ROBITT\_template

Rob Boyd

18 May 2022

# 1. Iteration

### 1.1 ROBITT iteration number

|  |  |  |
| --- | --- | --- |
| Iteration | Comments |  |
| 1 |  |  |

# 2. Research statement and pre-bias assessments

## Statistical population of interest

**2.1 Define the statistical target population about which you intend to make inferences.**

|  |  |  |
| --- | --- | --- |
| Domain | Extent | Resolution |
| Geographic | United Kingdom (UK) | 1km grid cells |
| Temporal | 1970-2020 | Annual increments |
| Taxonomic (or other relevant organismal domain such as functional group etc.) | All species of soldierfly | Species |
| Environmental | Environmental space in the UK | 1km to match the geographic resolution |

# Inferential goals

**2.2 What are your inferential goals?**

To estimate temporal trends in species' occupancy (proportion of occupied grid cells). The individual species trends will be "averaged" to construct a multispecies indicator of change.

## Data provenance

**2.3 From where were your data acquired (please provide citations, including a DOI, wherever possible)? What are their key features in respect of the inferential aims of your study (see the guidance document for examples)?**

The data are presence-only records of soldierfly occurrences recorded in the UK from 1990-2020.

dat <- read.csv("W:/PYWELL\_SHARED/Pywell Projects/BRC/\_BRC\_dataflow/Research Datasets/Soldierflies/2022/Research dataset/Soldierflies\_2022.csv")  
  
names(dat)

## [1] "research\_dataset\_id" "research\_dataset\_name"   
## [3] "raw\_dataset\_id" "raw\_dataset\_name"   
## [5] "source\_dataset\_id" "original\_dataset\_id"   
## [7] "source\_dataset" "additional\_source\_dataset\_info"  
## [9] "citation\_req" "date\_of\_capture"   
## [11] "capture\_method" "capture\_purpose"   
## [13] "permission\_info" "data\_filters"   
## [15] "source\_TVK" "source\_name"   
## [17] "source\_taxon\_author" "source\_qualifier"   
## [19] "source\_startdate" "source\_enddate"   
## [21] "source\_location" "recommended\_tvk"   
## [23] "recommended\_name" "taxon\_qualifier"   
## [25] "species\_long" "recommended\_authority"   
## [27] "startdate" "enddate"   
## [29] "grid\_reference" "hectad"   
## [31] "monad" "latitude"   
## [33] "longitude" "taxon\_rank"   
## [35] "taxon\_group\_one" "taxon\_group\_two"   
## [37] "uksi\_data" "rd\_comments"

dat$raw\_dataset\_name[1]

## [1] "soldierflies\_and\_allies\_indicia\_2022\_03\_15"

dat$citation\_req[1]

## [1] "Soldierflies and Allies Recording Scheme (15-MAR-2022). Records via iRecord."

dat$date\_of\_capture[1]

## [1] "15-MAR-2022"

## Data processing

**2.4 Provide details of, and the justification for, all of the steps that you have taken to clean the data described above prior to analyses.**

I modified the data described above in three ways. First, I removed records that were not resolved to one day. Second, I removed records that were duplicated in terms of date, grid cell and species name. And finally, I reprojected records collected in Northern Ireland from the Irish national grid (OSNI 1951) to the British national grid (OSGB 1936).

## process species occurrence data  
  
# first remove data not resolved to one day  
  
dat <- dat[- which(dat$startdate != dat$enddate)]  
  
# then remove duplicates (in terms of species name, date and monad)  
  
dat <- dat[- which(duplicated(dat[, c("recommended\_name", "startdate", "monad")]))]  
  
# drop columns that are not needed for analysis   
  
dat <- dat[, c("recommended\_name", "monad", "startdate")]  
  
# extract coordinates from grid references (needed by occAssess)  
  
coords <- BRCmap::gr\_let2num(gridref = dat$monad,  
 centre = TRUE,  
 return\_projection = TRUE)  
  
dat <- cbind(dat, coords)  
  
# check if there are any coordinates on the OSNI projection  
  
table(coords$PROJECTION)

##   
## OSGB OSNI UTM30   
## 130467 359 40

# if yes then reproject these onto OSGB  
  
if ("OSNI" %in% coords$PROJECTION) {  
   
 GBCRS <- sp::CRS("+proj=tmerc +lat\_0=49 +lon\_0=-2 +k=0.9996012717 +x\_0=400000 +y\_0=-100000 +ellps=airy +datum=OSGB36 +units=m +no\_defs")  
   
 NICRS <- sp::CRS("+proj=tmerc +lat\_0=53.5 +lon\_0=-8 +k=1 +x\_0=200000 +y\_0=250000 +ellps=airy +towgs84=482.5,-130.6,564.6,-1.042,-0.214,-0.631,8.15 +units=m +no\_defs")  
   
 datNI <- dat[which(dat$PROJECTION == "OSNI"), ]  
   
 datGB <- dat[which(dat$PROJECTION == "OSGB"), ]  
  
 NIcoords <- datNI[, c("EASTING", "NORTHING")]  
  
 sp::coordinates(NIcoords) <- c("EASTING", "NORTHING")  
  
 sp::proj4string(NIcoords) <- NICRS  
  
 NIcoords <- sp::spTransform(NIcoords, GBCRS)  
  
 datNI[,c("EASTING", "NORTHING")] <- data.frame(NIcoords)  
   
 dat <- rbind(datGB, datNI)  
   
}

## Warning in showSRID(uprojargs, format = "PROJ", multiline = "NO", prefer\_proj =  
## prefer\_proj): Discarded datum OSGB 1936 in Proj4 definition

## Warning in showSRID(uprojargs, format = "PROJ", multiline = "NO", prefer\_proj  
## = prefer\_proj): Discarded datum Unknown based on Airy 1830 ellipsoid in Proj4  
## definition

# remove more columns that aren't needed  
  
dat <- dat[, c("recommended\_name", "startdate", "EASTING", "NORTHING")]  
  
head(dat)

## recommended\_name startdate EASTING NORTHING  
## 2 Hermetia illucens 2020-08-16 416500 160500  
## 4 Anthrax anthrax 2012-06-12 613500 158500  
## 5 Anthrax anthrax 2020-05-23 613500 158500  
## 6 Anthrax anthrax 2020-05-17 613500 157500  
## 7 Anthrax anthrax 2020-05-18 613500 158500  
## 8 Anthrax anthrax 2019-06-06 613500 158500

# create a new column for year (needed by occAssess). Note we'll keep date as it will allow  
# us to look specifically at repeat visits later   
  
dat$year <- substr(dat$startdate, 1, 4)  
  
# create identifier and sptialUncertainty fields (again, needed by occAssess)  
  
dat$identifier <- "all\_data"  
  
dat$spatialUncertainty <- 1000  
  
head(dat)

## recommended\_name startdate EASTING NORTHING year identifier  
## 2 Hermetia illucens 2020-08-16 416500 160500 2020 all\_data  
## 4 Anthrax anthrax 2012-06-12 613500 158500 2012 all\_data  
## 5 Anthrax anthrax 2020-05-23 613500 158500 2020 all\_data  
## 6 Anthrax anthrax 2020-05-17 613500 157500 2020 all\_data  
## 7 Anthrax anthrax 2020-05-18 613500 158500 2020 all\_data  
## 8 Anthrax anthrax 2019-06-06 613500 158500 2019 all\_data  
## spatialUncertainty  
## 2 1000  
## 4 1000  
## 5 1000  
## 6 1000  
## 7 1000  
## 8 1000

## now create a second dataset with just the repeat visits   
  
repeats <- dat[which(duplicated(dat[, c("EASTING", "NORTHING", "startdate")])), ]  
  
repeats$identifier <- "repeat\_visits" # set identifier to distinguish from the rest  
  
# append to dat for analysis with occAssess  
  
dat <- rbind(dat, repeats)

# 3. Bias assessment and mitigation

## Assessment resolutions

**3.1 At what geographic, temporal and taxonomic resolutions (i.e. scales or grain sizes) will you conduct your bias assessment?**

I conducted the bias assessment at spatial and temporal resolutions of 1km and one year to match the statistical population about which I want to draw inferences (Table 2). However, it was not possible to assess the data at the species level; presence-only data say nothing about the spatial and temporal distribution of sampling where the focal species was not observed. Rather, I used the target group approach (Phillips et al., 2009) to approximate sampling effort, which is to say, I treated the spatial and temporal distribution of records for the whole taxonomic group (target group) as a proxy for the spatial and temporal distributions of sampling effort. In other words, if at least one species was recorded in some grid cell and at some time, then I assume that all species were searched for.

## Geographic domain

**3.2 Are the data sampled from a representative portion of geographical space in the domain of interest?**

**3.3 Are your data sampled from the same portions of geographic space across time periods?**

**3.4 If the answers to the above questions revealed any potential geographic biases, or temporal variation in geographic coverage, please explain, in detail, how you plan to mitigate them.**

## Environmental domain

**3.5 Are your data sampled from a representative portion of environmental space in the domain of interest?**

**3.6 Are your data sampled from the same portion of environmental space across time periods?**

**3.7 If the answers to the above questions revealed any potential environmental biases, or temporal variation in environmental coverage, please explain, in detail, how you plan to mitigate them.**

## Taxonomic domain (or other organismal domain, e.g., phylogenetic, trait space etc.)

**Is the sampled portion of the taxonomic (or phylogenetic, trait or other space if more relevant) space representative of the taxonomic (or other) domain of interest?**

**3.9 Do your data pertain to the same taxa/taxonomic domain across time periods?**

**3.10 If the answers to the above questions revealed any potential taxonomic biases, or temporal variation in taxonomic coverage, please explain, in detail, how you plan to mitigate them.**

## Other potential biases

**3.11 Are there other potential temporal biases in your data that relate to variables other than ecological states?**

**3.12 Are you aware of any other potential biases not covered by the above questions that might cause problems for your inferences?**

**3.13 If questions 3.11 or 3.12 revealed any important potential biases, please explain how you will mitigate them.**

# 4. Supporting references