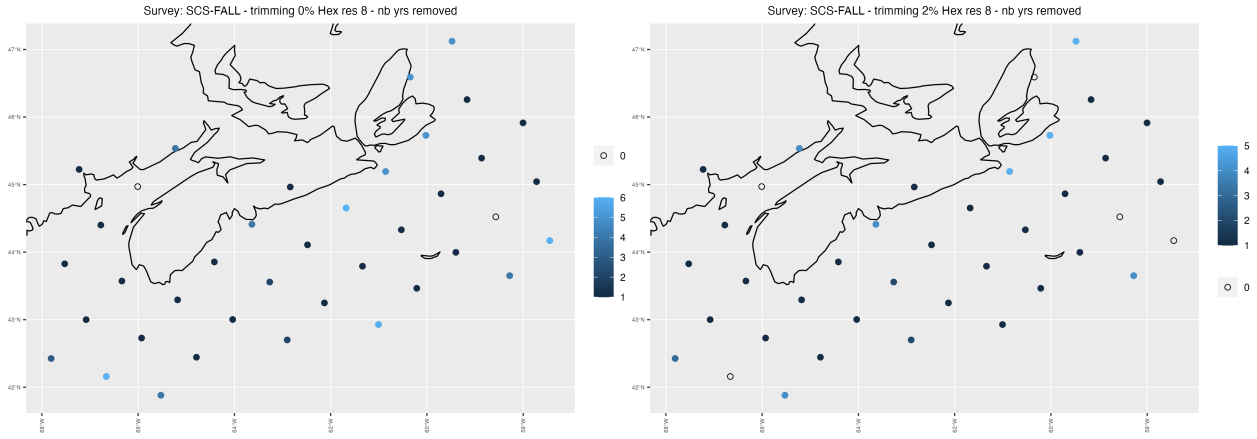
Map of hauls retained and removed per flagging method and threshold





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

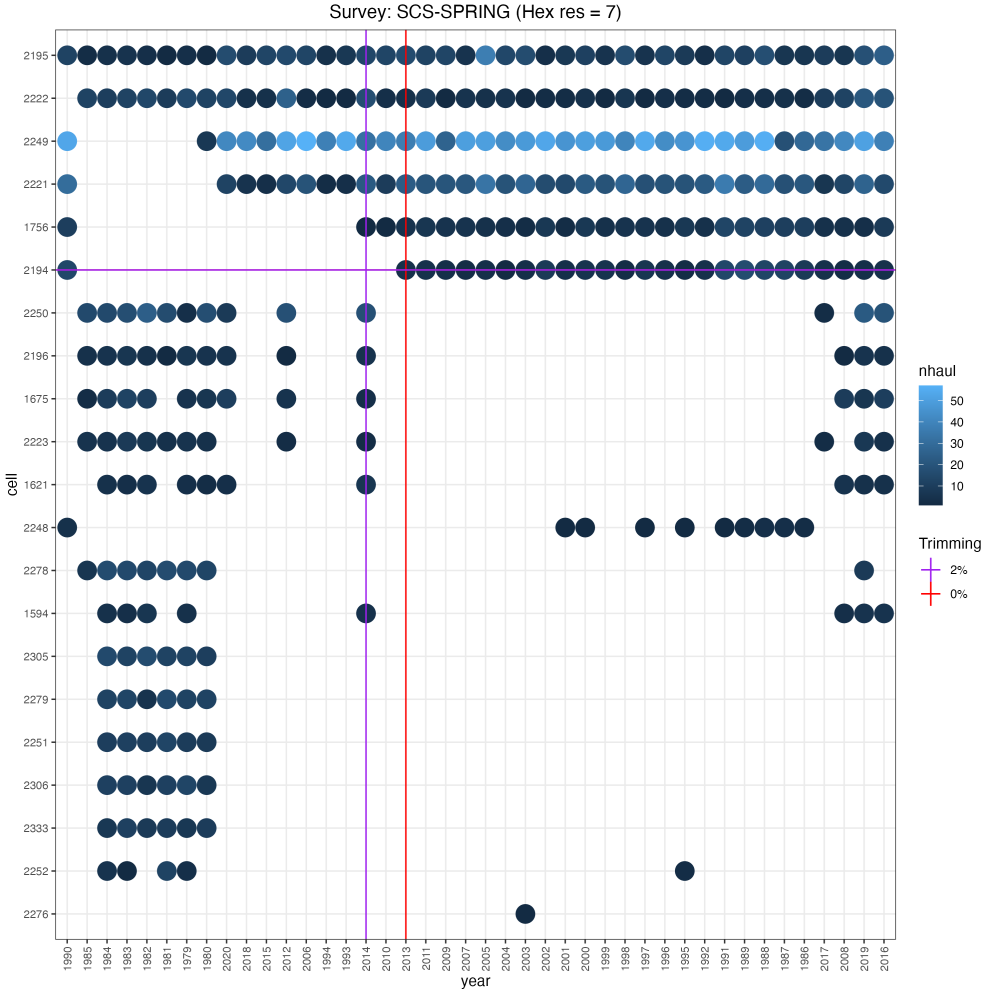
summary	grid cell 7, 0% threshold	grid cell 7, 2% threshold	grid cell 8, 0% threshold	grid cell 8, 2% threshold
number of hauls removed	210.0	187	315.0	237.0
percentage of hauls removed	20.2	18	30.3	22.8

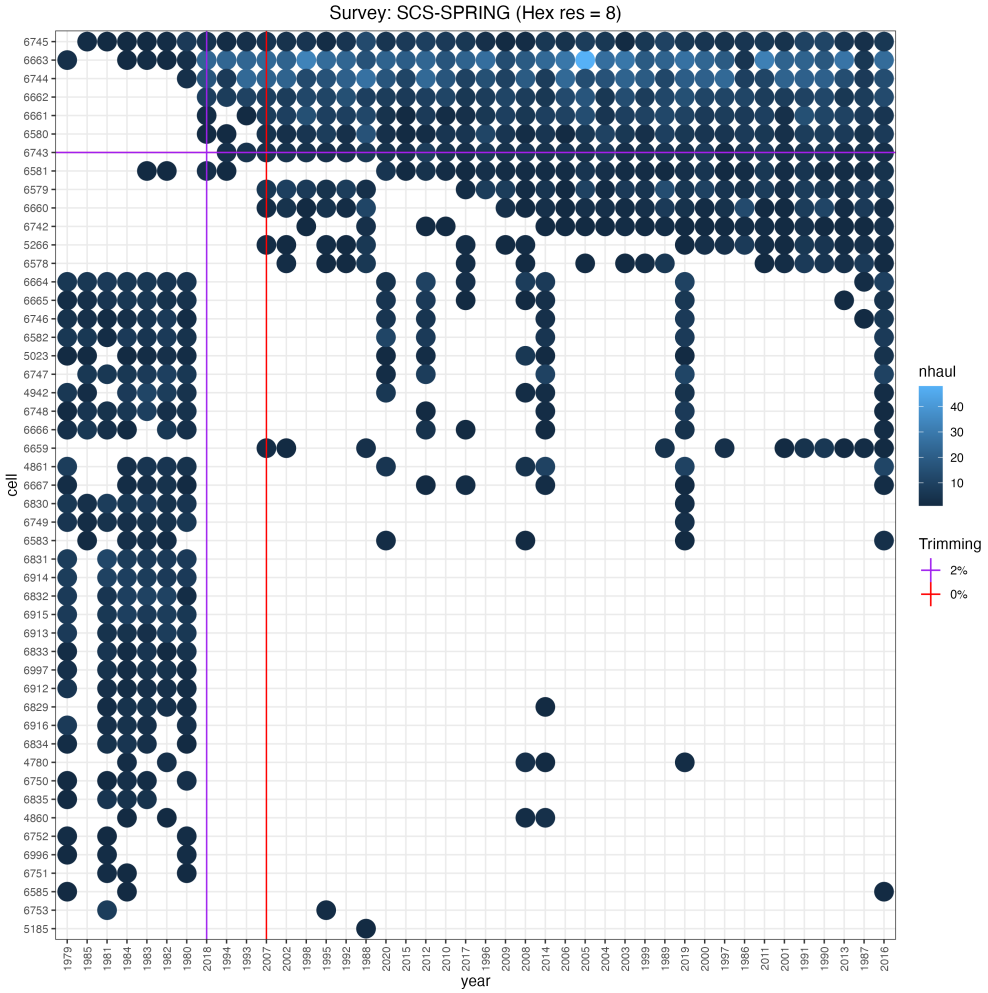
12. Spatio-temporal standardization: SCS-SPRING

a. Standardization method 1

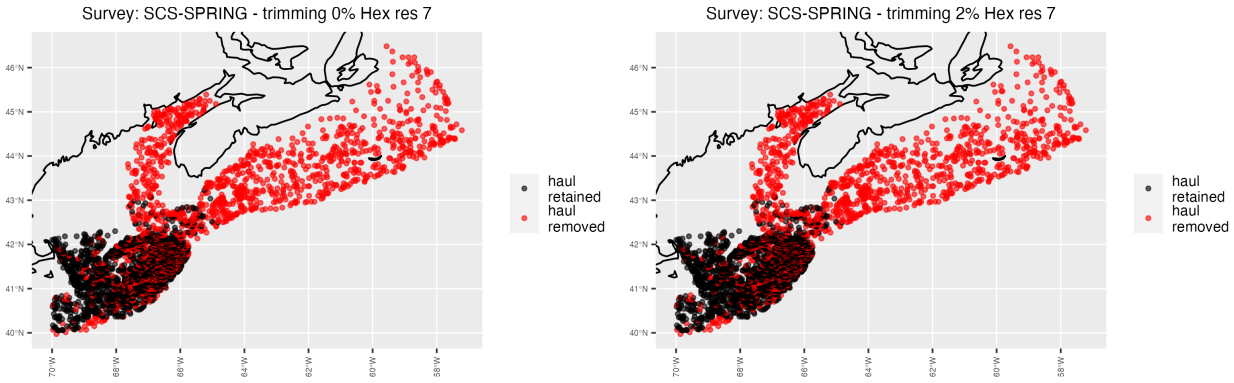
This standardization method was adapted from https://github.com/zoekitchel/trawl_spatial_turnover/blob/master/data_prep_code/species/explore_NorthSea_trimming.Rmd
It was run for hex resolution 7 and 8.

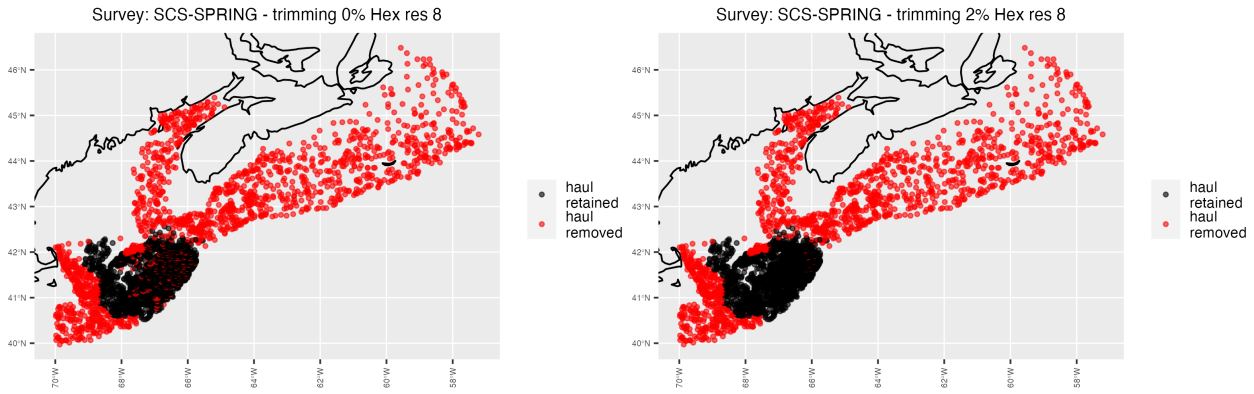
Plot of number of cells x years with overlaid flagging options



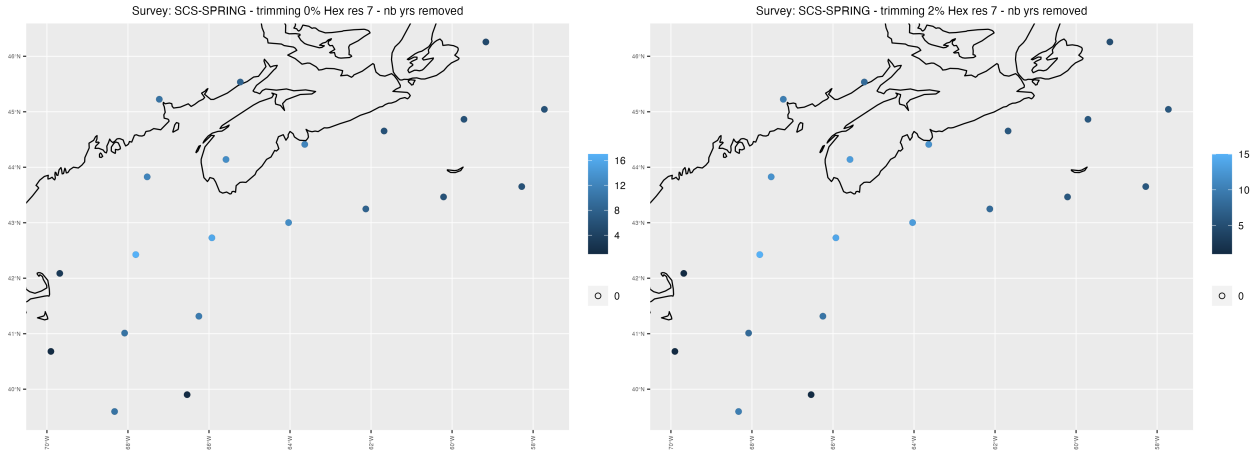


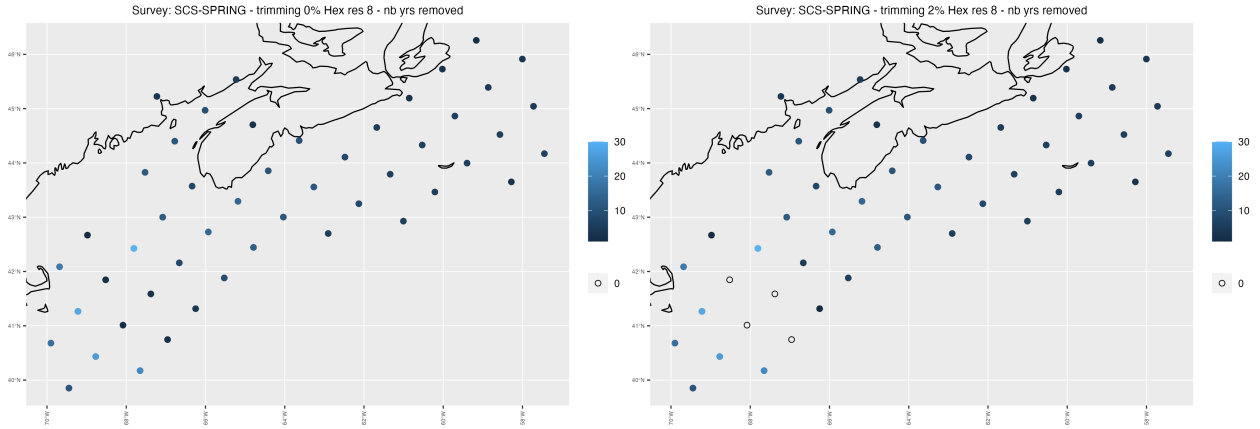
Map of hauls retained and removed per flagging method and threshold





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

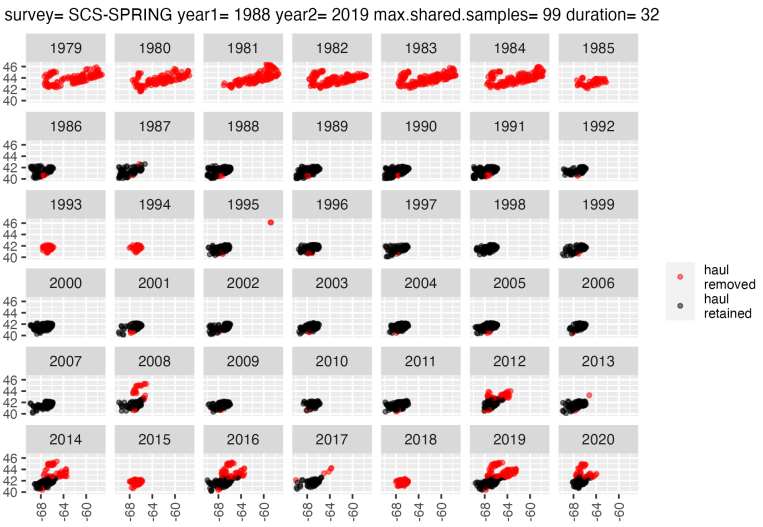




b. Standardization method 2

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	1804.0	1652.0	1786.0	1620.0	13983.0
percentage of hauls removed	43.6	39.9	43.1	39.1	35.2