ROBITT\_template

Rob Boyd

18 May 2022

# Background

In this document I use ROBITT to assess the suitability of data from the Soldierflies and Allies recording scheme for the estimation of temporal trends in species' distributions. ROBITT comprises a number of questions designed to elicit information on the potential for biases in the data to undermine trend estimation. Generally, I answer each question using text and figures. I also provide the R code used to generate the figures and some simple statistics. Readers should not feel compelled to read the R code; the document is discernible without it.

In the BRC we have been developing a workflow to estimate trends in species' distributions from the type of data assessed here. This workflow is built on what is called an occupancy-detection model. This model uses data from repeat visits to the same location (here 1km grid cell) in the same year to estimate species' detection probabilities. For this reason I conduct this ROBITT assessment both on the dataset as a whole, and on the subset of data that are repeat visits.

This Rmarkdown document relies on data stored on UK CEH's W (shared) drive. If you do not have access to this drive, then it will not work. All members of the BRC should have access.

# 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 1) estimate temporal trends in species' occupancy (proportion of occupied grid cells), and 2) combine the individual species' trends to produce a multispecies indicator.

## 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. The data were collated by the Soldierfly and Allies recording scheme and accessed on 15th March 2022 via iRecord.

Most of the data were collected by volunteer naturalists with no defined sampling protocol, so are best described as "opportunistic".

# the data can be accessed from the W drive accessbible to members of the BRC  
  
dat <- read.csv("W:/PYWELL\_SHARED/Pywell Projects/BRC/\_BRC\_dataflow/Research Datasets/Soldierflies/2022/Research dataset/Soldierflies\_2022.csv")  
  
# available fields   
  
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"

# some information on the provenance of the data  
  
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). This step enables the analysis of the dataset as a whole.

# Colin Harrower's BRCmap package is needed for processing the data (https://github.com/colinharrower/BRCmap)  
  
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, the package used for analysis downstream)  
  
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  
  
any(coords$PROJECTION == "OSNI")

## [1] TRUE

# 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)  
   
}  
  
# 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 (visits to the same site in the same year but on different dates)  
  
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 because 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, I assumed that all species were searched for where and when at least one species was recorded.

## 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 grid cells with at least one record, and grid cells that were visited on more than one date in the same year, are nonrandomly distributed in all years.

mask <- raster::raster("W:/PYWELL\_SHARED/Pywell Projects/BRC/Rob Boyd/TSDA/SDMs/Data/SDMOutputs\_Jan\_Feb\_2021/Bryophytes/Bry\_986\_LPT\_1.asc") # mask layer needed to delimit the geographic domain. This is just a raster of the UK at 1km  
  
# 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

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))

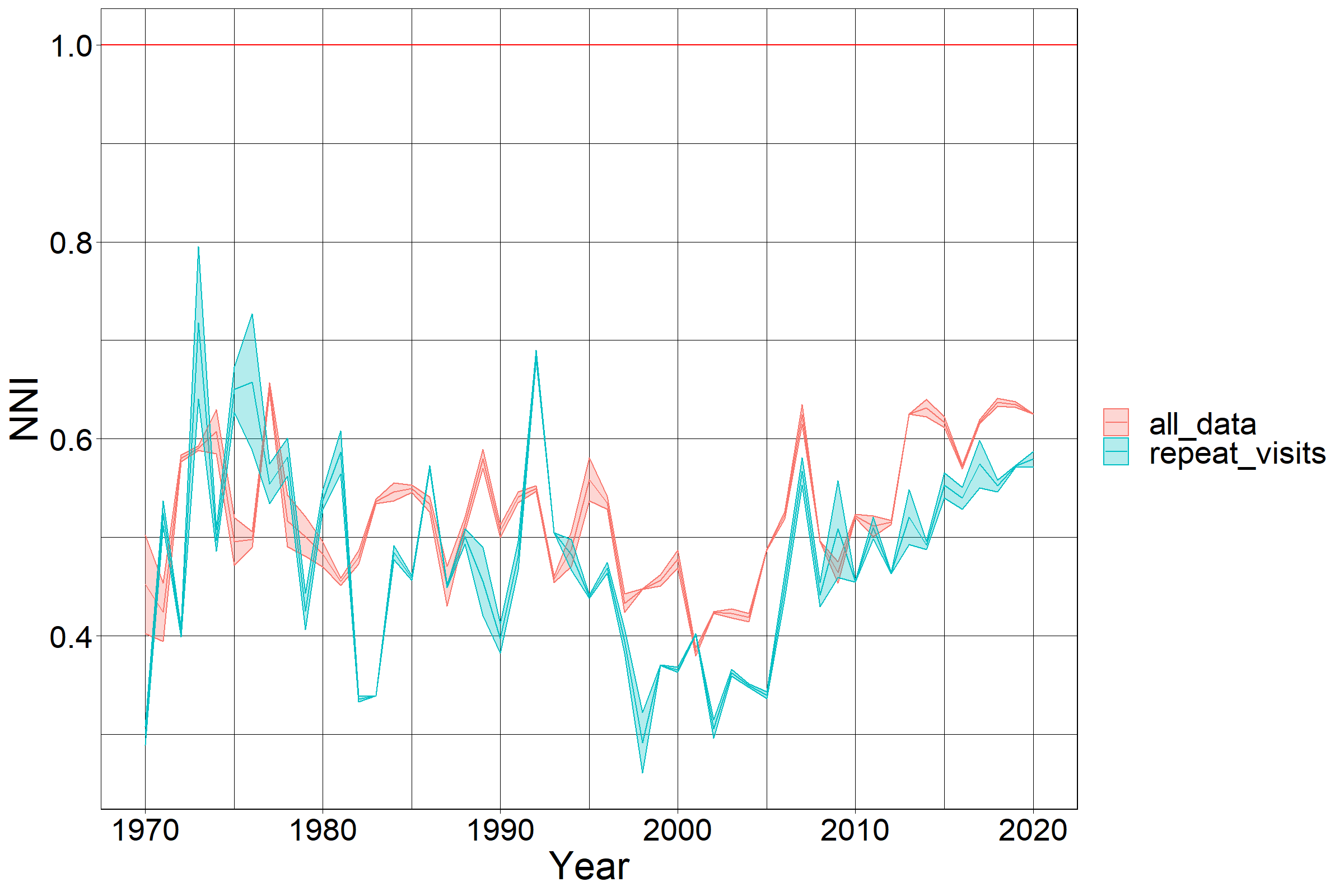


Figure 1. Nearest neighbour index calculated for each year. The shaded band denotes the 5th to 95th percentiles calculated by bootstrapping over five simulated distributions.

Fig. 2 shows the geographic distribution of 1km grid cells 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)  
  
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))

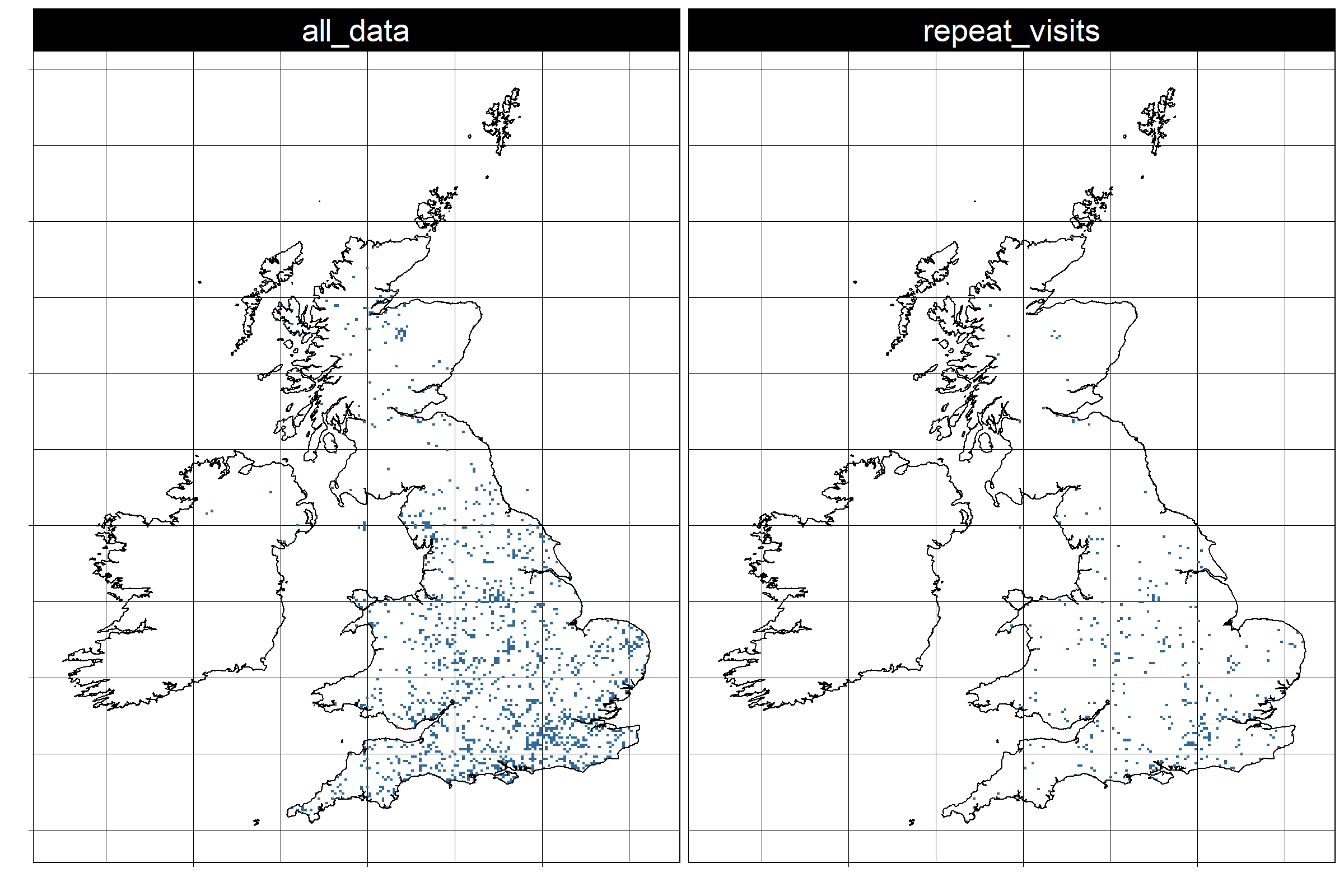


Figure 2. 1km grid cells in which at least one species was recorded between 1990 and 2020.

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

To establish whether the data have been sampled from the same portion of geographic space over time, I mapped the proportion of the 51-year time series in which grid cells were sampled, and the proportion of years in which they were visited more than once (Fig. 3). The maps clearly show that the most grid cells were sampled in a very low proportion of years; indeed, across grid cells with at least one record, the average proportion of years with records is less than 4%. This indicates that the geographic distribution of sampling has changed over time, which has the potential to confound estimates of temporal trends in species' range sizes.

spatCov <- occAssess::assessSpatialCov(periods = periods,  
 dat = dat,  
 species = "recommended\_name",   
 year = "year",  
 identifier = "identifier",  
 x = "EASTING",   
 y = "NORTHING",  
 spatialUncertainty = "spatialUncertainty",  
 res = 1000,  
 output = "nPeriods",  
 logCount = FALSE,  
 returnRaster = TRUE)  
  
  
# plot   
  
rasterVis::gplot(spatCov$rasters) +  
 ggplot2::geom\_tile(ggplot2::aes(fill = value / 51)) +  
 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::scale\_fill\_gradient2(low = "blue", mid = "green", high = "red", midpoint = 0.4, na.value = myCol) +   
 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::theme(strip.text.x = ggplot2::element\_text(size = 20))

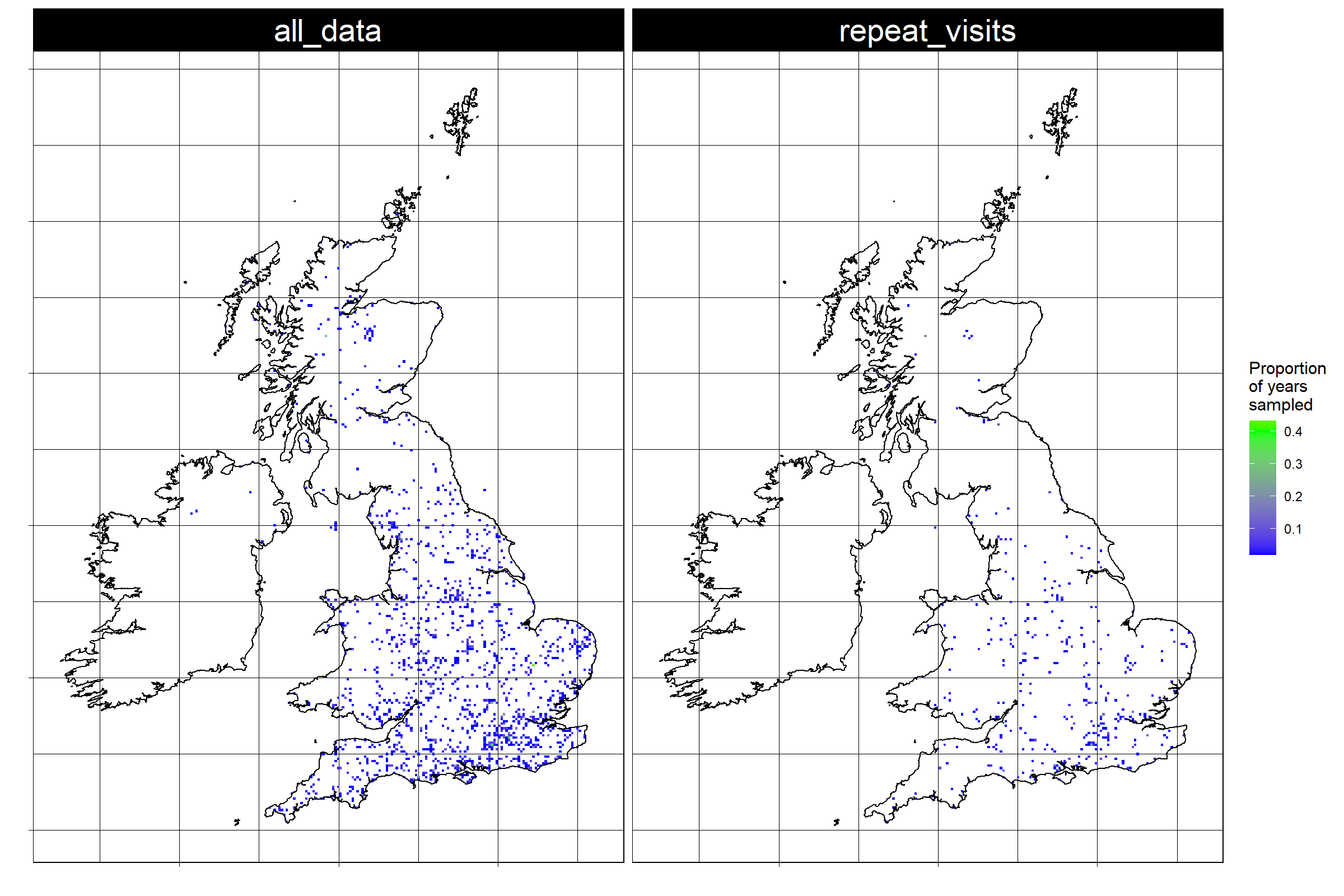


Figure 3. The proportion of the 51-year time series in which a record was made in each grid cell (left panel) and in which each grid cell was visited more than once (right panel).

# calculate the mean proportion of years in which grid cells were visited once and visited more than once   
  
mean(raster::getValues(spatCov$rasters$all\_data), na.rm=T) / 51

## [1] 0.035112

mean(raster::getValues(spatCov$rasters$repeat\_visits), na.rm=T) / 51

## [1] 0.031805

**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.**

Mitigating the geographic biases described above will require susbtantial effort. New analytical approaches and forms of data manipulation will need to be tested. In the meantime, we should caveat any trends derived from these data. This should include an acknowledgement that the data are not representative of the UK as a whole-just England- and that information on detectability from repeat vists comes from an even smaller subset of locations, also mainly in England.

## Environmental domain

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

In this section I began by defining environmental space based on the standard 19 worldclim bioclimatic variables (e.g. temperature of the wettest month, isothermality, etc.). These data were originally derived from the [Met Office's HadUK grid dataset](https://www.metoffice.gov.uk/research/climate/maps-and-data/data/haduk-grid/datasets). I randomly placed 10,000 points across the UK, and then extracted the values of the 19 bioclimatic variables at each of their locations. I then conducted a principal component analysis of these data, and defined environmental space in the UK as their joint distribution on the first two principal components. I then mapped the distribution of the data from each year in environmental space (ellipses in Fig. 4). It is clear that the data are not representative of environmental space in the UK; however, what is less clear is whether this reflects a bias, or whether it reflects the fact soldierflies and allies are truly concentrated in that portion of environmental space.

# load bioclim data (usual 19 variables at 1km in the UK)  
  
biovars <- raster::stack(  
 list.files("W:/PYWELL\_SHARED/Pywell Projects/BRC/Rob Boyd/TSDA/SDMs/Data/bioVars/",  
 full.names = T,  
 pattern = ".asc"))  
  
# extract bioclim vars in grid cells with records  
  
envDat <- raster::extract(biovars, dat[, c("EASTING", "NORTHING")])  
  
# create a "background" sample for comparison: this is supposed to represent climatic space across the UK  
  
backgroundEnvDat <- raster::sampleRandom(biovars, size = 10000,  
 xy = F)  
  
envBias <- occAssess::assessEnvBias(dat = dat,  
 species = "recommended\_name",   
 year = "year",  
 identifier = "identifier",  
 x = "EASTING",   
 y = "NORTHING",  
 spatialUncertainty = "spatialUncertainty",  
 envDat = envDat,  
 backgroundEnvDat = backgroundEnvDat,  
 xPC = 1,  
 yPC = 2,  
 periods = periods) # xPC and yPC indicate which principal components to set as the x and y axes,respectively  
  
envBias$plot +  
 ggplot2::guides(fill = "none",  
 colour = "none") +  
 ggplot2::theme(strip.text.x = ggplot2::element\_text(size = 20))

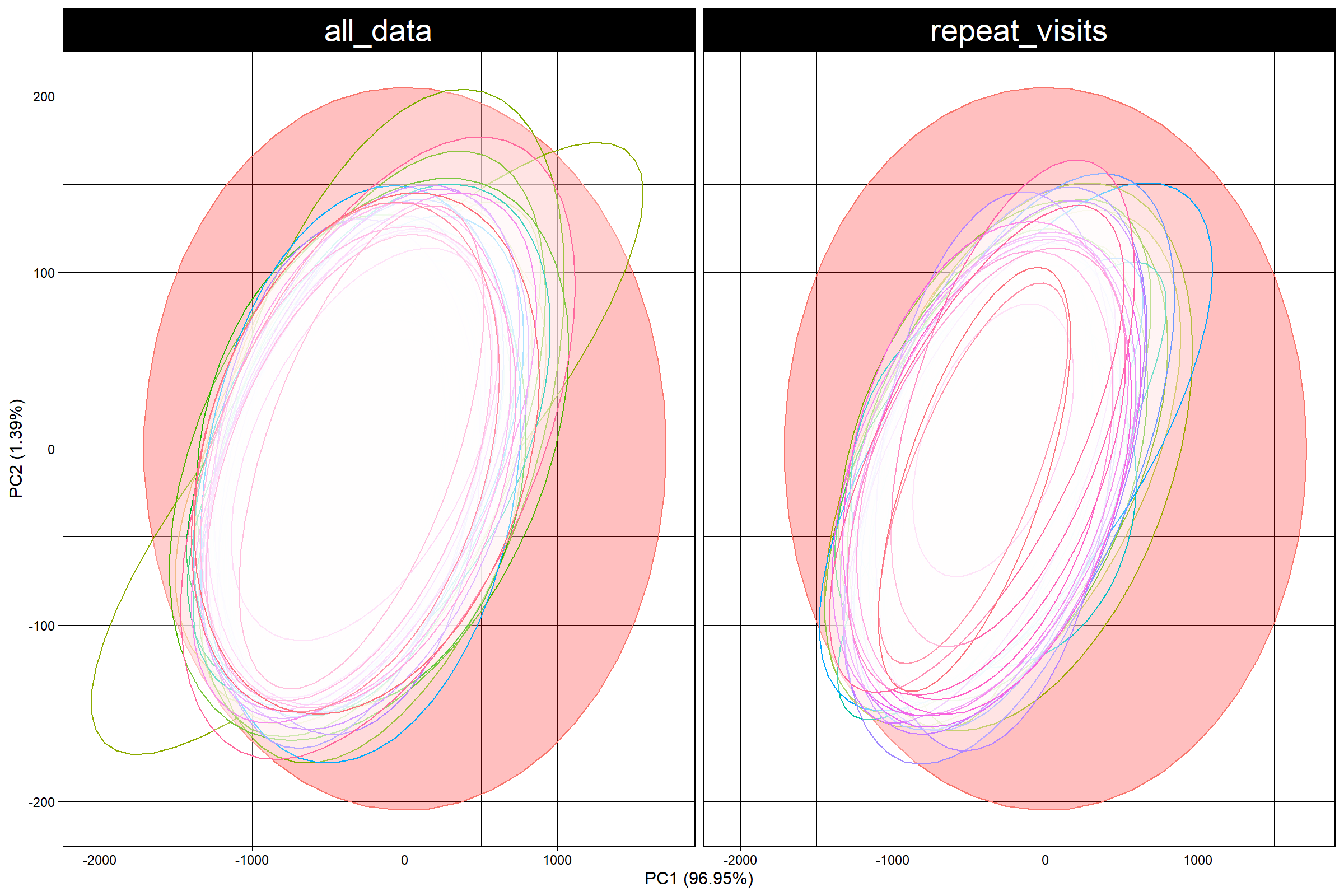


Figure 4. The distribution of records in environmental space. Environmental space is defined as the first two principal components of the standard 19 bioclimatic variables at 10,000 randomly-selected locations in the UK. The data from each year are shown as one ellipse for ease of interpretation, but note that, strictly speaking, using ellipses in this way is most appropriate where the data are normally distributed along both axes, which I did not check.

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

Fig 4. shows the distribution of the data in climate space in each year. Although the data were not collected in a representative portion of environmental space, they were, generally, collected from the same portion of climate space across years.

Whilst the data appear to have been sampled from a reasonably consistent portion of climate space over time, there are other features of the environment to consider. It widely suspected, for example, that recording has shifted towards urban centres over time. To test this I extracted the proportion of each 1km grid cell in the UK that was classed as "built up" in 2007 from UK CEH's land cover map (Morton et al., 2007). I then extracted these proportions for grid cells that were visited once and more than once in each year and plotted them in Fig. 4. Two things stand out: 1) there has been a striking increase in urban recording since ca. 2011, and 2) cells visited more than once tend to be more built up than those visited once only, across all years.

# load raster showing proportion of each 1km cell was is "built up" in 2007  
  
lc <- raster::raster("W:/PYWELL\_SHARED/Pywell Projects/BRC/Rob Boyd/TSDA/SDMs/Data/targetLandCover2007/selectedCovs/built\_up.asc")  
  
# for each record add the proportion of the cell on which it was recorded that was built up in 2007  
  
dat$built\_up <- raster::extract(lc,   
 cbind(dat$EASTING, dat$NORTHING))  
  
# plot  
  
ggplot2::ggplot(data=dat[dat$year>=1970, ], ggplot2::aes(x=year, y=built\_up, fill=identifier)) +  
 ggplot2::geom\_boxplot() +  
 ggplot2::theme\_linedraw() +  
 ggplot2::theme(axis.text.x = ggplot2::element\_text(angle = 90)) +  
 ggplot2::labs(x = "",  
 y = "Proportion of cell classed as built up",  
 fill = "")

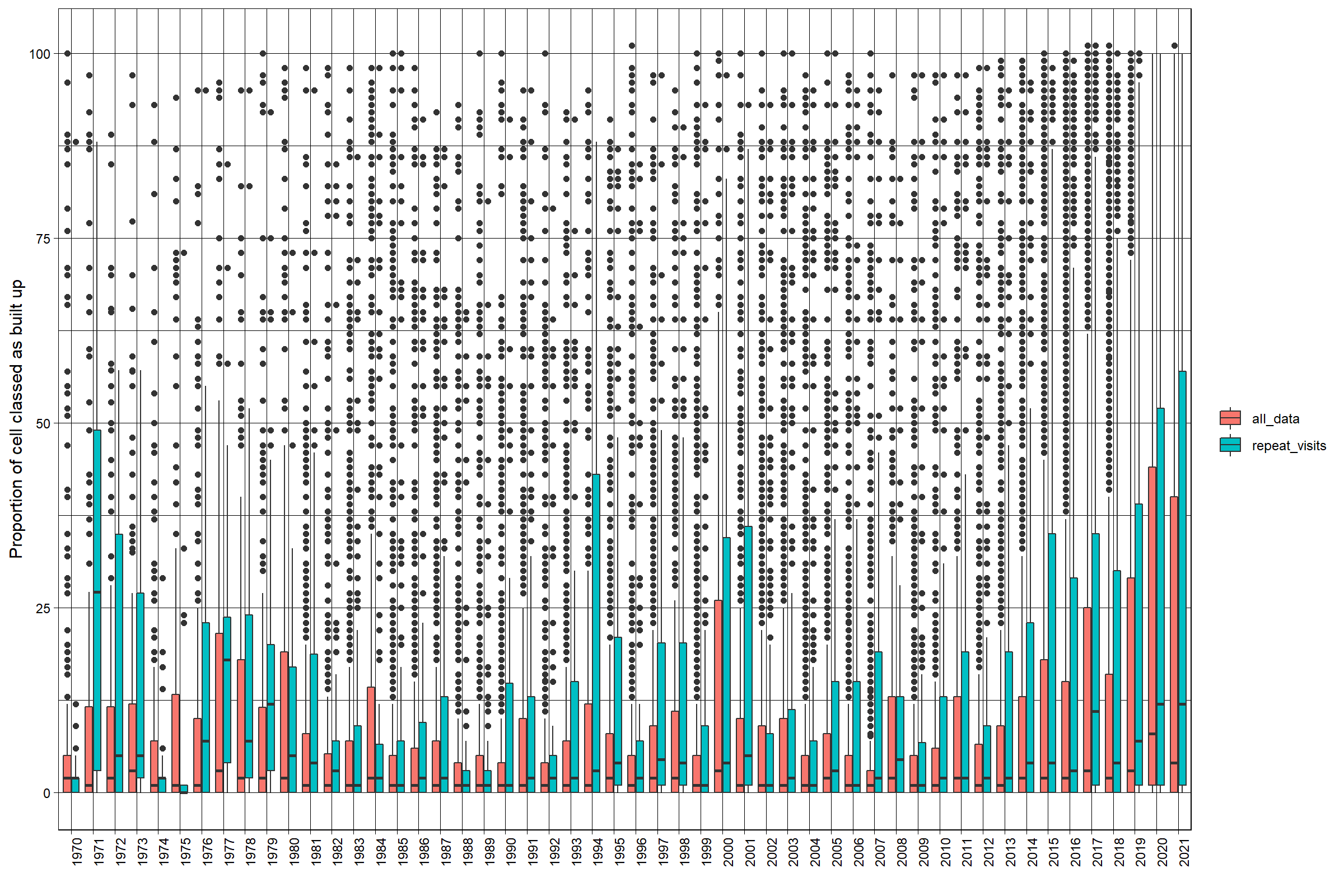


Figure 4. Extent to which cells that were visited (all\_data) and visited more than once (repeat\_vists) were built up (urban or suburban) across years. Boxes show the medians and interquartile ranges for each year; points denote outliers.

**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.**

Whilst the data were not collected from a representative portion of climate space in the UK, it is not clear whether this reflects sampling bias or the true distributions of the species. I will speak to Martin Harvey, the national expert for this taxon group, to see if there is more information that can be brought to bear on this.

What is clear is that there has been a shift towards recording in urban areas over time, including repeat visits. It is easy to imagine that species' trends and detection probabilities might be different in urban areas than elsewhere. As was the case for geographic biases, there are no quick fixes for this problem. We will need to develop new analytical approaches or forms of data manipulation. In the meantime, we should clearly acknowledge the fact that any trends derived from these data will likely increasingly reflect urban areas over time.

## 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