

SAÉ 3-01-EMS Recueil et analyse de données par échantillonnage ou plan d'expérience

TP/TD Echantillonnage stratifié - Données EVHOE

B. Alglave et L. Bellanger

Objectif du TD/TP: Estimer l'abondance de merlus à partir d'un sondage stratifié dans le Golfe de Gascogne et en Mer Celtique.

Dans le cadre de l'études d'une population marine, les chercheurs estiment l'abondance d'une population (e.g. le nombre de merlus dans le Golfe de Gascogne et Mer Celtique en 2018) à partir d'un échantillon récolté de façon stratifié (Cf. théorie de l'estimation).

Pour cela, ils réalisent chaque année des campagnes océanographiques qui permettent de récolter des données et d'obtenir des estimations de l'abondance.

Ces données sont cruciales pour le suivi des espèces marines afin d'évaluer le bon état écologique des populations exploitées. Pour ce TP, nous allons étudier la campagne EVHOE dont un descriptif est donné au lien: <https://campagnes.flotteoceanographique.fr/series/8/fr/>.

Données

Les données EVHOE (Evaluation Halieutique Ouest de l'Europe) sont des données échantillonées chaque année en Octobre/Novembre. Cette campagne cible les espèces benthodémersales du golfe de Gascogne (GdG) et de Mer Celtique (MC). L'échantillonnage est stratifié suivant les classes de profondeur et les grandes unités écologiques du GdG et de MC (voir le shapefile `Agreed_Strata_EVHOE_Polyg_WGS84.shp` et l'objet `evhoe_shp`).

Les poissons sont échantillonnées à l'aide d'un chalut ; ils sont comptés, pesés, sexés pour tout ou partie du trait de chalut. Les données entre 2008 et 2019 sont stockés dans le fichier `EVHOE_2008_2019.RData`. Il est constitué de trois data frame:

- `Save_Datras$datras_HH.full` regroupe les principales informations de chaque trait de chalut (e.g. localisation, période de relevé)
 - Year: année
 - long: longitude
 - lati: latitude
 - StNo: numéro de station
 - HaulNo: numéro du trait de chalut
 - Depth: profondeur
 - Distance: distance parcourue pour un trait de chalut (en mètres). Il y a des NA dans cette colonne (données manquantes). Dans ce cas, on prend la moyenne de la distance des autres traits de chaluts pour remplacer les NA.
- `Save_Datras$datras_sp.HL.full` regroupe le poids et les abondances sur l'ensemble d'un trait de chalut de chaque combinaison 'trait de chalut x espèce x classe de taille x sexe' (données ré-haussées)

- Year: année
 - long: longitude
 - lati: latitude
 - StNo: numéro de station
 - HaulNo: numéro du trait de chalut
 - scientificname: nom scientifique
 - LngtClass: classe de taille
 - TotalNo: comptages (nombre d'individus par combinaison de facteur)
- Save_Datras\$datras_sp.CA.full regroupe les données de mesures individuelles d'un sous-échantillon du trait de chalut. Une ligne correspond à un individu. Ces données regroupent les données individuelles de taille, de poids, de sexe. Nous n'utiliserons pas ces données dans ce projet.

Chargement des données

```
# Charger les données EVHOE et les strates de la campagne
load("data/EVHOE_2008_2019.RData")
evhoe_shp <- st_read("data/STRATES/Agreed_Strata_EVHOE_Polyg_WGS84.shp") %>%
  dplyr::select(STRATE)

## Reading layer 'Agreed_Strata_EVHOE_Polyg_WGS84' from data source
##   '/home/balglave/Desktop/Teaching/EVHOE_data/data/STRATES/Agreed_Strata_EVHOE_Polyg_WGS84.shp'
##   using driver 'ESRI Shapefile'
## Simple feature collection with 29 features and 7 fields
## Geometry type: MULTIPOLYGON
## Dimension:      XY
## Bounding box:  xmin: -11.6487 ymin: 43.61471 xmax: -1.109647 ymax: 52.19023
## Geodetic CRS:  WGS 84

evhoe_shp$area_strata <- as.numeric(st_area(evhoe_shp)) # calcul de l'aire de chaque strate (en m²)

# Regrouper les strates constituées de plusieurs polygones (Cn2)
evhoe_shp <- evhoe_shp %>%
  group_by(STRATE) %>%
  summarise(area_strata = sum(area_strata))

# Tracé de côte
mapBase <- map("worldHires", fill = T, plot = F)
mapBase <- st_as_sf(mapBase) %>% filter(ID %in% c("France", "Spain", "UK", "Ireland"))

# Espèce pour l'analyse
species <- "Merluccius_merluccius"
```

Aire géographique et strates de la campagne

```
xlims <- c(-12, 0)
ylims <- c(42, 52)
```



Figure 1: Récolte des données EVHOE.

```

Strata_plot <- ggplot(evhoe_shp)+  

  geom_sf(aes(fill=STRATE))+  

  geom_sf(data=mapBase)+  

  coord_sf(xlim = xlims, ylim = ylims, expand = FALSE)+  

  theme_bw() +  

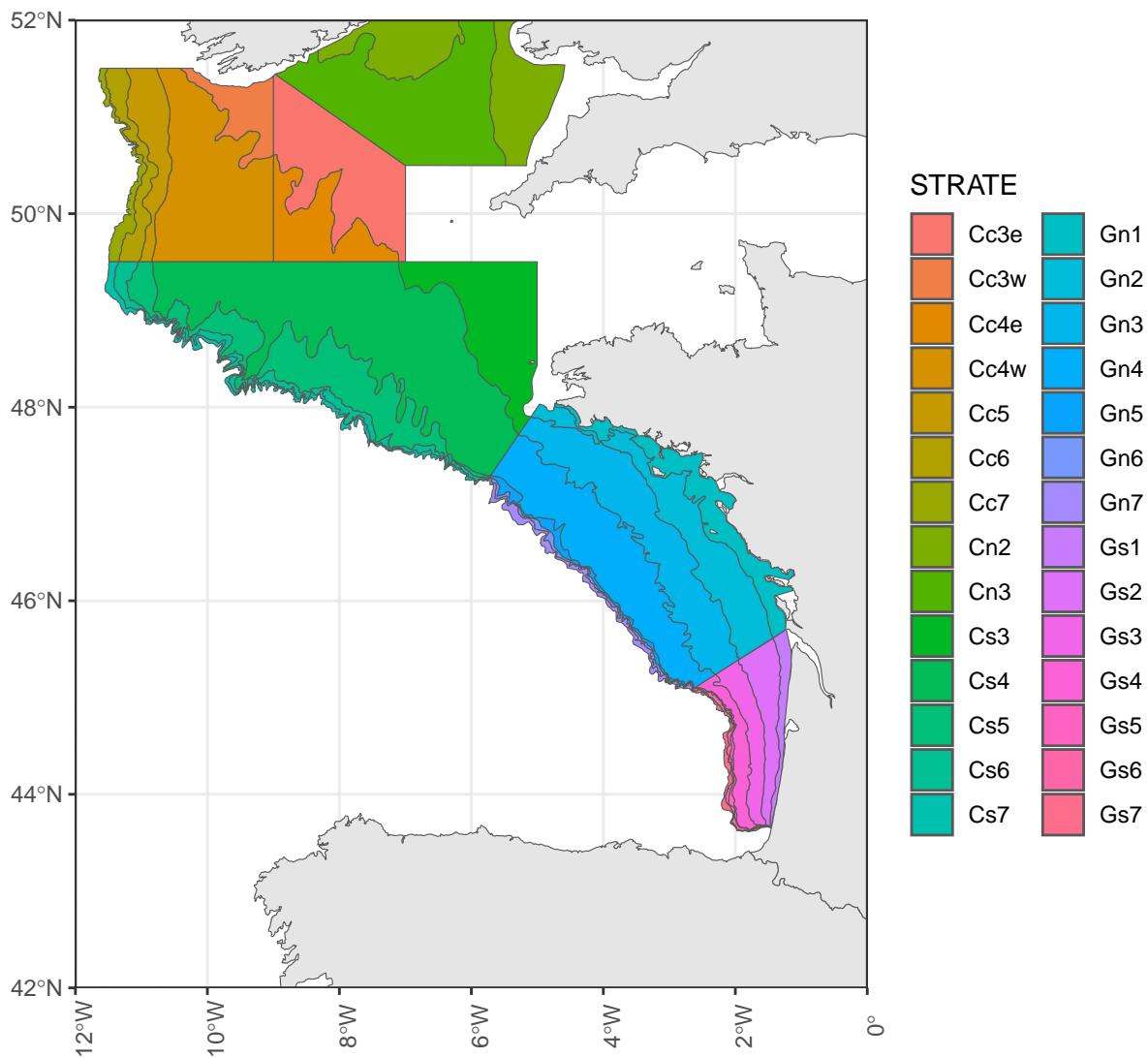
  theme(axis.text.x = element_text(angle = 90),  

        plot.title = element_text(hjust = 0.5, face = "bold", size=14),  

        panel.spacing.x = unit(4, "mm"))+  

  ylab("")+xlab("")

```



Données de traits de chaluts

```

# Pour le calcul de la taille de l'échantillon:  

# --> l'ouverture du chalut est de 10 m  

# l'aire balayé par le chalut correspond  

# à la distance parcourue (Distance) x l'ouverture du chalut (10m)

```

```

Haul_df <- Save_Datras$datras_HH.full %>%
  dplyr::select(Year, long, lati, StNo, HaulNo, Depth, Distance) %>%
  mutate(Area_swept = Distance * 10) # aire chalutée par chaque trait de chalut
Haul_df$Area_swept[which(is.na(Haul_df$Area_swept))] <- mean(Haul_df$Area_swept, na.rm = T)

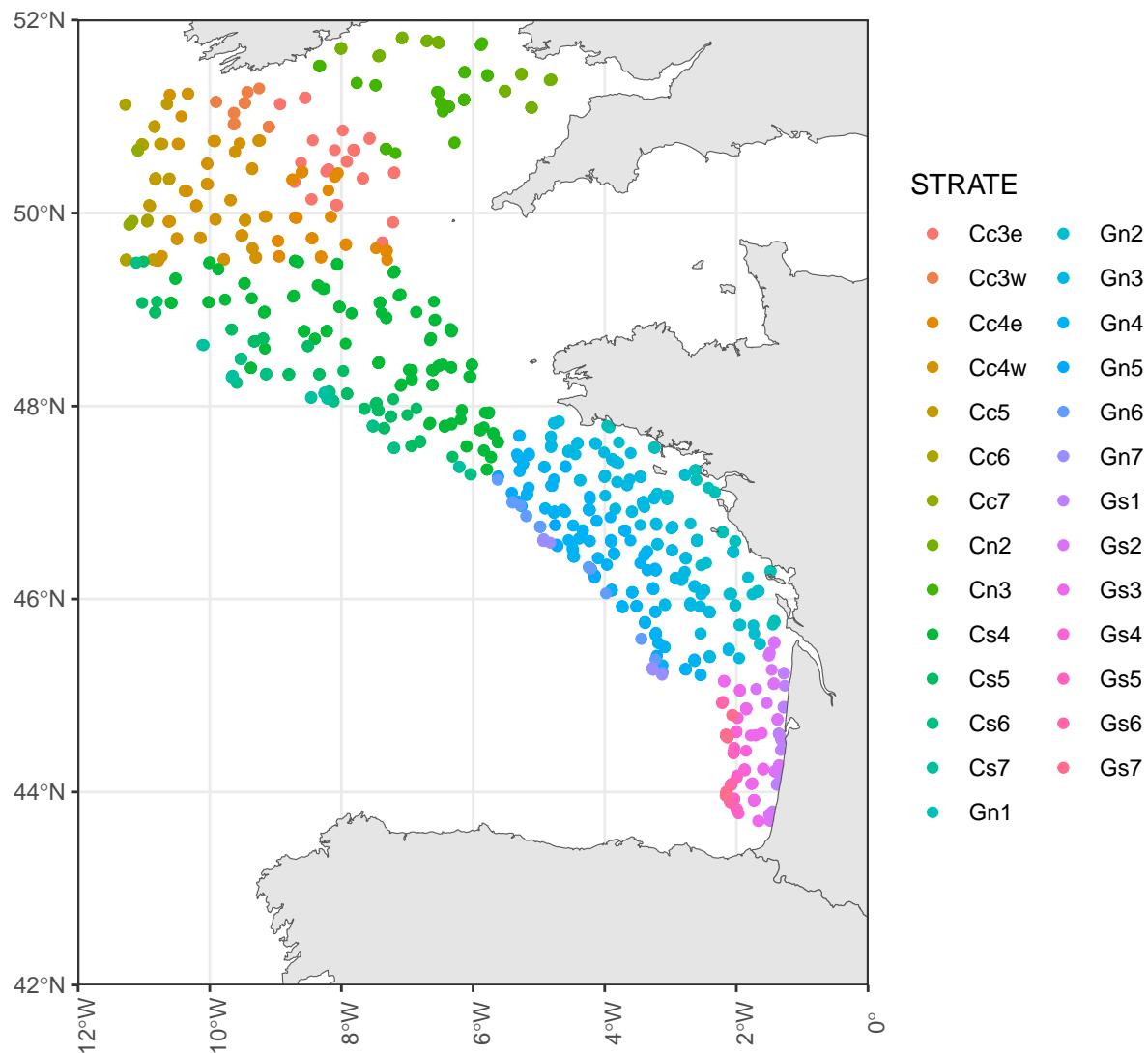
year_vec <- unique(Haul_df$Year)
year_vec <- year_vec[order(year_vec)]
n_year <- length(year_vec)

# Converti en sf et jointure avec le shapefile EVHOE
Haul_sf <- st_as_sf(Haul_df, coords=c("long", "lati"), crs = st_crs(evhoe_shp))
Haul_sf_2 <- st_intersection(Haul_sf, evhoe_shp)

## Warning: attribute variables are assumed to be spatially constant throughout
## all geometries

# Plot des points échantillonnés
Haul_plot <- ggplot(Haul_sf_2) +
  geom_sf(aes(col=STRATE)) +
  geom_sf(data=mapBase) +
  coord_sf(xlim = xlims, ylim = ylims, expand = FALSE) +
  theme_bw() +
  theme(axis.text.x = element_text(angle = 90),
        plot.title = element_text(hjust = 0.5, face = "bold", size=14),
        panel.spacing.x = unit(4, "mm")) +
  ylab("") + xlab("")

```



Données de captures

```
# Niveau d'aggrégation --> trait de chalut, espèce, classe de taille, sexe
Catch_df <- Save_Datras$datras_sp.HL.full %>%
  dplyr::select(Year, long, lati, StNo, HaulNo, scientificname, LngtClass, TotalNo) %>%
  group_by(Year, long, lati, StNo, HaulNo, scientificname) %>%
  dplyr::summarise(TotalNo = sum(TotalNo))

## `summarise()` has grouped output by 'Year', 'long', 'lati', 'StNo', 'HaulNo'.
## You can override using the '.groups' argument.

# Pivotter le tableau de données pour avoir les espèces en colonne
Catch_df_2 <- full_join(Catch_df, Haul_df) %>%
  mutate(haul_id = paste0(StNo, "-", HaulNo, "-", Year)) %>%
  tidyr::pivot_wider(names_from = scientificname, values_from = TotalNo)

## Joining with `by = join_by(Year, long, lati, StNo, HaulNo)`

# Joindre avec le shapefile EVHOE
Catch_sf_2 <- st_as_sf(Catch_df_2,
  coords = c("long", "lati"),
  crs=st_crs(evhoe_shp)) %>%
  st_intersection(evhoe_shp)

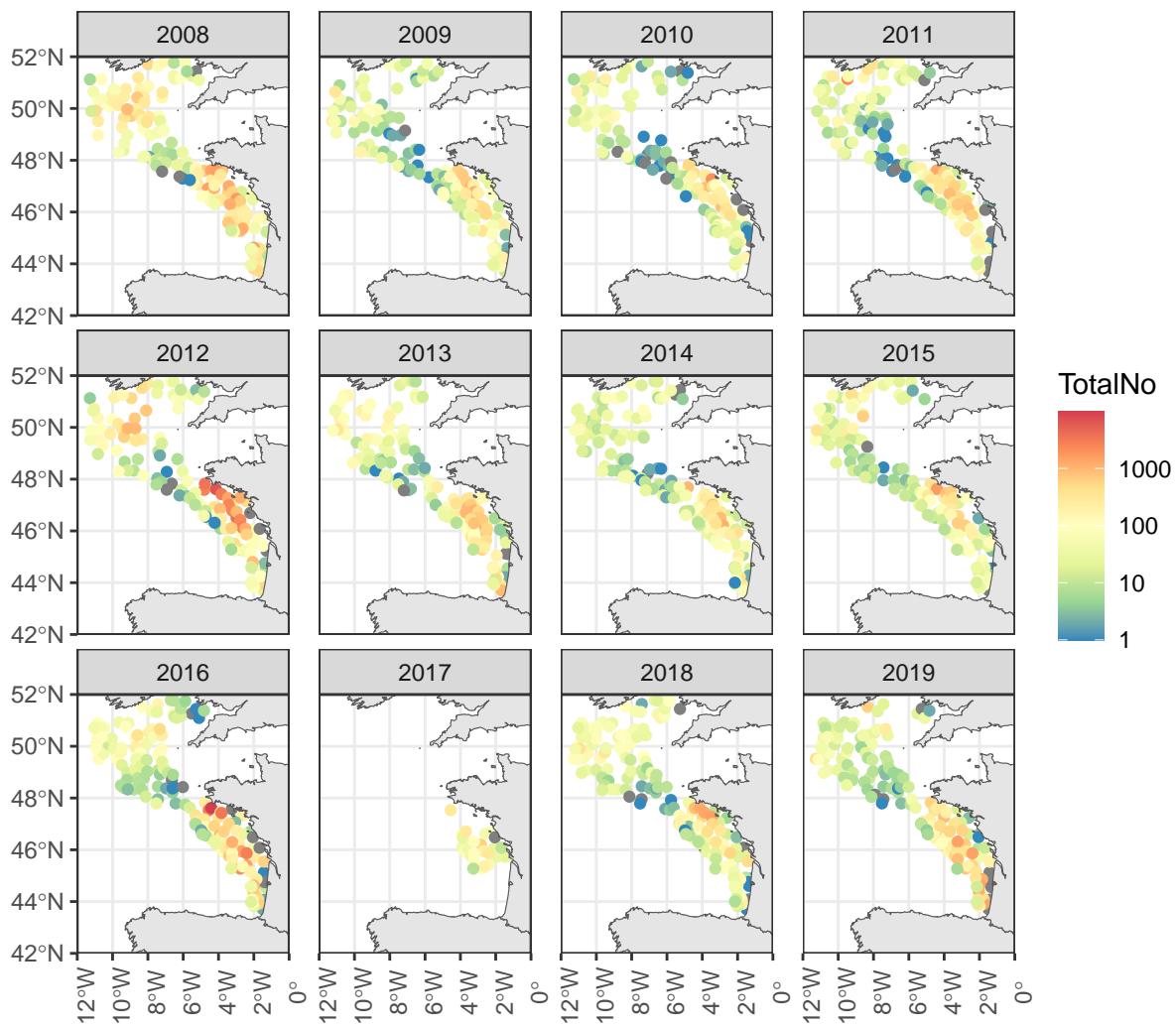
## Warning: attribute variables are assumed to be spatially constant throughout
## all geometries

# Filtrer sur les espèces
Catch_sf_3 <- Catch_sf_2 %>%
  select_at(vars(Year, StNo, HaulNo, STRATE, area_strata, Area_swept, contains(species)))
colnames(Catch_sf_3)[which(colnames(Catch_sf_3) == species)] <- "TotalNo"
Catch_sf_3$TotalNo[which(is.na(Catch_sf_3$TotalNo))] <- 0

# Plot
Evhoe_plot <- ggplot(Catch_sf_3)+
  geom_sf(aes(col=TotalNo))+
  scale_color_distiller(palette="Spectral", trans = 'log10')+
  facet_wrap(~Year)+
  geom_sf(data=mapBase)+
  coord_sf(xlim = xlims, ylim = ylims, expand = FALSE)+
  theme_bw()+
  theme(axis.text.x = element_text(angle = 90),
        plot.title = element_text(hjust = 0.5, face = "bold", size=14),
        panel.spacing.x = unit(4, "mm"))+
  ggtitle("Données de captures (en effectif)")+
  ylab("")+xlab("")

## Warning in scale_color_distiller(palette = "Spectral", trans = "log10"): log-10
## transformation introduced infinite values.
```

Données de captures (en effectif)



Définition de la population statistique étudiée, les individus statistiques et la variable étudiée.

De façon simple, il faut considérer que le **domaine d'étude** (= la “population statistique”) est le Golfe de Gascogne/Mer Celtique (GdG/MC).

Sur ce domaine, les scientifiques réalisent des relevés en certains points géographiques (= trait de chalut).

Les **unités statistiques** de l'étude statistique sont les traits de chalut (dit autrement des points géographiques échantillonnés dans le GdG/MC).

L'**échantillon** est donc constitué de l'ensemble des traits de chaluts qui ont été échantillonnés suivant un plan d'échantillonnage stratifié à allocation proportionnelle (i.e. nombre d'échantillon dans chaque strate est proportionnelle à l'aire de chaque strate). La stratification est basée sur les classes de profondeur et les grandes zones géographiques du GdG/MC. La **taille de l'échantillon** (n) correspond à la surface chalutée (`Swept_area`).

La **taille de la population** (N) correspond à l'ensemble des traits de chaluts possible soit l'ensemble du domaine d'étude.

Attention: le nombre de merlus ne sont pas les individus statistiques !

Pour chaque trait de chalut on observe un effectif de merlu. Le nombre de merlu observé pour un trait de chalut est la **variable d'intérêt** (y).

Le paramètre d'intérêt (t_y) est le nombre total de merlu sur le domaine d'étude.

A partir de la variable d'intérêt, on peut formuler un estimateur de t_y noté \hat{t}_y et calculer l'estimations.

Attention à bien différencier la population de merlu et la population statistique qui est ici étudiée (l'ensemble du GdG/MC dans lequel on échantillonne un ensemble de localisation = les traits de chaluts).

Questions

- Faire une analyse exploratoire des données (e.g. résumés statistiques de la variable étudiée, nombre de strate, représentation graphique de la variable étudiée) et comprendre le lien entre les jeux de données. Y a-t-il des données manquantes dans les data frames `Haul_df` et `Catch_df`? Si oui, comment sont-elles imputées ?
- Redonner l'expression d'un estimateur stratifié et sa variance pour le total (\hat{t}_y). Faire le lien entre les notations de l'estimateur et les données du TP (par exemple à quoi correspondent n_h , Y_i , h dans le cas d'application?).

Indications: La taille de l'échantillon pour une strate h correspond à la surface chalutée dans la strate donnée (`Area_swept`). Pour la taille d'une strate h (N_h), on prendra la surface de la strate (variable `area_strata` dans le shapefile `evhoe_shp` et le data frame `Haul_sf_2`).

- Ci-dessous l'abondance du merlu estimé pour l'année 2012 à l'aide des fonctions `HTstrata` et `varest`. En vous basant sur ces codes, estimer l'abondance de merlu pour chaque année. Représenter sur la même figure les estimations annuelles d'abondance et les intervalles de confiance associés.

Indications: Pour chaque année, calculer une estimation de l'abondance et calculer l'écart-type associé avec les fonctions `HTstrata` et `varest` (Cf. TD/TP sampling). Les représenter simultanément sur le même graphique.

```

Catch_sf_i <- Catch_sf_3 %>%
  filter(Year == 2012)

samp_per_strata <- Catch_sf_i %>%
  data.frame() %>%
  dplyr::select(-geometry) %>%
  group_by(Year, STRATE) %>%
  summarise(Area_swept_strata = sum(Area_swept))

## `summarise()` has grouped output by 'Year'. You can override using the
## `.` argument.

Catch_sf_i2 <- inner_join(Catch_sf_i, samp_per_strata)

## Joining with `by = join_by(Year, STRATE)`

Est_Ab <- HTstrata(y = Catch_sf_i2$TotalNo,
                     pik = Catch_sf_i2$Area_swept_strata / (Catch_sf_i2$area_strata),
                     strata = as.numeric(factor(Catch_sf_i2$STRATE)))

Var_Est_Ab <- varest(Ys=Catch_sf_i2$TotalNo,
                      pik=as.numeric(Catch_sf_i2$Area_swept_strata / (Catch_sf_i2$area_strata)))

## Estimation du total de l'abondance
Est_Ab

##           [,1]
## [1,] 2234911616

## Ecart type associé
sqrt(Var_Est_Ab)

## [1] 376764164

## Coefficient de variation
sqrt(Var_Est_Ab) / Est_Ab

##           [,1]
## [1,] 0.1685812

```

- Evaluer l'impact de la stratification sur l'estimation de l'abondance du merlu.

Indications: regrouper les strates suivant leur zone géographique *i.e.* les strates doivent être regroupées suivant leur deux premières lettres (Cc, Cn, Cs, Gn, Gs). Calculer des estimations de l'abondance par année pour cette nouvelle stratification et comparer aux estimations précédentes. Pour regrouper les strates sous R, il est possible d'utiliser le package **stringr** (package d'opération sur les chaînes de caractères).