GMEX: Gulf of Mexico US 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 Mexico (US) bottom trawl survey provided by Jeff Rester, Coordinator - Gulf States Marine Fisheries Commission - Habitat Focus Team - Gulf of Mexico Program & David Hanisko, Research Fisheries Biologist, National Marine Fisheries Service, Southeast Fisheries Science Center. It contains annual data from 1982 and up to 2019.

Data cleaning in R

```
##############################################################################
#### R code to clean trawl survey Gulf of Mexico
#### Public data Ocean Adapt
#### Contacts: Jeff Rester jrester@gsmfc.org Coordinator - Gulf States Marine Fisheries
####                Commission - Habitat Focus Team - Gulf of Mexico Program
####           David Hanisko david.s.hanisko@noaa.gov Research Fisheries Biologist,
####                National Marine Fisheries Service, Southeast Fisheries Science Center
#### Coding: Michelle Stuart, Dan Forrest, Zoë Kitchel November 2021
##############################################################################
#Relevant Organizations
#Gulf States Marine Fisheries Commission: https://www.gsmfc.org/seamap-gomrs.php
```

```
#Southeast Area Monitoring and Assessment Program Reports:
#https://www.fisheries.noaa.gov/southeast/funding-and-financial-services/
#southeast-area-monitoring-and-assessment-program-seamap

#Helpful reference document
#https://sedarweb.org/docs/wpapers/SEDAR7_DW1.pdf
#Many different survey events included in the files we pull in
#Most consistent through time are Summer SEAMAP 1987-on and Fall SEAMAP 1988-on
#All other surveys are excluded in the following code

#-----------------------------------------------------------------------------------------#
#### 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")
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 Mexico 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 ####
#-----------------------------------------------------------------------------------------#

gmex_station_raw <- read_lines(
  "https://github.com/pinskylab/OceanAdapt/raw/master/data_raw/gmex_STAREC.csv")

# remove oddly quoted characters
gmex_station_clean <- str_replace_all(gmex_station_raw, "\\\\\\"", "")
write_lines(gmex_station_clean, file = "gmex_station_raw.txt")

gmex_station <- read_csv(file = "gmex_station_raw.txt",
                         col_types = cols(.default = col_character())) %>%
  #output of new names...49 means The message is telling you that some of
  #the columns have no names and it's giving them one
  select('STATIONID', 'CRUISEID', 'CRUISE_NO', 'P_STA_NO', 'TIME_ZN',
         'TIME_MIL', 'S_LATD', 'S_LATM', 'S_LOND', 'S_LONM', 'E_LATD',
         'E_LATM', 'E_LOND', 'E_LONM', 'DEPTH_SSTA', 'MO_DAY_YR',
         'VESSEL_SPD', 'COMSTAT','TEMP_SSURF','TEMP_BOT')
```

```
#delete this file we temporarily made
file.remove("gmex_station_raw.txt")

problems <- problems(gmex_station) %>%
  filter(!is.na(col))

stopifnot(nrow(problems) == 0)

gmex_station <- type_convert(gmex_station, col_types = cols(
  STATIONID = col_integer(),
  CRUISEID = col_integer(),
  CRUISE_NO = col_integer(),
  P_STA_NO = col_integer(),
  TIME_ZN = col_integer(),
  TIME_MIL = col_character(),
  S_LATD = col_integer(),
  S_LATM = col_double(),
  S_LOND = col_integer(),
  S_LONM = col_double(),
  E_LATD = col_integer(),
  E_LATM = col_double(),
  E_LOND = col_integer(),
  E_LONM = col_double(),
  DEPTH_SSTA = col_double(),
  MO_DAY_YR = col_date(format = ""),
  VESSEL_SPD = col_double(),
  COMSTAT = col_character()
))


gmex_tow <-read_csv(
  "https://github.com/pinskylab/OceanAdapt/raw/master/data_raw/gmex_INVREC.csv",
  col_types = cols(
  INVRECID = col_integer(),
  STATIONID = col_integer(),
  CRUISEID = col_integer(),
  VESSEL = col_integer(),
  CRUISE_NO = col_integer(),
  P_STA_NO = col_integer(),
  GEAR_SIZE = col_integer(),
  GEAR_TYPE = col_character(),
  MESH_SIZE = col_double(),
  OP = col_character(),
  MIN_FISH = col_integer(),
  WBCOLOR = col_character(),
  BOT_TYPE = col_character(),
  BOT_REG = col_character(),
  TOT_LIVE = col_double(),
  FIN_CATCH = col_double(),
  CRUS_CATCH = col_double(),
  OTHR_CATCH = col_double(),
  T_SAMPLEWT = col_double(),
  T_SELECTWT = col_double(),
```

```
  FIN_SMP_WT = col_double(),
  FIN_SEL_WT = col_double(),
  CRU_SMP_WT = col_double(),
  CRU_SEL_WT = col_double(),
  OTH_SMP_WT = col_double(),
  OTH_SEL_WT = col_double(),
  COMBIO = col_character()
))

gmex_tow <- gmex_tow %>%
  select('STATIONID', 'CRUISE_NO', 'P_STA_NO', 'INVRECID', 'GEAR_SIZE',
         'GEAR_TYPE', 'MESH_SIZE', 'MIN_FISH', 'OP') %>%
  filter(GEAR_TYPE=='ST') #ST = shrimp trawl (this is what OceanAdapt does too,
                                                    #preserves 90% of tows)

problems <- problems(gmex_tow) %>%
  filter(!is.na(col))
stopifnot(nrow(problems) == 0)

gmex_spp <-read_csv(
  "https://github.com/pinskylab/OceanAdapt/raw/master/data_raw/gmex_NEWBIOCODESBIG.csv",
  col_types = cols(
  Key1 = col_integer(),
  TAXONOMIC = col_character(),
  CODE = col_integer(),
  TAXONSIZECODE = col_character(),
  isactive = col_integer(),
  common_name = col_character(),
  tsn = col_integer(),
  tsn_accepted = col_integer()
)) %>%
  select(-tsn_accepted)

# problems should be 0 obs
problems <- problems(gmex_spp) %>%
  filter(!is.na(col))
stopifnot(nrow(problems) == 0)

gmex_cruise <-read_csv(
  "https://github.com/pinskylab/OceanAdapt/raw/master/data_raw/gmex_CRUISES.csv",
  col_types = cols(.default = col_character())) %>%
  select(CRUISEID, VESSEL, TITLE, SOURCE)

# problems should be 0 obs
problems <- problems(gmex_cruise) %>%
  filter(!is.na(col))
stopifnot(nrow(problems) == 0)

gmex_cruise <- type_convert(gmex_cruise,
                            col_types = cols(
                              CRUISEID = col_integer(),
                              VESSEL = col_integer(),
                              TITLE = col_character()))
```

```
temp <- tempfile()
download.file(
  "https://github.com/pinskylab/OceanAdapt/raw/master/data_raw/gmex_BGSREC.csv.zip", temp)
gmex_bio <- read.csv(unz(temp, "gmex_BGSREC.csv")) %>%
  select('CRUISEID', 'STATIONID', 'VESSEL', 'CRUISE_NO', 'P_STA_NO',
         'GENUS_BGS','CNT','CNTEXP', 'SPEC_BGS', 'BGSCODE', 'BIO_BGS', 'SELECT_BGS') %>%
  # trim out young of year records (only useful for count data) and those with
  #UNKNOWN species
  filter(BGSCODE != "T" | is.na(BGSCODE),
         GENUS_BGS != "UNKNOWN" | is.na(GENUS_BGS))  %>%
  # remove the few rows that are still duplicates
  distinct()

# problems should be 0 obs
problems <- problems(gmex_bio) %>%
  filter(!is.na(col))
stopifnot(nrow(problems) == 0)

gmex_bio <- type_convert(gmex_bio, cols(
  CRUISEID = col_integer(),
  STATIONID = col_integer(),
  VESSEL = col_integer(),
  CRUISE_NO = col_integer(),
  P_STA_NO = col_integer(),
  GENUS_BGS = col_character(),
  SPEC_BGS = col_character(),
  BGSCODE = col_character(),
  BIO_BGS = col_integer(),
  SELECT_BGS = col_double()
))


# make two combined records where 2 different species share the same species code
newspp <- tibble(
  Key1 = c(503,5770),
  TAXONOMIC = c('ANTHIAS TENUIS AND WOODSI', 'MOLLUSCA AND UNID.OTHER #01'),
  CODE = c(170026003, 300000000),
  TAXONSIZECODE = NA,
  isactive = -1,
  common_name = c('threadnose and swallowtail bass', 'molluscs or unknown'),
  tsn = NA)

# remove the duplicates that were just combined
gmex_spp <- gmex_spp %>%
  distinct(CODE, .keep_all = T)

# add the combined records on to the end. trim out extra columns from gmexspp
gmex_spp <- rbind(gmex_spp[,1:7], newspp) %>%
  select(CODE, TAXONOMIC) %>%
  rename(BIO_BGS = CODE)

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

```r
#-------------------------------------------------------------------------------#

# merge tow information with catch data, but only for shrimp trawl tows (ST)
gmex <- left_join(gmex_bio, gmex_tow, by = c("STATIONID", "CRUISE_NO", "P_STA_NO")) %>%
  # add station location and related data
  left_join(gmex_station, by = c("CRUISEID", "STATIONID", "CRUISE_NO", "P_STA_NO")) %>%
  # add scientific name
  left_join(gmex_spp, by = "BIO_BGS") %>%
  # add cruise title
  left_join(gmex_cruise, by = c("CRUISEID", "VESSEL"))


gmex <- gmex %>%
  # Trim to high quality SEAMAP summer trawls (1987-on) and SEAMAP fall trawls (1988-on)
  #based off the subset used by Jeff Rester's GS_TRAWL_05232011.sas, but including fall
  #Keeps Fall SEAMAP groundfish Survey,Fall SEAMAP Groundfish Survey,
  #Fall SEAMAP Groundfish Suvey, Summer SEAMAP Groundfish Survey, Summer SEAMAP
  #Groundfish Suvey
  filter(
    ((grepl("Summer", TITLE) & year(as.Date(MO_DAY_YR)) >= 1987) |
       (grepl("Fall", TITLE) & year(as.Date(MO_DAY_YR)) >= 1988)) &
           GEAR_SIZE == 40 &
           MESH_SIZE == 1.63 &
           # OP has no letter value
           !grepl("[A-Z]", OP)) %>%
  mutate(
    # Create a unique haulid
    haulid = paste(formatC(VESSEL, width=3, flag=0), formatC(CRUISE_NO, width=3, flag=0),
                   formatC(P_STA_NO, width=5, flag=0, format='d'), S_LATD, S_LOND,sep=''),
    # Extract year where needed
    year = year(MO_DAY_YR),
    month = month(MO_DAY_YR),
    day = day(MO_DAY_YR),
    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 = ifelse(
      grepl("Summer", TITLE), "Summer",
       ifelse(grepl("Fall", TITLE), "Fall", NA

      )),
    # Calculate decimal lat and lon, depth in m, where needed
    S_LATD = ifelse(S_LATD == 0, NA, S_LATD),
    S_LOND = ifelse(S_LOND == 0, NA, S_LOND),
    E_LATD = ifelse(E_LATD == 0, NA, E_LATD),
    E_LOND = ifelse(E_LOND == 0, NA, E_LOND),
    latitude = rowMeans(cbind(S_LATD + S_LATM/60, E_LATD + E_LATM/60), na.rm=T),
    longitude = -rowMeans(cbind(S_LOND + S_LONM/60, E_LOND + E_LONM/60), na.rm=T),
    # convert fathoms to meters
    depth = DEPTH_SSTA * 1.8288,
    # Add "strata" (define by lat, lon and depth bands) where needed
    # degree bins, # degree bins, # 100 m bins
```

```
    stratum = paste(floor(latitude)+0.5, floor(longitude)+0.5,
                    floor(depth/100)*100 + 50, sep= "-")
  )

# fix speed
# Trim out or fix speed and duration records
# trim out tows of 0, >60, or unknown minutes
gmex <- gmex %>%
  filter(MIN_FISH <= 60 & MIN_FISH > 0 & !is.na(MIN_FISH)) %>%
  # fix typo according to Jeff Rester: 30 = 3
  mutate(VESSEL_SPD = ifelse(VESSEL_SPD == 30, 3, VESSEL_SPD)) %>%
  # trim out vessel speeds 0, unknown, or >5 (need vessel speed to calculate area trawled)
  filter(VESSEL_SPD <= 5 & VESSEL_SPD > 0  & !is.na(VESSEL_SPD))

# while comsat (text comment field) is still present
# Remove a tow when paired tows exist (same lat/lon/year but different haulid,
#only Gulf of Mexico)
# identify duplicate tows at same year/lat/lon
dups <- gmex %>%
  group_by(year, latitude, longitude, season) %>% ###add season here if necessary
  filter(n() > 1) %>%
  group_by(haulid) %>%
  filter(n() == 1)

# remove the identified tows from the dataset
gmex <- gmex %>%
  filter(!haulid %in% dups$haulid & !grepl("PORT", COMSTAT))

#sum wtcpue for duplicates (all columns the same except for BGSID which
#we don't pull in, and doesn't  have any significance other than telling us that
#these are indeed independent observations. we're not sure why this occurs
#in the raw data files, but it was the recommended technique by Jeff in 2012)

gmex <- gmex %>%
  group_by(haulid, stratum, year, latitude, longitude, depth, BIO_BGS, SOURCE,
           MIN_FISH, GEAR_TYPE, SPEC_BGS,GENUS_BGS,
           STATIONID, TEMP_BOT, TEMP_SSURF, TAXONOMIC, VESSEL_SPD, GEAR_SIZE,
           month, day, quarter, season) %>%
   summarise(SELECT_BGS = sum(SELECT_BGS, na.rm = T), #sum weights across duplicates
             CNTEXP = sum(CNTEXP, na.rm = T)) #sum counts across duplicates

gmex <- gmex %>%
  rename(sub_area = SOURCE,
         haul_dur.min = MIN_FISH,
         gear = GEAR_TYPE,
         haul_id = haulid,
         station = STATIONID,
         sbt = TEMP_BOT,
         sst = TEMP_SSURF,
         verbatim_name = TAXONOMIC,
         num = CNTEXP,
         wgt = SELECT_BGS
         ) %>%
```

```
  mutate(
  # adjust for area towed
# kg per 1,000,000m2 (1km^2). calc area trawled in m2:
    #     knots * 1.8 km/hr/knot * 1000 m/km * minutes *
    #       1 hr/60 min * width of gear in feet * 0.3 m/ft # biomass per standard tow
    wgt_cpue = 1000000*wgt/
      (VESSEL_SPD * 1.85200 * 1000 * haul_dur.min / 60 * GEAR_SIZE * 0.3048),
# count per 1,000,000m2 (1km^2). calc area trawled in m2: knots *
#                  1.8 km/hr/knot * 1000 m/km * minutes * 1 hr/60 min *
#                  width of gear in feet * 0.3 m/ft # biomass per standard tow
    num_cpue = 1000000*num/
      (VESSEL_SPD * 1.85200 * 1000 * haul_dur.min / 60 * GEAR_SIZE * 0.3048),
   #area_swept in km^2: knots * 1.8 km/hr/knot * minutes *
#                  1 hr/60 min * width of gear in feet * 0.0003 km/ft
   area_swept = VESSEL_SPD * 1.85200 * haul_dur.min / 60 * GEAR_SIZE * 0.0003048,
   #kg per hour: 60 minutes/hour * kg / minutes fished
   wgt_h = 60*wgt/haul_dur.min,
   #count per hour:  60 minutes/hour * abundance/minutes fished
   num_h  = 60*num/haul_dur.min
  ) %>%
  # remove non-fish
  filter(
    verbatim_name != '' | !is.na(verbatim_name),
    # remove unidentified verbatim_name
    !verbatim_name %in%
      c('UNID CRUSTA', 'UNID OTHER', 'UNID.FISH',
        'CRUSTACEA(INFRAORDER) BRACHYURA', 'MOLLUSCA AND UNID.OTHER #01',
        'ALGAE', 'MISCELLANEOUS INVERTEBR', 'OTHER INVERTEBRATES')
  ) %>%
  mutate(
  # adjust verbatim_name names
    verbatim_name = ifelse(GENUS_BGS == 'PELAGIA' &
                             SPEC_BGS == 'NOCTUL', 'PELAGIA NOCTILUCA', verbatim_name),
    BIO_BGS = ifelse(verbatim_name == "PELAGIA NOCTILUCA", 618030201, BIO_BGS),
    verbatim_name = ifelse(GENUS_BGS == 'MURICAN' &
                             SPEC_BGS == 'FULVEN', 'MURICANTHUS FULVESCENS', verbatim_name),
    BIO_BGS = ifelse(verbatim_name == "MURICANTHUS FULVESCENS", 308011501, BIO_BGS),
    verbatim_name = ifelse(grepl("APLYSIA", verbatim_name), "APLYSIA", verbatim_name),
    verbatim_name = ifelse(grepl("AURELIA", verbatim_name), "AURELIA", verbatim_name),
    verbatim_name = ifelse(grepl("BOTHUS", verbatim_name), "BOTHUS", verbatim_name),
    verbatim_name = ifelse(grepl(
      "CLYPEASTER", verbatim_name), "CLYPEASTER", verbatim_name),
    verbatim_name = ifelse(grepl("CONUS", verbatim_name), "CONUS", verbatim_name),
    verbatim_name = ifelse(grepl("CYNOSCION", verbatim_name), "CYNOSCION", verbatim_name),
    verbatim_name = ifelse(grepl(
      "ECHINASTER", verbatim_name), "ECHINASTER", verbatim_name),
    verbatim_name = ifelse(grepl(
      "OPISTOGNATHUS", verbatim_name), "OPISTOGNATHUS", verbatim_name),
    verbatim_name = ifelse(grepl(
      "OPSANUS", verbatim_name), "OPSANUS", verbatim_name),
    verbatim_name = ifelse(grepl(
      "ROSSIA", verbatim_name), "ROSSIA", verbatim_name),
    verbatim_name = ifelse(grepl(
```

```
      "SOLENOCERA", verbatim_name), "SOLENOCERA", verbatim_name),
    verbatim_name = ifelse(grepl(
      "TRACHYPENEUS", verbatim_name), "TRACHYPENEUS", verbatim_name)
  ) %>%
  # add survey column
  mutate(survey = "GMEX",
         country = "United States",
         continent = "n_america",
         stat_rec = NA,
         verbatim_aphia_id = NA,
         #haul duration in hours is haul duration minutes * 1 hour/60 minutes
         haul_dur = haul_dur.min/60
  ) %>%
  ungroup() %>%
  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_aphia_id)


#-------------------------------------------------------------------------------------#
#### 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
gmex_survey_code <- "GMEX"

gmex <- gmex %>%
  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(gmex$taxa2),
                         input_survey = gmex_survey_code,
                         save = F, output=NA, fishbase=T) # takes 10 mins

#Portunus spinimanus                                    no match
#Trachypeneus                                           no match
#Portunus spinicarpus                                   no match
#Podochela sidneyi                                      no match
#Parthenope granulata                                   no match
#Stenocionops furcata                                   no match
#Ventricolaria                                          no match
#Astroscopus y-graecum                                  no match      ##fish
```

```
#Actaea rufopunctata                                    no match
#Coelenterata                                           no match
#Ventricolaria                                          no match
#Panopeus bermudensis                                   no match
#Photichthyidae                                         no match
#Parthenope punctata                                    no match
#Enidae                                                 no match
#Unionidae                                              no match
#Iliacantha intermedia                                  no match
#Abisa                                                  no match
#Mithrax acuticornis                                    no match
#Portunus floridanus                                    no match
#Nereidae                                               no match
#Macrobrachium ohione                                   no match
#Parthenope fraterculus                                 no match
#Pinnotheres maculatum                                  no match
#Portunus ordwayi                                       no match
#Podochela gracilipes                                   no match
#Portunus depressifrons                                 no match
#Mithrax forceps                                        no match
#Mellita sexiesperforata                                no match
#Hypoconcha sabulosa                                    no match
#Calappa angusta                                        no match
#Processa tenuipes                                      no match
#Cyclois bairdii                                        no match
#Pecten tereinus                                        no match
#Lophopanopeus distinctus                               no match
#Helix                                                  no match
#Cheiraster echinulatus                                 no match
#Corillidae                                             no match
#Chirostylus spinifer                                   no match
#Pisidiidae                                             no match
#Pomacea                                                no match


#all invertebrates except for Astroscopus y-graecum

#add new row for this species
ast_ygr <- c("Astroscopus y-graecum", 159252,3704,
             "Astroscopus y-graecum","Animalia","Chordata",
             "Actinopteri","Perciformes","Uranoscopidae","Astroscopus","Species","GMEX")

clean_auto.missing <- rbind(clean_auto, ast_ygr)

#-------------------------------------------------------------------------------------------#
#### INTEGRATE CLEAN TAXA in GMEX survey data ####
#-------------------------------------------------------------------------------------------#
clean_taxa <- clean_auto.missing %>%
  select(-survey)

clean_gmex <- left_join(gmex, clean_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 = "NOAA",
         timestamp = lubridate::my("03/2021"),
         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`)


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

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

#explanations for duplications of haulid x species
  #Etropus crossotus and Etropus intermedius both fix to Etropus crossotus
  #Monacanthus hispidus, Monacanthus setifer, and Stephanolepis hispida
      #all fix to Stephanolepis hispida
  #Ophidion beani and Ophidion holbrooki both fix to Ophidion holbrookii
  #Anthias tenuis and Anthias tenuis and woodsi both fix to Choranthias tenuis
  #Multiple genuses resolve together (Cynoscion, Bothus, Opsanus)

#User decisions with what to do with repeats due to taxonomic classifications
#depend on goals of data use, and therefore are maintained in FishGlob data product


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

# Just run this routine should be good for all
write_clean_data(data = clean_gmex, survey = "GMEX", 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_gmex$survey))

#run flag_spp function in a loop
for (r in regions) {
  flag_spp(clean_gmex, 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_gmex, 7)

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

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


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

```r
survey_units <- sort(unique(clean_gmex$survey_unit))
survey_std <- clean_gmex %>%
  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 = "GMEX_std",
                 overwrite = T, rdata=TRUE)
```

1. Overview of the survey data table

survey	source	timestamp	haul_id	country	sub_area	continent
GMEX	NOAA	2021-03-01	004167454822988	United States	US	n_america
GMEX	NOAA	2021-03-01	004167454822988	United States	US	n_america
GMEX	NOAA	2021-03-01	004167454822988	United States	US	n_america
GMEX	NOAA	2021-03-01	004167454822988	United States	US	n_america
GMEX	NOAA	2021-03-01	004167454822988	United States	US	n_america

stat_rec	station	stratum	year	month	day	quarter	season
NA	53286	29.5–88.5-50	1987	6	11	2	Summer
NA	53286	29.5–88.5-50	1987	6	11	2	Summer
NA	53286	29.5–88.5-50	1987	6	11	2	Summer
NA	53286	29.5–88.5-50	1987	6	11	2	Summer
NA	53286	29.5–88.5-50	1987	6	11	2	Summer

latitude	longitude	haul_dur	area_swept	gear	depth
29.89083	-88.40833	0.9833333	0.0666098	ST	56.87568
29.89083	-88.40833	0.9833333	0.0666098	ST	56.87568
29.89083	-88.40833	0.9833333	0.0666098	ST	56.87568
29.89083	-88.40833	0.9833333	0.0666098	ST	56.87568
29.89083	-88.40833	0.9833333	0.0666098	ST	56.87568

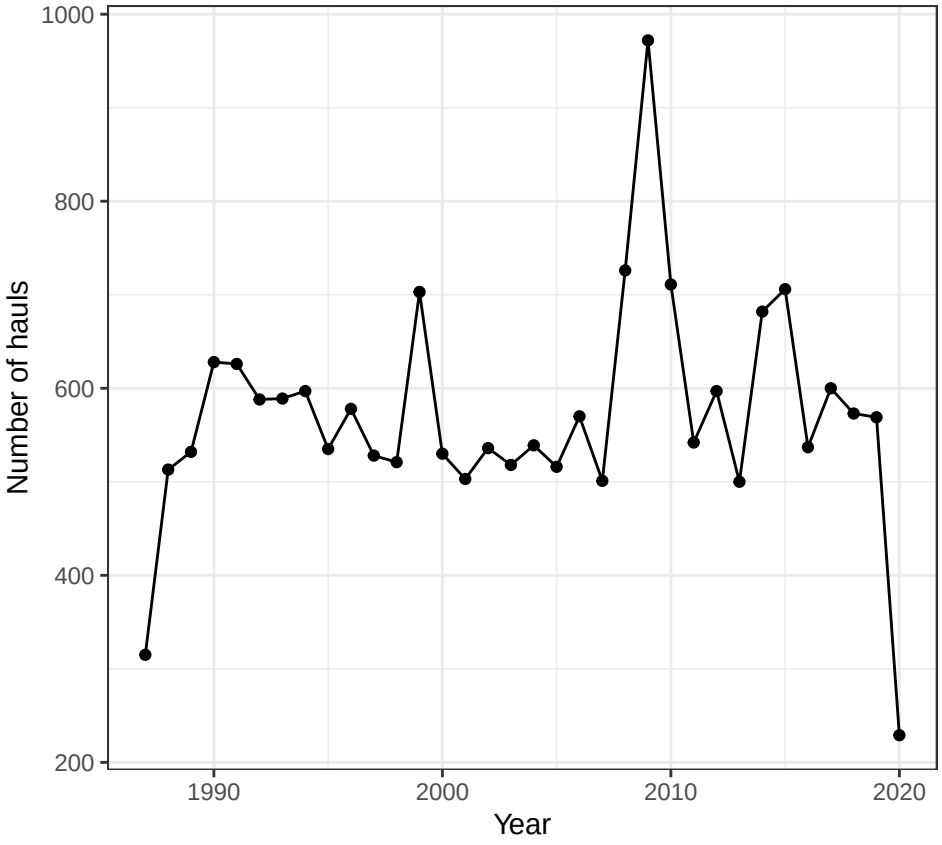
sbt	sst	num	num_cpue	num_cpua	wgt
26.07	28.07	20	20.33898	300.2562	0.358
26.07	28.07	12	12.20339	180.1537	0.447
26.07	28.07	55	55.93220	825.7047	0.447
26.07	28.07	394	400.67797	5915.0479	11.626
26.07	28.07	126	128.13559	1891.6143	5.366

wgt_cpue	wgt_cpua	verbatim_name	verbatim_aphia_id	accepted_name
0.3640678	5.374587	ANCHOA HEPSETUS	NA	Anchoa hepsetus
0.4545763	6.710727	SYNODUS FOETENS	NA	Synodus foetens
0.4545763	6.710727	PRIONOTUS RUBIO	NA	Prionotus rubio
11.8230508	174.538953	DIPLECTRUM BIVITTATUM	NA	Diplectrum bivittatum
5.4569492	80.558750	CENTROPRISTIS PHILADELPHICUS	NA	Centropristis philadelphica

aphia_id	SpecCode	kingdom	phylum	class	order	family
158698	1133	Animalia	Chordata	Teleostei	Clupeiformes	Engraulidae
158758	2719	Animalia	Chordata	Teleostei	Aulopiformes	Synodontidae
276289	4025	Animalia	Chordata	Teleostei	Perciformes	Triglidae
276176	3318	Animalia	Chordata	Teleostei	Perciformes	Serranidae
159347	3317	Animalia	Chordata	Teleostei	Perciformes	Serranidae

2. Summary of sampling intensity

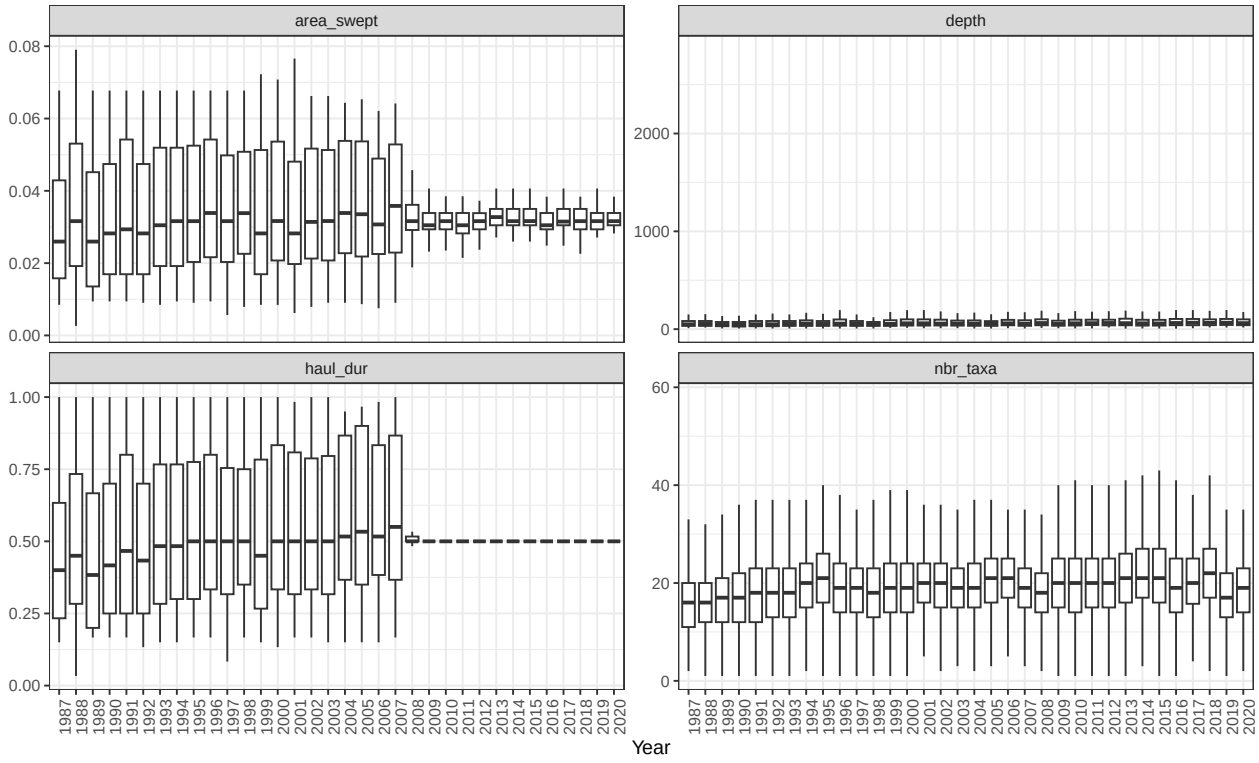
Number of hauls per year performed during the survey after data processing.



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:

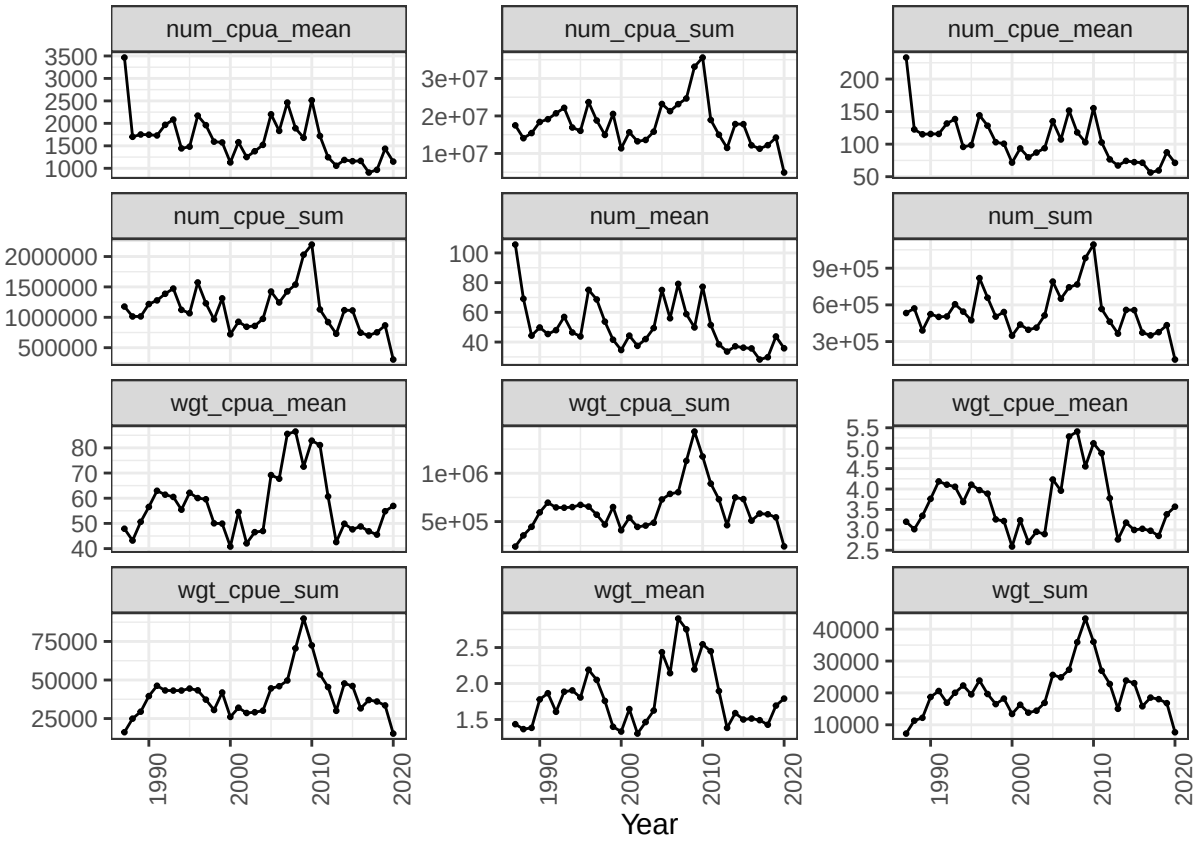
- *area_swept*, swept area by the bottom trawl gear km^2
- *depth*, sampling depth in m
- *haul_dur*, haul sampling duration *hour*
- *number of marine fish taxa*, taxa were cleaned following the last version of taxonomy from the World Register of Marine Species (https://www.marinespecies.org/, October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



4. Summary of biological variables

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

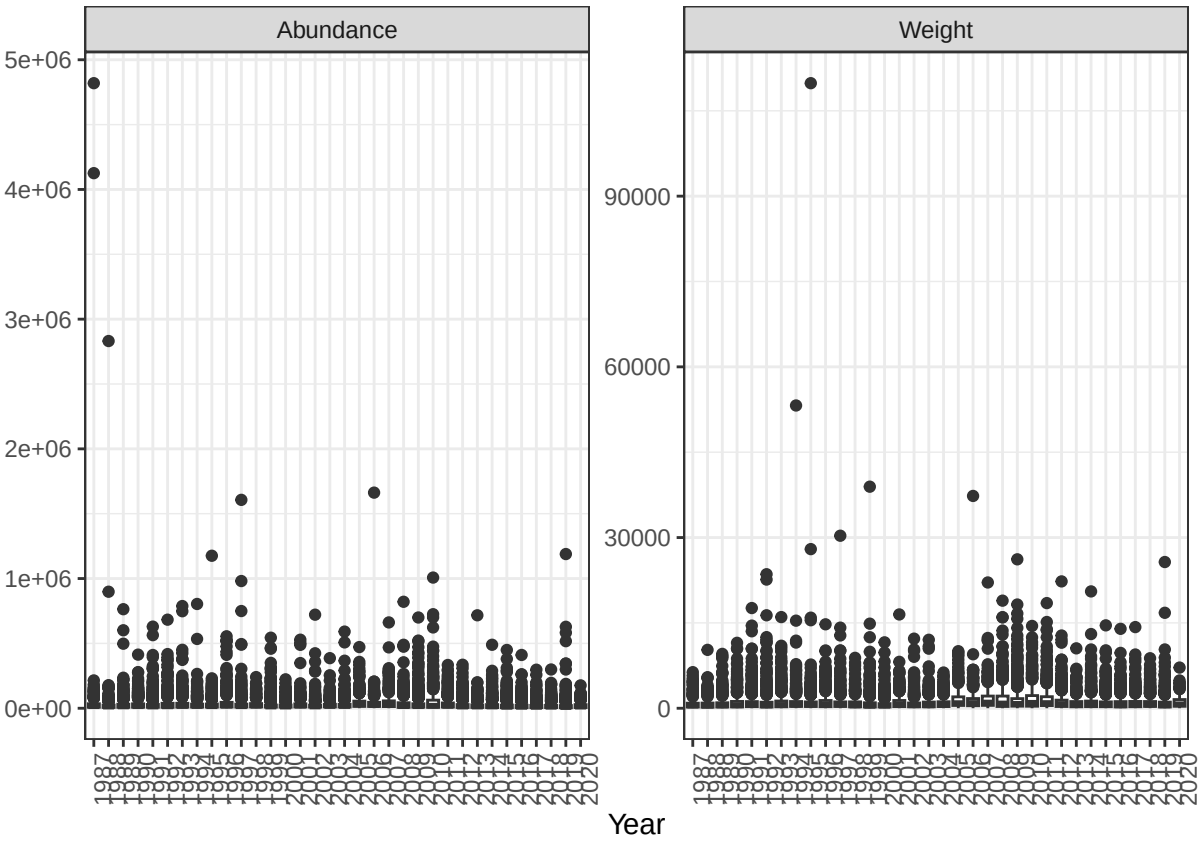
- *num_cpua*, number of individuals (abundance) in $\frac{individuals}{km^2}$
- *num_cpue*, number of individuals (abundance) in $\frac{individuals}{h}$
- *num*, number of individuals (abundance)
- *wgt_cpua*, weight in $\frac{kg}{km^2}$
- *wgt_cpue*, weight in $\frac{kg}{h}$
- *wgt*, weight in kg



5. Extreme values

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

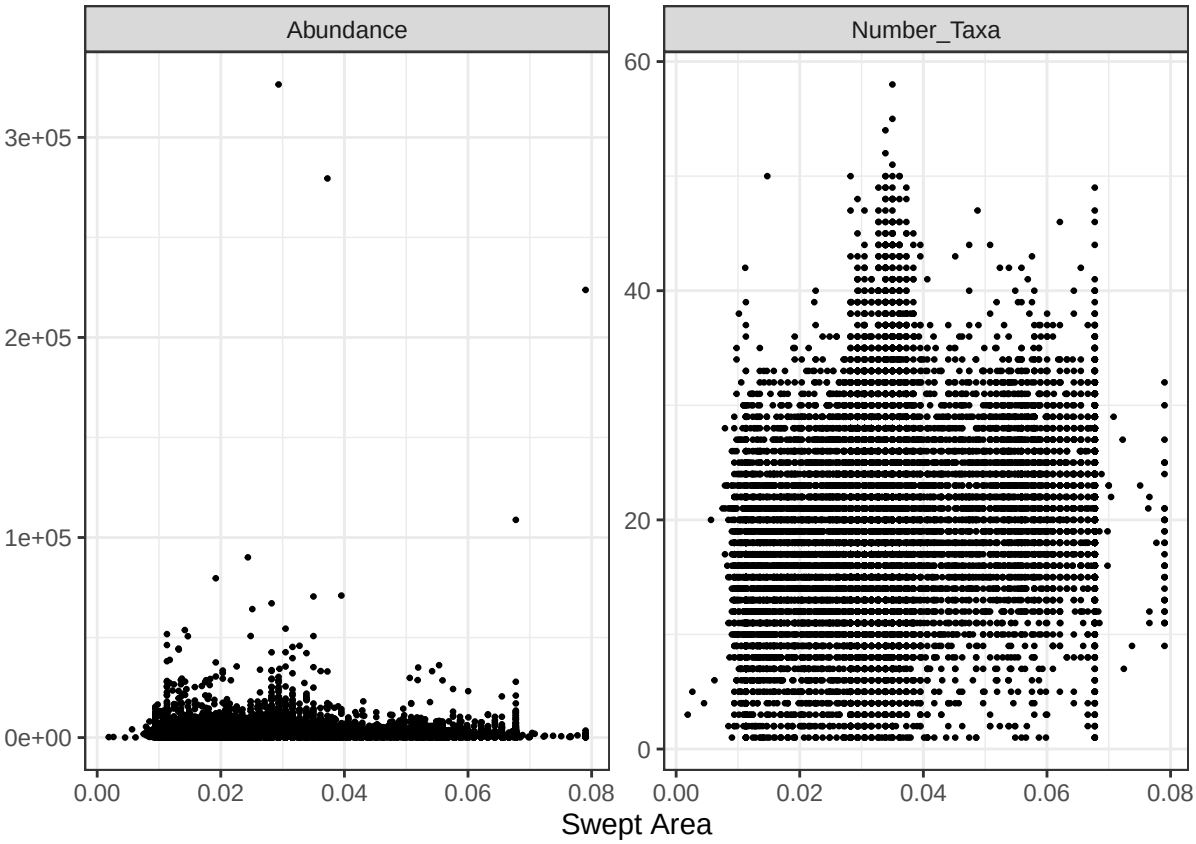
- *num_cpue*, number of individuals (abundance) in $\frac{individuals}{km^2}$
- *wgt_cpue*, weight in $\frac{kg}{km^2}$



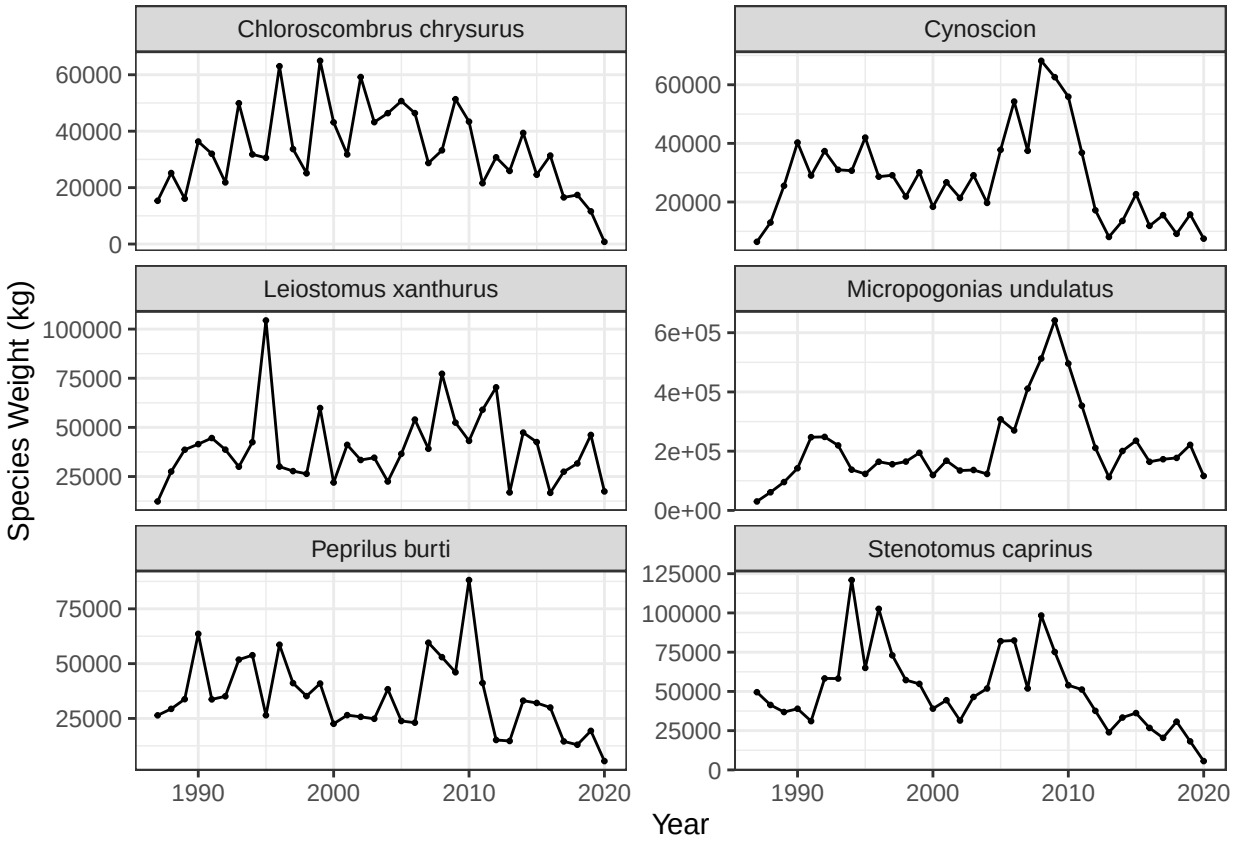
6. Summary of variables against swept area

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

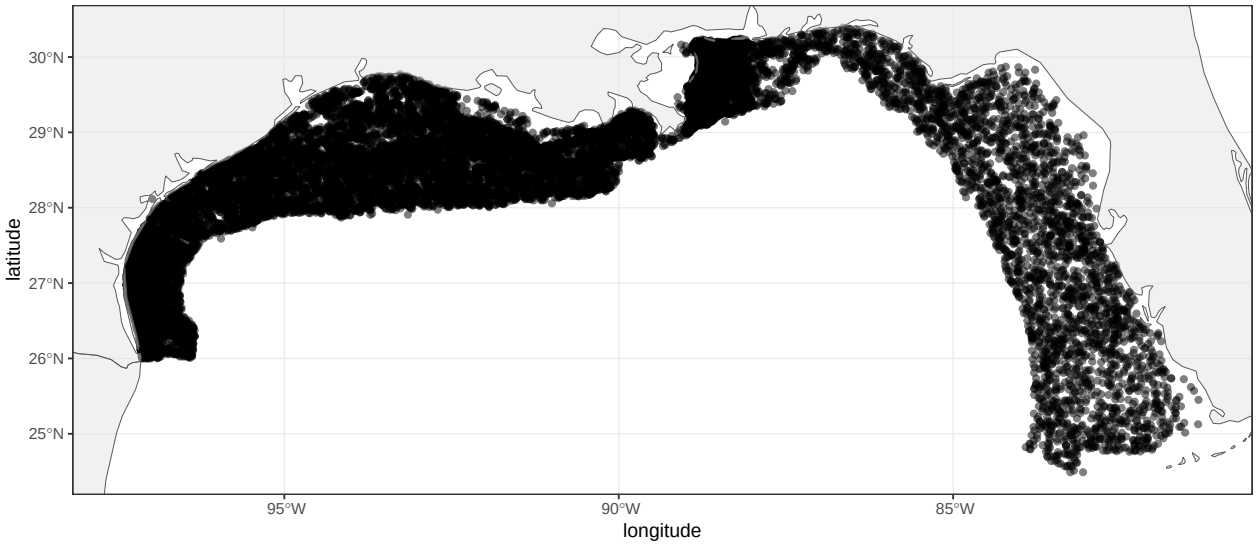


7. Abundance or Weight trends of the six most abundant species



8. Distribution mapping

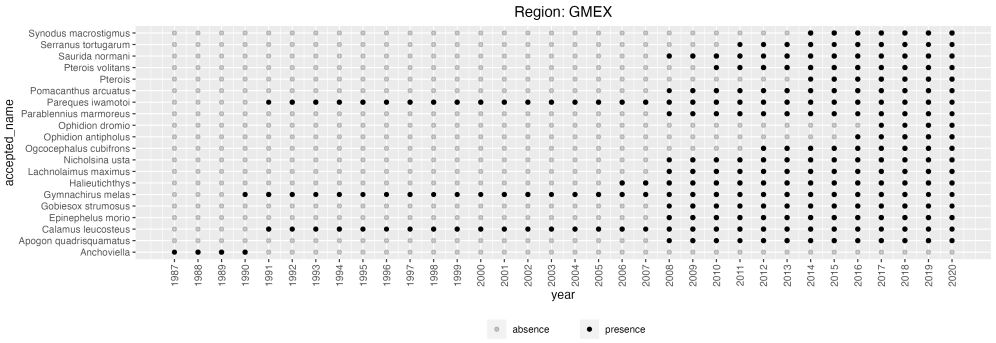
Map of the sampling distribution in space. Note that we only show one year per coordinate.



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

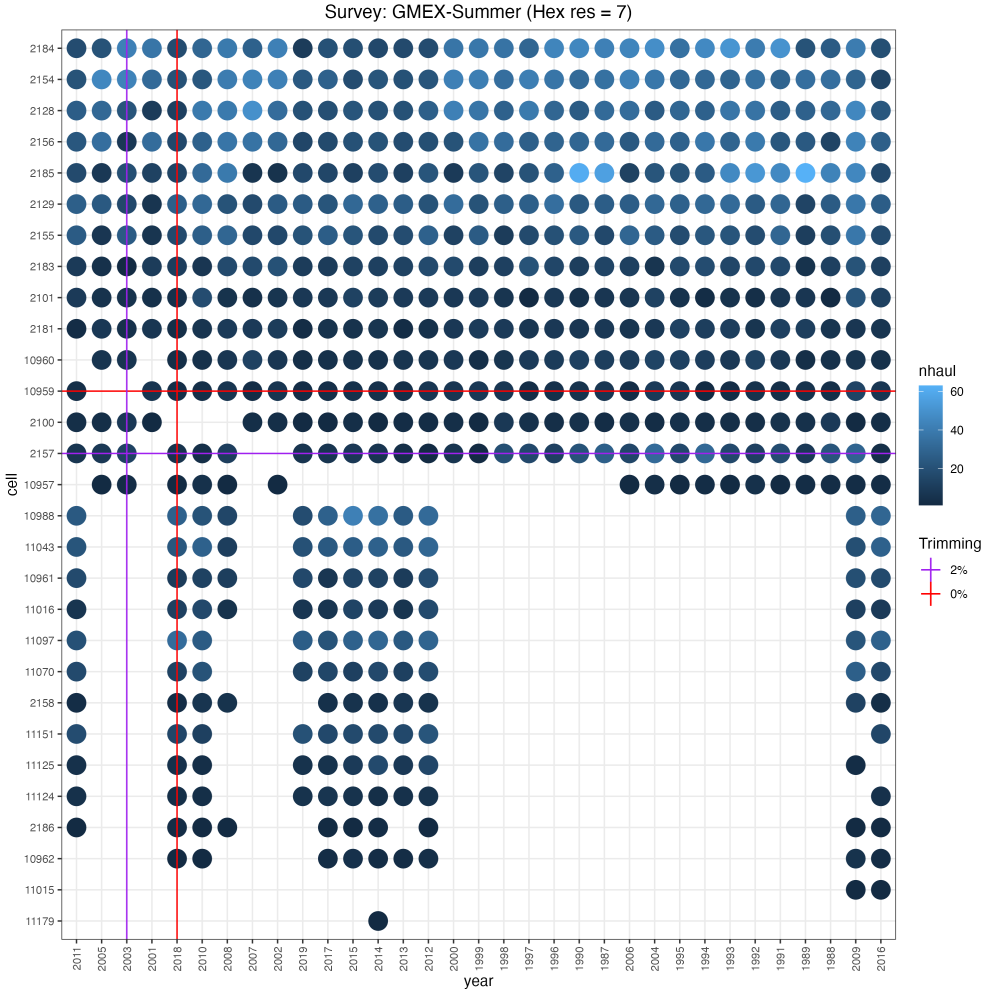
Total number of species	814.0
Percentage of species flagged	2.5

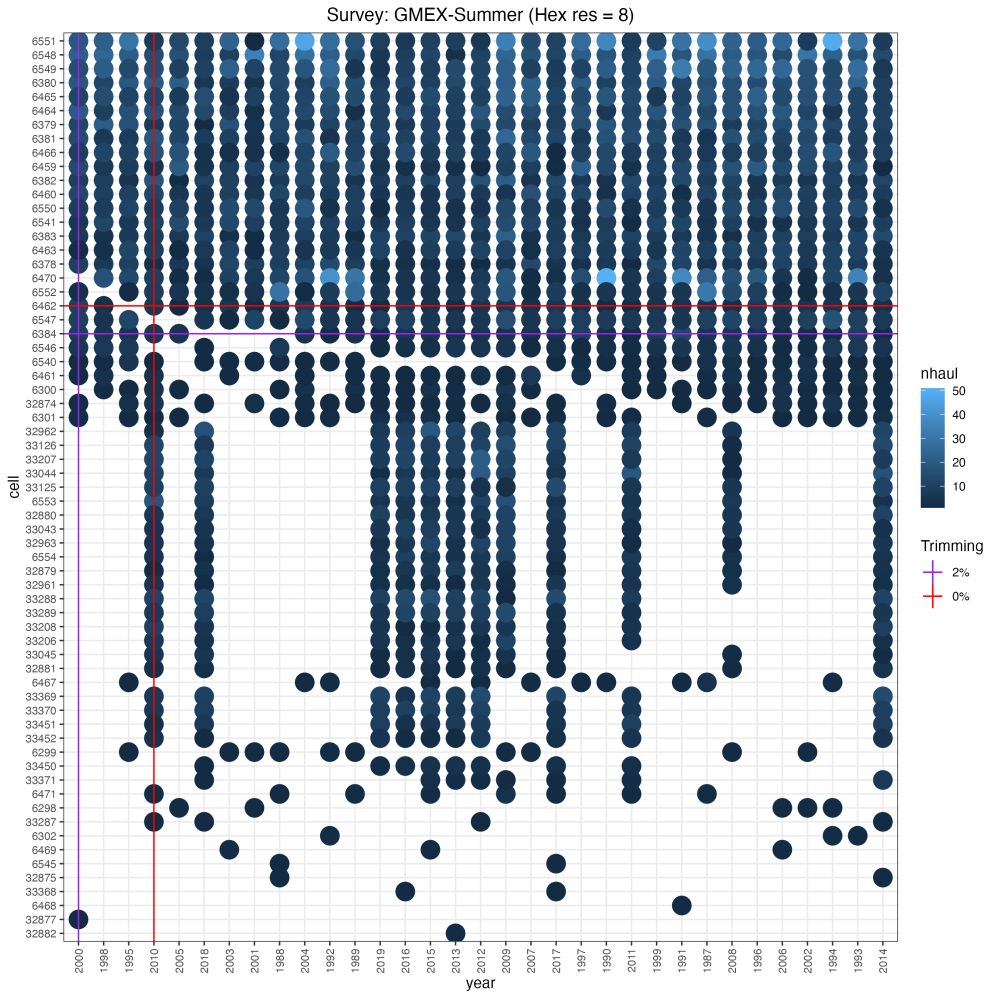
10. Spatio-temporal standardization: GMEX-Summer

a. Standardization method 1

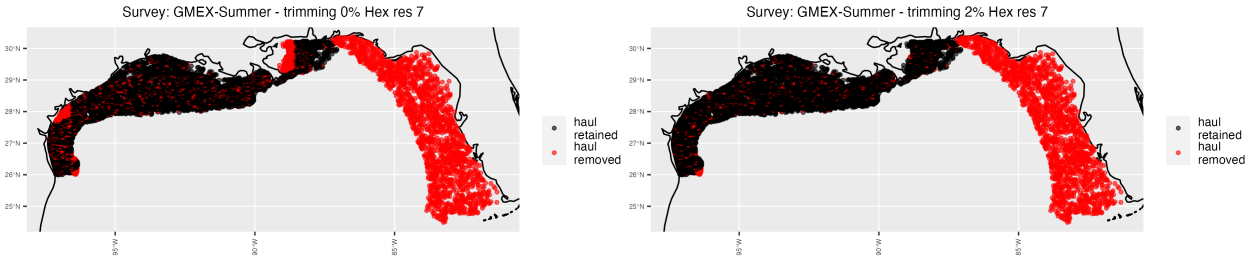
This standardization method was adapted from https://github.com/zoekitchel/trawl_spatial_turnover/blob/master/data_prep_code/species/explore_NorthSea_trimming.Rmd
It was run for hex resolution 7 and 8.

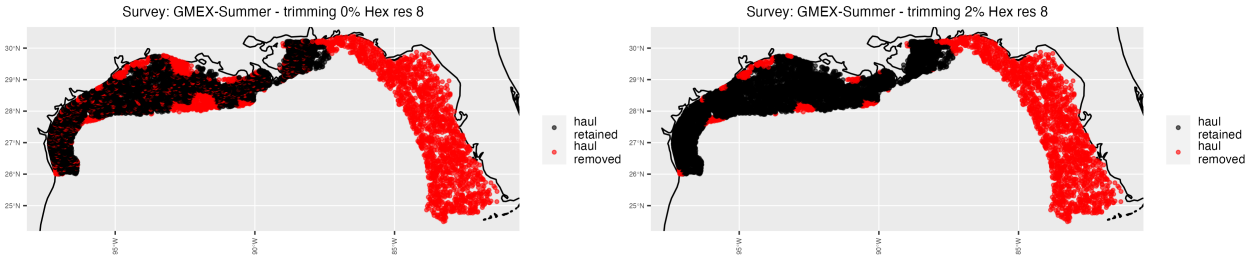
Plot of number of cells x years with overlaid flagging options



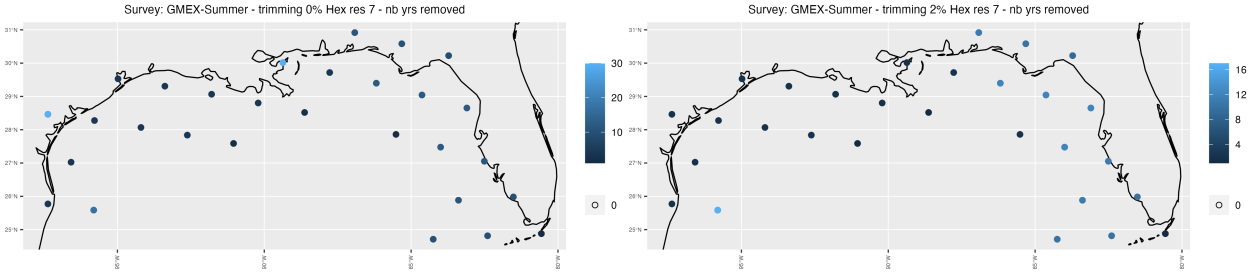


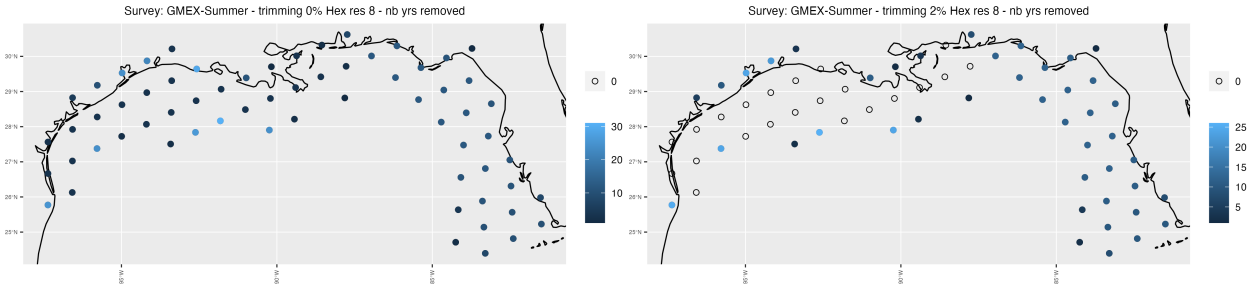
Map of hauls retained and removed per flagging method and threshold





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

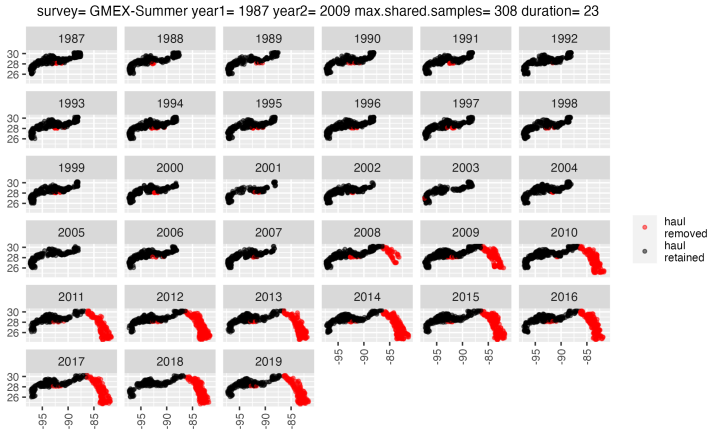




b. Standardization method 2

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

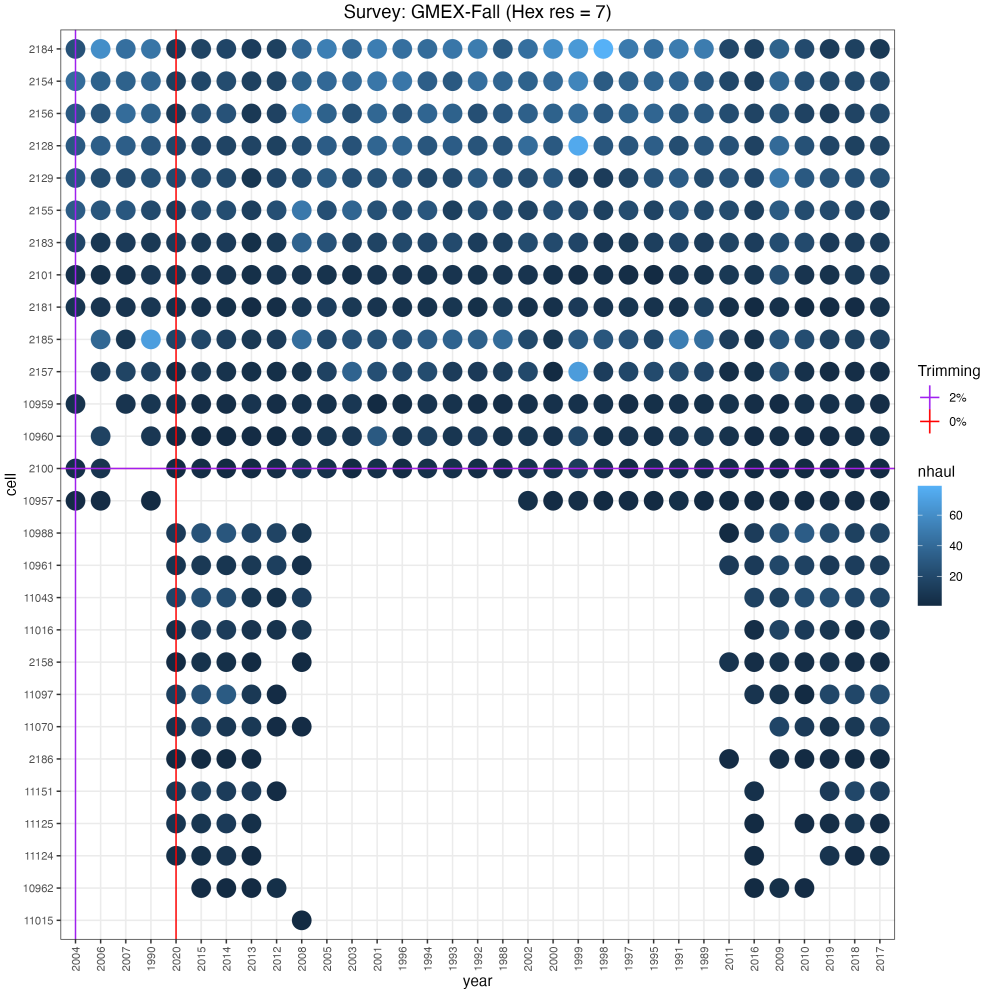
summary	grid cell 7, 0% threshold	grid cell 7, 2% threshold	grid cell 8, 0% threshold	grid cell 8, 2% threshold	method 2 (biotime)
number of hauls removed	2969.0	2157.0	3247.0	2177.0	40791.0
percentage of hauls removed	30.1	21.9	32.9	22.1	22.6

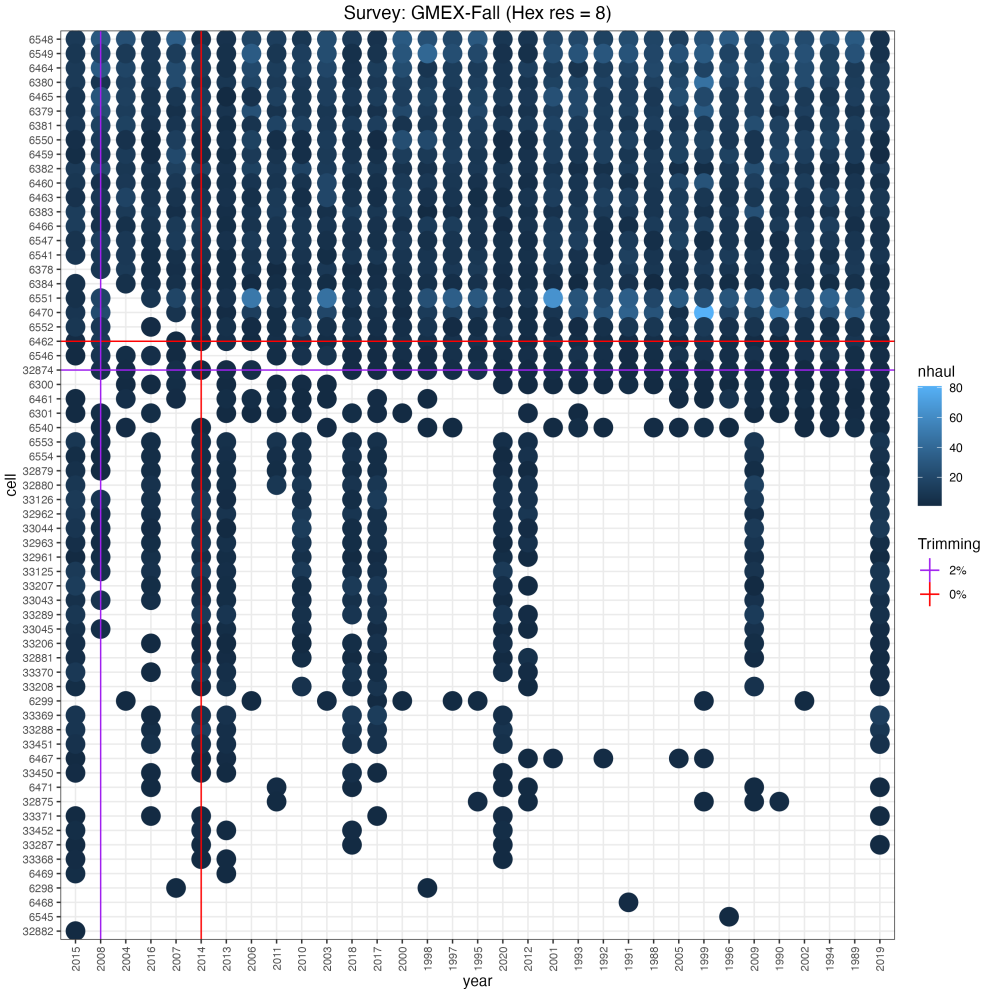
11. Spatio-temporal standardization: GMEX-Fall

a. Standardization method 1

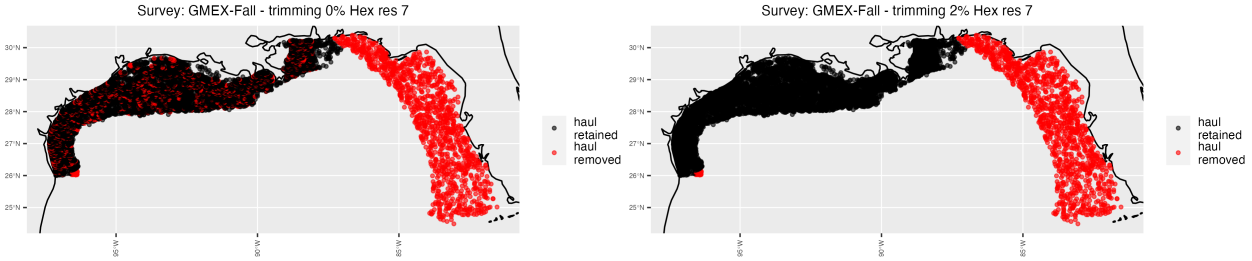
This standardization method was adapted from https://github.com/zoekitchel/trawl_spatial_turnover/blob/master/data_prep_code/species/explore_NorthSea_trimming.Rmd
It was run for hex resolution 7 and 8.

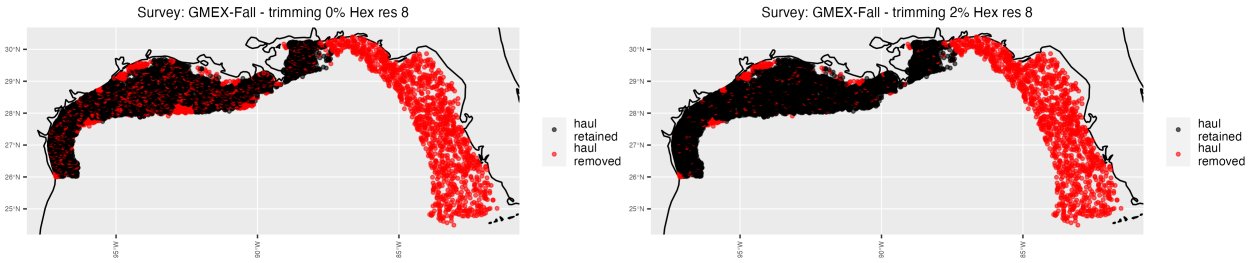
Plot of number of cells x years with overlaid flagging options



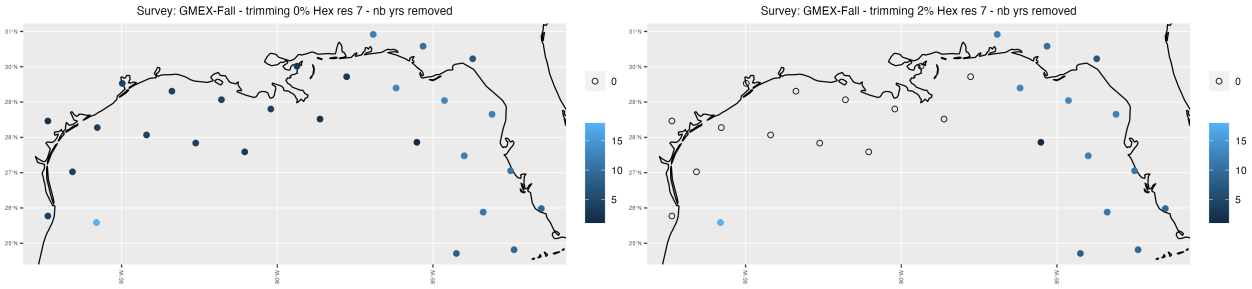


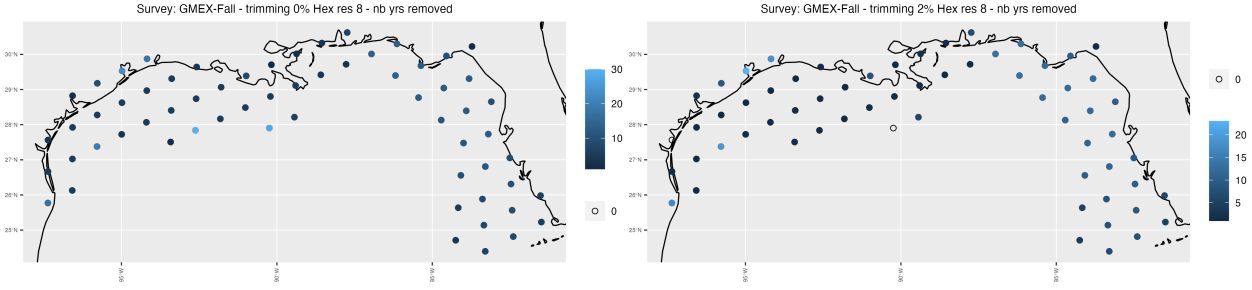
Map of hauls retained and removed per flagging method and threshold





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

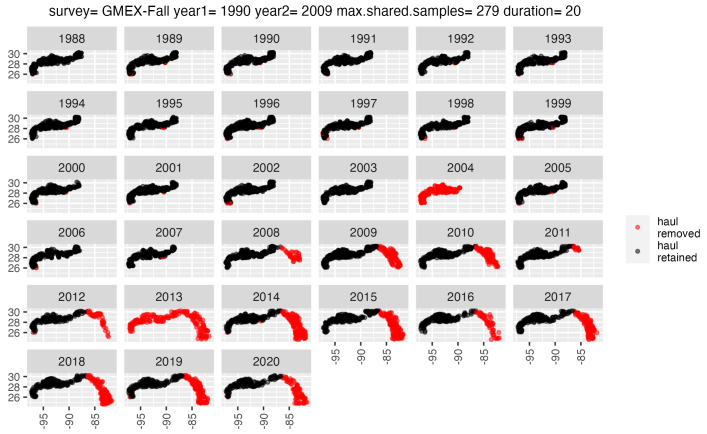




b. Standardization method 2

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	2342.0	1234.0	2675	1587.0	35446.0
percentage of hauls removed	24.5	12.9	28	16.6	18.3