Data mobilisation from GBIF to the EBV Data Portal

Notebook 2 - Data exploration of species listed in the Birds Directive Annex I

true

2024-08-23

Introduction

In this notebook we explore the occurrence data of species listed in the Annex I of the Birds Directive available in GBIF until June 2024. To do this, a Birds occurrence cube was previously created using the occurrence cube software developed by GBIF under the Biodiversity Building Blocks for Policy (B3) project. Details of the data query in GBIF are available at doi: 10.15468/dl.uh84tp. The cube generation script is also part of this repository.

Note: This series of notebooks is part of the results of Task 3.3 of the Biodiversity Building Blocks for Policy project funded by the European Union's Horizon Europe Research and Innovation Programme (ID No 101059592). Additional notebooks exploring the results and calculating simple metrics are also available in the same repository.

Load library and input data

```
rm(list=ls())
gc()
```

```
##           used (Mb) gc trigger  (Mb) max used  (Mb)
## Ncells 1793751 95.8    3515202 187.8  3515202 187.8
## Vcells 3778001 28.9   10146329  77.5  8388608  64.0
```

```
# load requiered libraries
library(here)
library(b3gbi) # for csv occurrence cubes
library(ggplot2)
library(dplyr)
library(lubridate)
library(stringr)
```

After loading the necessary libraries, we load the occurrence cube of birds obtained previously through the GBIF API.

```r
# File name from the JSON query
occcube <- "0062979-240626123714530"

# Load occurrence cube using b3gbi
cin <- process_cube(here(paste0("output/datacubes/csv/birds/", occcube,".csv")))

# Load species taxonomy only with 'accepted' scientific names in the GBIF backbone taxonomy
tax <- read.csv(here("input/data/birds/taxonomy/list193birds_directive_annexi_allaccepted_usagekey_gbif
```

Data Analysis

Calculate Total Number of Occurrence First we calculate the total number of occurence by family.

```r
cdata <- cin[["data"]]
# rename columns
colnames(cdata)[colnames(cdata) == "order"] <- "order_" # Note that if the name only remains as "order"
colnames(cdata)[colnames(cdata) == "taxonKey"] <- "acceptedUsageKey"

# Aggregate occurrences at the family level
cag <- cdata %>%
  group_by(family) %>%
  summarize(totalOcc = sum(obs), Order = first(order_))

# Sort in ascending order
cag <- cag[order(cag$totalOcc), ]

# Save file
write.csv(cag, here("output/summary_data/csv/birds/summary_birds_total_occurrences_byfamily+order.csv"))
```

Now, we calculate the total number of occurence by species.

```r
# Aggregate occurrences at species level
cag_sp <- cdata %>%
  group_by(acceptedUsageKey) %>%
  summarize(totalOcc = sum(obs))

# Sort in ascending order
cag_sp <- cag_sp[order(cag_sp$totalOcc), ]

# Check family with one record
position <- which(cdata$family == "Columbidae")
second_row <- cdata %>%
  slice(position)
as.data.frame(second_row) # taxon key 2495406
```

```
##   yearMonth     cellCode     basisofrecord                            datasetkey
## 1   2010-12 10kmE271N196 HUMAN_OBSERVATION 8ea87d00-f94e-4221-8f2f-2e9150132361
##   countrycode acceptedUsageKey scientificName     family class        order_
## 1          PT          2495406 Columba trocaz Columbidae  Aves Columbiformes
##   obs year xcoord ycoord resolution
## 1   1 2010 271000 196000       10km
```

```r
# Rename columns for joining occurence cube with GBIF Backbone taxonomy. Remember our species identifie

# Merge data sets and sort in ascending order
xout <- cag_sp %>%
  inner_join(tax[,c("scientificName", "key", "kingdom", "phylum", "class", "order", "family", "acceptedU

xout <- xout[order(xout$totalOcc), ]

# Save file
write.csv(xout, here("output/summary_data/csv/birds/summary_birds_total_occurrences_acceptedusagekey.csv
```

```r
# Find what species have no records in GBIF
noocc <- anti_join(tax, cag_sp, by = "acceptedUsageKey")
write.csv(noocc, here("output/summary_data/csv/birds/birds_no_occurrences_acceptedusagekeys.csv"))
print(noocc[,c("scientificName", "acceptedUsageKey")])
```

Identify birds without species occurrences through the GBIF occurrence cube software

```
##                                          scientificName acceptedUsageKey
## 1       Accipiter gentilis arrigonii (O.Kleinschmidt, 1903)          4408405
## 2                      Accipiter nisus granti Sharpe, 1890          4408410
## 3      Anser albifrons flavirostris Dalgety & P.Scott, 1948          6178325
## 4    Calidris alpina schinzii (C.L.Brehm & Schilling, 1822)          6177960
## 5             Certhia brachydactyla dorotheae Hartert, 1904          4408795
## 6                               Columba bollii Godman, 1872          2495427
## 7                           Columba junoniae Hartert, 1916          2495434
## 8                  Columba palumbus azorica Hartert, 1905          2495459
## 9          Dendrocopos major canariensis (A.F.Koenig, 1889)          6177296
## 10                 Dendrocopos major thanneri le Roi, 1911          6177279
## 11                                      Egretta alba alba          7190969
## 12               Fringilla coelebs ombriosa Hartert, 1913          6175538
## 13  Fringilla teydea Webb, Berthelot & Moquin-Tandon, 1836          2494442
## 14                          Gelochelidon nilotica nilotica          7192423
## 15                               Lagopus mutus helveticus          5227692
## 16                  Lagopus mutus pyrenaicus Hartert, 1921          5227710
## 17                      Parus ater cypriotes Dresser, 1888          5844831
## 18              Perdix perdix hispaniensis Reichenow, 1892          2473966
## 19                   Perdix perdix italica E.Hartert, 1917          2473961
## 20 Phalacrocorax aristotelis desmarestii (Payraudeau, 1826)          4408479
## 21                            Pyrrhula murina Godman, 1866          4408970
## 22                   Saxicola dacotiae (Meade-Waldo, 1889)          2492522
## 23                                 Tetrao tetrix tetrix          7191070
## 24   Troglodytes troglodytes fridariensis Williamson, 1951          5739334
## 25                       Uria aalge ibericus Bernis, 1948          6065721
```

As a remark, the list above refers mostly to subspecies that cannot be identified by the 'speciesKey' of our JSON query, and a few correspond to species from the Canary Islands or the Azores that are not covered by the EEA grid.

Prepare taxonomy for the EBV Data Portal Due to the absence of data for 25 species, we prepared a new file including only taxonomic information of 168 species based on our results.

```
# Filter taxonomy file for only species with data
# Import file for all Birds Species and following the ebvcube format

# Subset only for species with data occurrences in GBIF
spskey <- unique(cin[["data"]][["taxonKey"]]) # "taxonKey" is the name of the species key in b3bgi

gbif_spx <- tax %>%
  filter(acceptedUsageKey %in% spskey)

# Unmatched species
gbif_spx_missing <- tax %>%
  filter(acceptedUsageKey %in% setdiff(tax$acceptedUsageKey, spskey))

# Sort the right taxonomic order for filling the EBV Data Portal requierement before saving
gbif_spx2 <- gbif_spx[,c("kingdom", "phylum", "class", "order", "family", "genus", "species", "acceptedU

# Save files
write.csv(gbif_spx2[,1:8], here("input/data/birds/taxonomy/list168_BirdsAnnexI_gbif_accepted_ebvcube+ac
write.csv(gbif_spx2, here("input/data/birds/taxonomy/list168_BirdsAnnexI_gbif_accepted_ebvcube+acceptedU
write.csv(gbif_spx_missing, here("input/data/birds/taxonomy/list_missing_BirdsAnnexI_gbif_accepted_ebvcu
```

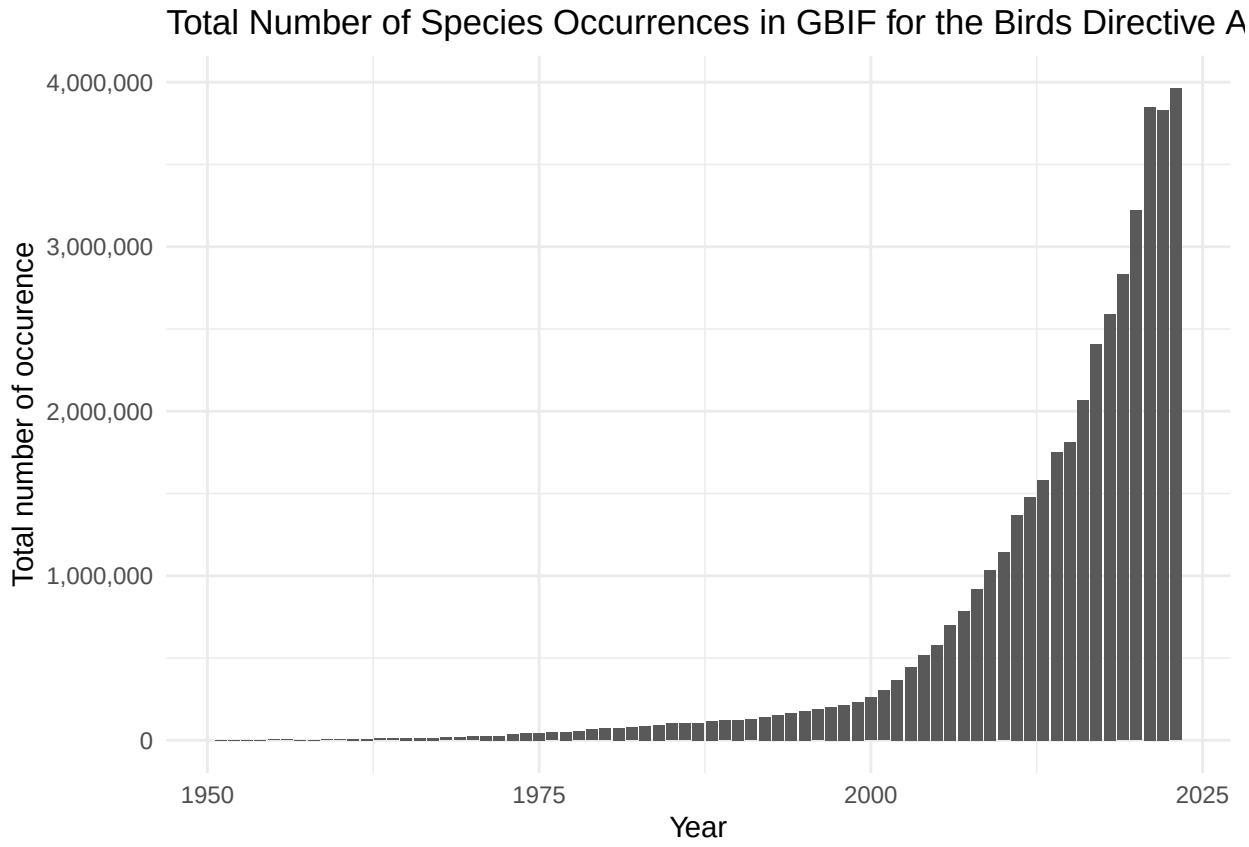
Data Exploration

We will start checking the number of records since 1950.

```
# Convert date from character to numeric
todates <- as.data.frame(str_split(cdata$yearMonth, "-", simplify = TRUE))
colnames(todates) <- c("year", "month")

# Add year and month columns separate to the initial dataset
cdata$year <- todates$year
cdata$month <- todates$month

# Filter data starting from 1950
cdata2 <- cdata %>%
  filter(year > 1950)

# Group data by year
cag_year <- cdata2 %>%
    group_by(year) %>%
  summarize(totalOcc = sum(obs))

# Group data by month
cag_month <- cdata %>%
    group_by(month) %>%
  summarize(totalOcc = sum(obs))
```

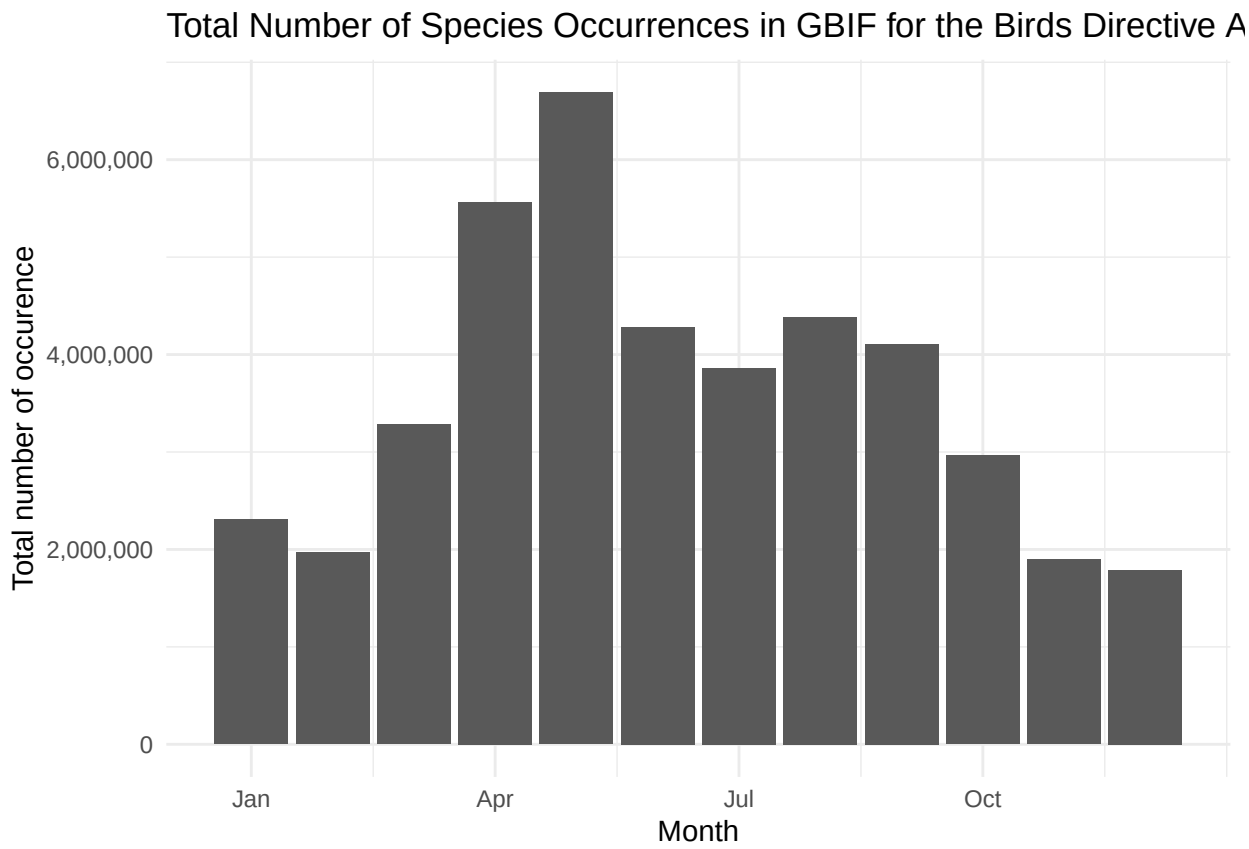
In the following code chunk we plot the GBIF occurrences available since 1950 for the Bird Directive species in Annex I.

```
ggplot(cag_year, aes(x = as.numeric(year), y = totalOcc)) +
  geom_bar(stat = "identity") +
  # guides(fill=guide_legend(ncol=1)) +
  scale_y_continuous(label = scales::label_comma(accuracy = 1)) +
  # coord_flip() +
  labs(title = "Total Number of Species Occurrences in GBIF for the Birds Directive Annex I since 1950"
       x = "Year",
       y = "Total number of occurence") +
       guides(color = guide_legend(override.aes = list(size = 5))) +
  theme_minimal()
```



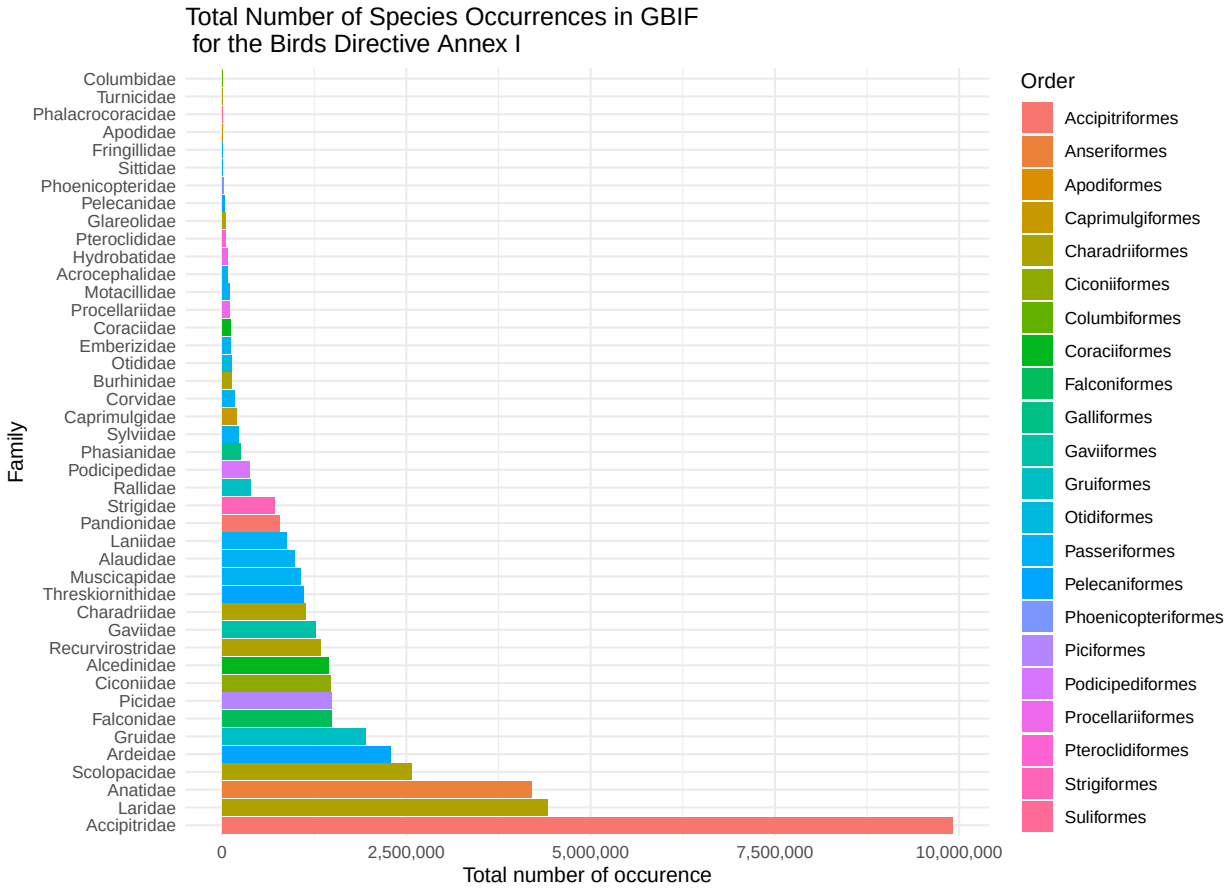
Now we plot the all GBIF occurrences per month for the Bird Directive species in Annex I.

```
ggplot(cag_month, aes(x = as.numeric(month), y = totalOcc)) +
  geom_bar(stat = "identity") +
  # guides(fill=guide_legend(ncol=1)) +
  scale_x_continuous(breaks = c(1,4,7,10), labels = c("Jan", "Apr", "Jul", "Oct")) +
  scale_y_continuous(label = scales::label_comma(accuracy = 1)) +
  labs(title = "Total Number of Species Occurrences in GBIF for the Birds Directive Annex I per month",
       x = "Month",
       y = "Total number of occurence") +
       guides(color = guide_legend(override.aes = list(size = 5))) +
  theme_minimal()
```



The next plot shows the occurrence of birds by family and the bars are coloured by the taxonomic order.

```
# Create a horizontal bar plot
ggplot(cag, aes(x = reorder(family, -totalOcc), y = totalOcc, fill = Order)) +
  geom_bar(stat = "identity") +
  guides(fill=guide_legend(ncol=1)) +
  scale_y_continuous(label = scales::label_comma(accuracy = 1)) +
  coord_flip() +
  labs(title = "Total Number of Species Occurrences in GBIF\n for the Birds Directive Annex I",
       x = "Family",
       y = "Total number of occurence") +
       guides(color = guide_legend(override.aes = list(size = 5))) +
  theme_minimal()
```

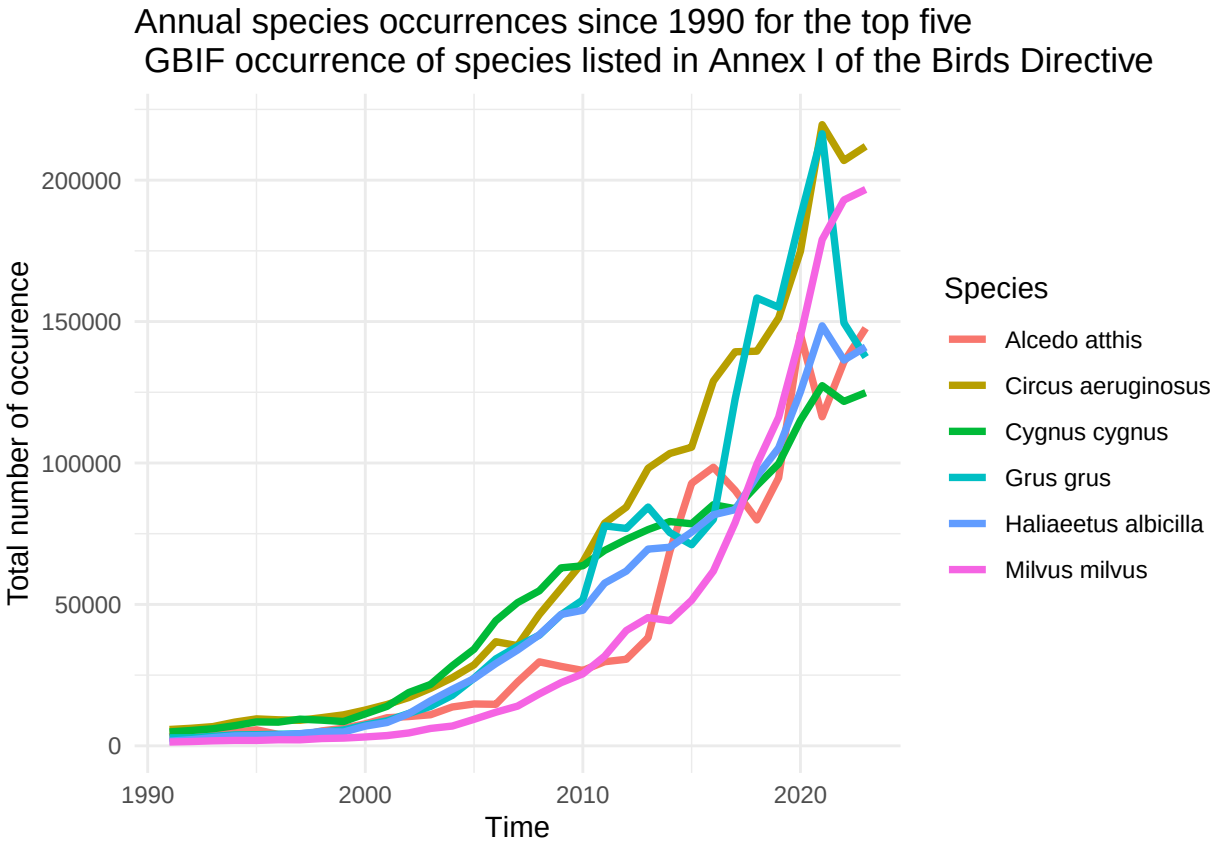


Our last plot shows the increasing number of records since 2000 for the five species with the highest records.

```
# Select the five species with more records
cag_sel <- cag_sp[(dim(cag_sp)[1]-5):dim(cag_sp)[1],]

# Subset data set
cag_top5 <- cdata %>%
  inner_join(cag_sel[,c("acceptedUsageKey")], by = "acceptedUsageKey") %>%
  filter(year > 1990)

# Aggregate occurrences by year
cag_top5year <- cag_top5 %>%
  group_by(year, scientificName) %>%
  summarize(totalOcc = sum(obs), scientificName = first(scientificName))
```

Lastly, this plot shows the five species with more records in GBIF since 2000.

```
ggplot(cag_top5year, aes(x = as.numeric(year), y = totalOcc, color = scientificName, name = "Species"))
  geom_line(linewidth = 1.3) +
  guides(color=guide_legend(ncol=1, title = "Species")) +
  labs(title = "Annual species occurrences since 1990 for the top five \n GBIF occurrence of species lis
       x = "Time",
       y = "Total number of occurence") +
  theme_minimal()
```


Annual species occurrences since 1990 for the top five
GBIF occurrence of species listed in Annex I of the Birds Directive
Total number of occurence
200000
150000
100000
50000
0
1990
2000
2010
2020
Time
Species
Alcedo atthis
Circus aeruginosus
Cygnus cygnus
Grus grus
Haliaeetus albicilla
Milvus milvus