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

Domain	Extent	Resolution
Environmental	Environmental space in the UK	Environmental 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. My code can be seen below, along with some metadata documenting the provenance of the data.

```
dat <- read.csv("W:/PYWELL_SHARED/Pywell Projects/BRC/_BRC_dataflow/Research Datasets/Soldierflies/2022/

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). My code can be seen below.

```
library(BRCmap)

## 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

  NICRS <- sp::CRS("+proj=tmerc +lat_0=53.5 +lon_0=-8 +k=1 +x_0=200000 +y_0=250000 +ellps=airy +towgs84=

  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
```

```r
# 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
```

```r
## now create a second dataset with just the repeat visits (visits to the same site in the same year bu

repeats <- dat[which(duplicated(dat[, c("EASTING", "NORTHING", "year")]) &
        !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?

To assess the geographic representativness of the data, I used what is called the Nearest Neighbour Index (NNI). The NNI is the ratio of the average nearest neighbour distances of the centroids of grid cells with records to the average nearest neighbour distances of simulated random distributions of the same density. Where the NNI is below 1, the data more clustered than a random distribution; where it is about 1, the data are approximately randomly distributed; and where it falls above 1, the data are overdispersed. Fig 1. clearly shows that the data are more clustered than a random distribution.

```
mask <- raster::raster("W:/PYWELL_SHARED/Pywell Projects/BRC/Rob Boyd/TSDA/SDMs/Data/SDMOutputs_Jan_Feb

# define time periods for analysis as required by occAssess

periods <- as.list(1970:2020)

NNI <- occAssess::assessSpatialBias(dat = dat,
                          periods = periods,
                          nSamps = 2,
                          degrade = TRUE,
                          mask = mask,
                          species = "recommended_name",
                          year = "year",
                          identifier = "identifier",
                          x = "EASTING",
                          y = "NORTHING",
                          spatialUncertainty = "spatialUncertainty",)
```

```
## Registered S3 method overwritten by 'spatstat.geom':
##   method     from
##   print.boxx cli

## Warning in FUN(X[[i]], ...): Fewer than 100 records in period 1 for
## repeat_visits . View this result with caution.

## Warning in FUN(X[[i]], ...): Fewer than 100 records in period 2 for
## repeat_visits . View this result with caution.

## Warning in FUN(X[[i]], ...): Fewer than 100 records in period 3 for
## repeat_visits . View this result with caution.

## Warning in FUN(X[[i]], ...): Fewer than 100 records in period 4 for
## repeat_visits . View this result with caution.

## Warning in FUN(X[[i]], ...): Fewer than 100 records in period 5 for
## repeat_visits . View this result with caution.
```

```
## Warning in FUN(X[[i]], ...): Fewer than 100 records in period 6 for
## repeat_visits . View this result with caution.

## Warning in FUN(X[[i]], ...): Fewer than 100 records in period 7 for
## repeat_visits . View this result with caution.

## Warning in FUN(X[[i]], ...): Fewer than 100 records in period 8 for
## repeat_visits . View this result with caution.

## Warning in FUN(X[[i]], ...): Fewer than 100 records in period 9 for
## repeat_visits . View this result with caution.

## Warning in FUN(X[[i]], ...): Fewer than 100 records in period 10 for
## repeat_visits . View this result with caution.

## Warning in FUN(X[[i]], ...): Fewer than 100 records in period 11 for
## repeat_visits . View this result with caution.

## Warning in FUN(X[[i]], ...): Fewer than 100 records in period 12 for
## repeat_visits . View this result with caution.

## Warning in FUN(X[[i]], ...): Fewer than 100 records in period 13 for
## repeat_visits . View this result with caution.

## Warning in FUN(X[[i]], ...): Fewer than 100 records in period 19 for
## repeat_visits . View this result with caution.

## Warning in FUN(X[[i]], ...): Fewer than 100 records in period 20 for
## repeat_visits . View this result with caution.

## Warning in FUN(X[[i]], ...): Fewer than 100 records in period 23 for
## repeat_visits . View this result with caution.

## Warning in FUN(X[[i]], ...): Fewer than 100 records in period 25 for
## repeat_visits . View this result with caution.

## Warning: `guides(<scale> = FALSE)` is deprecated. Please use `guides(<scale> =
## "none")` instead.
```

```
ggplot2::ggplot(data = NNI$data, ggplot2::aes(x = as.numeric(Period) + 1969, y = mean,
                            group = identifier, fill = identifier,
                            colour = identifier)) +
  ggplot2::geom_line() +
  ggplot2::theme_linedraw() +
  ggplot2::xlab("Year") +
  ggplot2::ylab("NNI") +
  ggplot2::geom_hline(yintercept = 1, colour = "red") +
  ggplot2::geom_ribbon(ggplot2::aes(ymin = lower, ymax = upper),
              alpha = 0.3) +
  ggplot2::labs(group = "",
       fill = "",
       colour = "") +
  ggplot2::theme(text = ggplot2::element_text(size = 25))
```

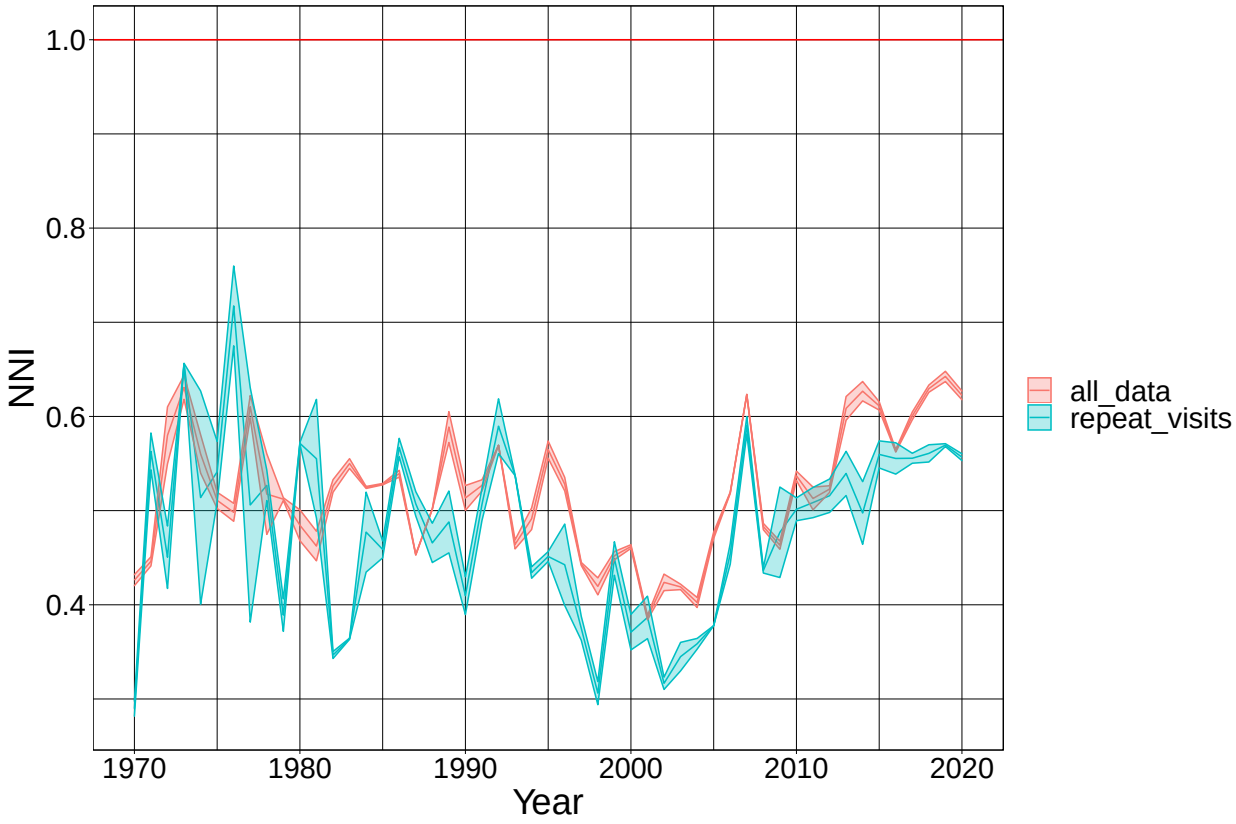

Fig. 2 shows the geographic distribution of 1km grid cells (monads) in which soldierflies have been recorded between 1970 and 2020. Looking at this figure it is clear that the majority of the data were recorded in England, whereas Scotland and Northern Ireland have been sampled to a lesser extent. This pattern is particularly evident when considering only grid cells with repeat visits (right hand panel in Fig. 2).

```
## load required data

# UK shapefile from BRCmap

data(UK)

shp <- UK[UK$REGION == "Great Britain", ]

shp2 <- UK[UK$REGION == "Ireland", ]

# fortify shapefile for use with ggplot2

mapGB <- ggplot2::fortify(shp)
```

```
## Regions defined for each Polygons
```

```
mapIr <- ggplot2::fortify(shp2)
```

```
## Regions defined for each Polygons
```

```
# map grid cells sampled at some point

spatCov <- occAssess::assessSpatialCov(periods = periods,
                                       dat = dat,
                                       species = "recommended_name",
                                       year = "year",
                                       identifier = "identifier",
                                       x = "EASTING",
                                       y = "NORTHING",
                                       spatialUncertainty = "spatialUncertainty",
                                       res = 1000,
                                       output = "overlap",
                                       minPeriods = 1,
                                       returnRaster = TRUE)
```

```
## Warning in FUN(X[[i]], ...): Some or all country names provided are not in
## unique(ggplot2::map_data(world)$region)

## Warning in FUN(X[[i]], ...): Some or all country names provided are not in
## unique(ggplot2::map_data(world)$region)
```

```
myCol <- rgb(255, 255, 255, max = 255, alpha = 0, names = "blue50")

rasterVis::gplot(spatCov$rasters) +
  ggplot2::geom_tile(ggplot2::aes(fill = value)) +
  ggplot2::facet_wrap(~variable) +
  ggplot2::geom_polygon(data = mapGB, ggplot2::aes(x = long,
                                                   y = lat, group = group), colour = "black",
                fill = myCol, inherit.aes = F) +
  ggplot2::geom_polygon(data = mapIr, ggplot2::aes(x = long,
                                                   y = lat, group = group), colour = "black",
                fill = myCol, inherit.aes = F) +
  ggplot2::theme_linedraw() +
  ggplot2::theme(axis.text.x=ggplot2::element_blank(),
        axis.text.y=ggplot2::element_blank()) +
  ggplot2::labs(fill = "Proportion
       of years
       sampled") +
  ggplot2::labs(x = "",
       y = "") +
  ggplot2::scale_fill_continuous(na.value = myCol) +
  ggplot2::guides(fill = "none") +
  ggplot2::theme(strip.text.x = ggplot2::element_text(size = 20))
```

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?

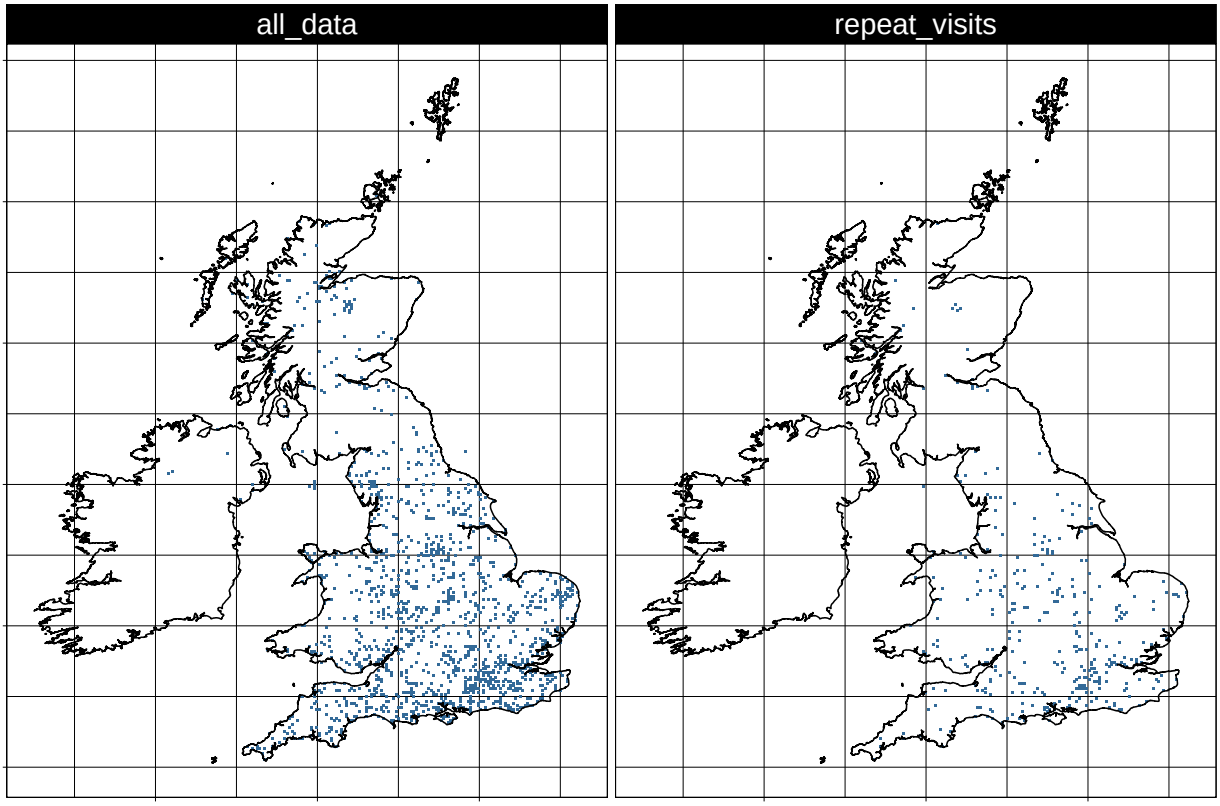

Figure 1: Figure 2. 1km grid cells in which at least one species was recorded between 1990 and 2020.

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