Supporting Information Appendix S4 - Background sampling for spatial sampling bias in citizen science data

S. Derville, L. Torres, C. Iovan, C. Garrigue

22 mai, 2018

Supporting information S4 to "Finding the right fit: Comparative cetacean distribution models using multiple data sources and statistical approaches", 2018, Diversity and Distributions. Contact information: solene.derville@ird.fr

Runs on R version 3.3.2 (2016-10-31) Platform: x86_64-pc-linux-gnu (64-bit)

The "POP" background sampling method was developped to model humpback whale habitat suitability while accounting for spatial sampling bias in the observations of cetaceans collected through a citizen science program in New Caledonia. The 500 m isobath was used to delineate waters surrounding the main lagoons and was buffered out by 15 km in order to include all Fish Aggregation Devices located off the outer edge of the barrier reef. Within this buffer, human densities were approximated using a Gaussian Kernel Density Estimate of the main cities weighted by their respective population size (source: INSEE). The resulting density values were scaled between 1 and 100 and were used as weights for the random sampling of control points. Outside the buffer zone around the lagoons, the probability of sampling was set to 1.

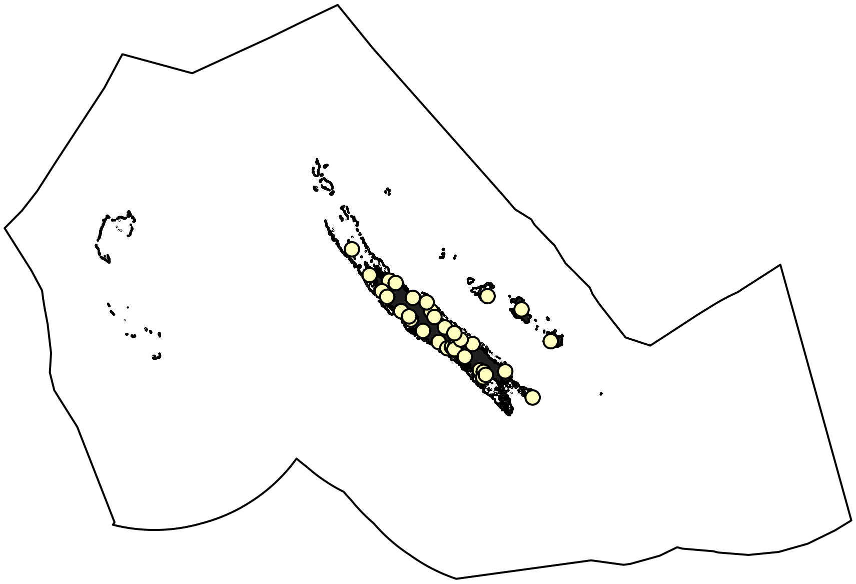
R Packages

```r
library(raster)
library(rgdal)
library(ggplot2)
library(reshape2)
library(plyr)
library(maps)
library(rgeos)
library(ks)
library(gridExtra)
library(sp)
library(geoR)
library(oce)
library(maptools)
library(pryr)
library(geosphere)
library(viridis)
```

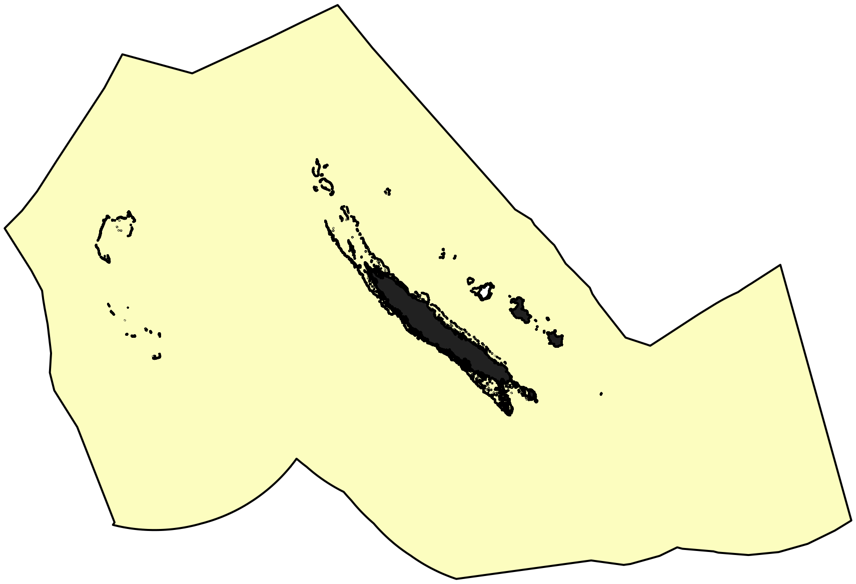
1- Defining the study area

Brief description of input data

SpatialPoint shapefile containing the position and population size of all major human settlements in New Caledonia **pop_shp_merc** (yellow points on the figure)



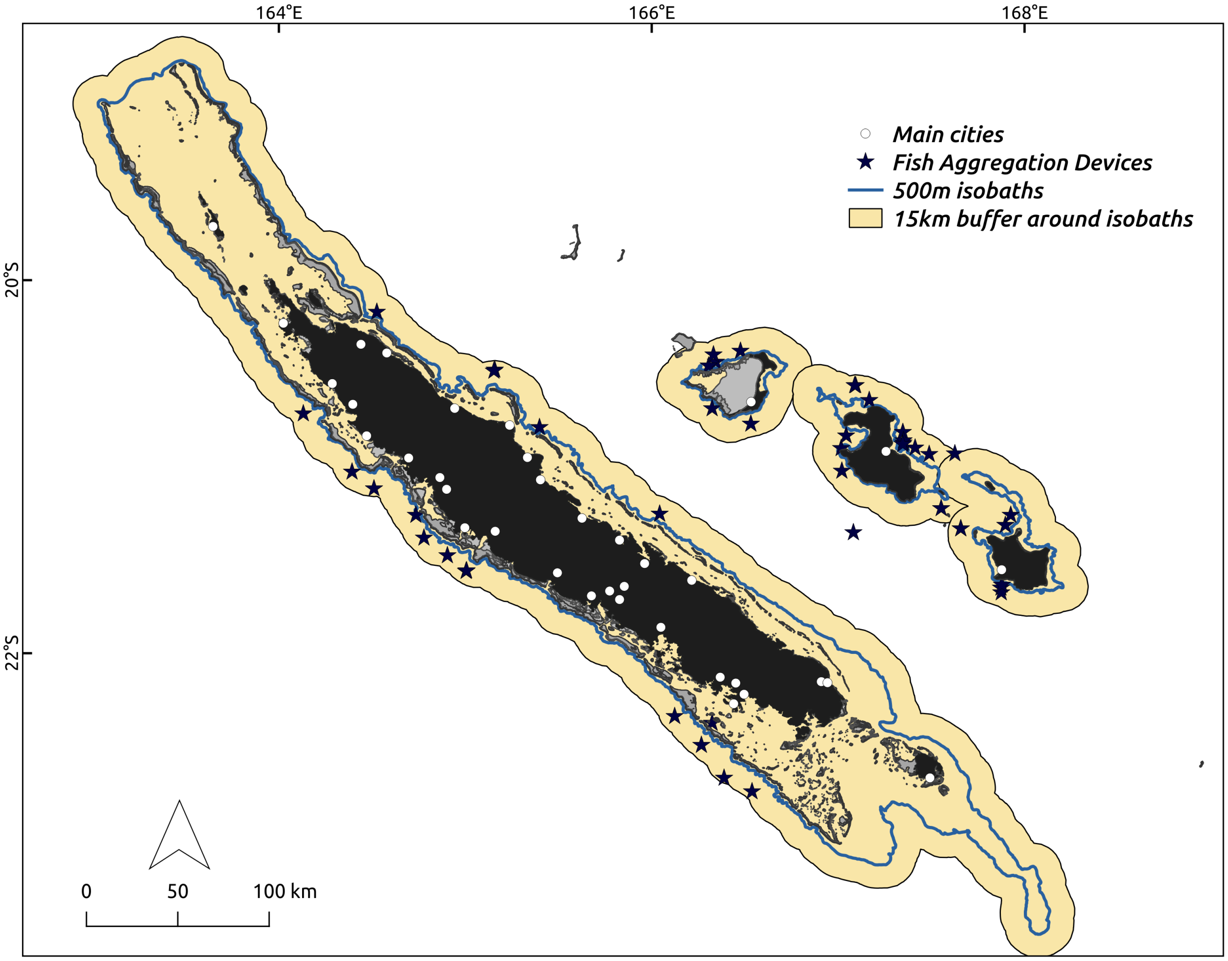
SpatialPolygon shapefile of all New Caledonian waters (excludes lands, islands, reefs and areas outside the Economic Exclusive Zone) **ocean_shp_merc**



Area of greatest human density

In the New Caledonian archipelago, lagoons surrounding the main islands are preferential areas of use for people as they concentrate activities such as recreational fishing, leisure boating, scuba diving etc. Human activities may also extend a little further than the barrier reef as several Fish Aggregation Devices are scattered around the archipelago, usually a few kilometers from the barrier.

The 500m isobath was used to delineate the lagoons around the main islands. A 15km buffer was created around this isobath in order to include Fish Aggregation Devices: the SpatialPolygon **lagoon_buf_merc** (in yellow in figure below)



2- Account for distance to cities: the POP background

Generate weighted kernels around cities and within the lagoon_buffer (isobath 500m + 15km buffer). Those kernels are scaled to [1,100]. All the rest of the eez is set to a weight.pop = 1. Kernel DEnsity Estimates are created over the locations of cities weighted by the population size of each.

Population size of cities was provided by INSEE (Institut national de la statistique et des études économiques; open-source download at https://www.insee.fr/fr/statistiques/). The latest estimates dating from 2014 were used.

```
# generate kernel with default smoothing method and weights = to population per city
grid_sp <- rasterToPoints(env_stack$B0.5km_ras, spatial=T)
k <- ks::kde(as.matrix(coordinates(pop_shp_merc)), eval.points = grid_sp@coords,
             w = pop_shp_merc@data$pop)

# convert to raster format
grid <- data.frame(x = k$eval.points[, 1], y = k$eval.points[, 2], z = k$estimate)
grid$z <- grid$z * 100 / max(grid$z, na.rm=T) # rescale to 0 - 100
grid[grid$z<1,]$z <- 1 # set everything <1 to 1
coordinates(grid) = ~ x + y
proj4string(grid)=CRS("+proj=merc +lon_0=165 +lat_1=-21.5 +k=1 +x_0=0 +y_0=0
                      +ellps=WGS84 +datum=WGS84 +units=m +no_defs")
```

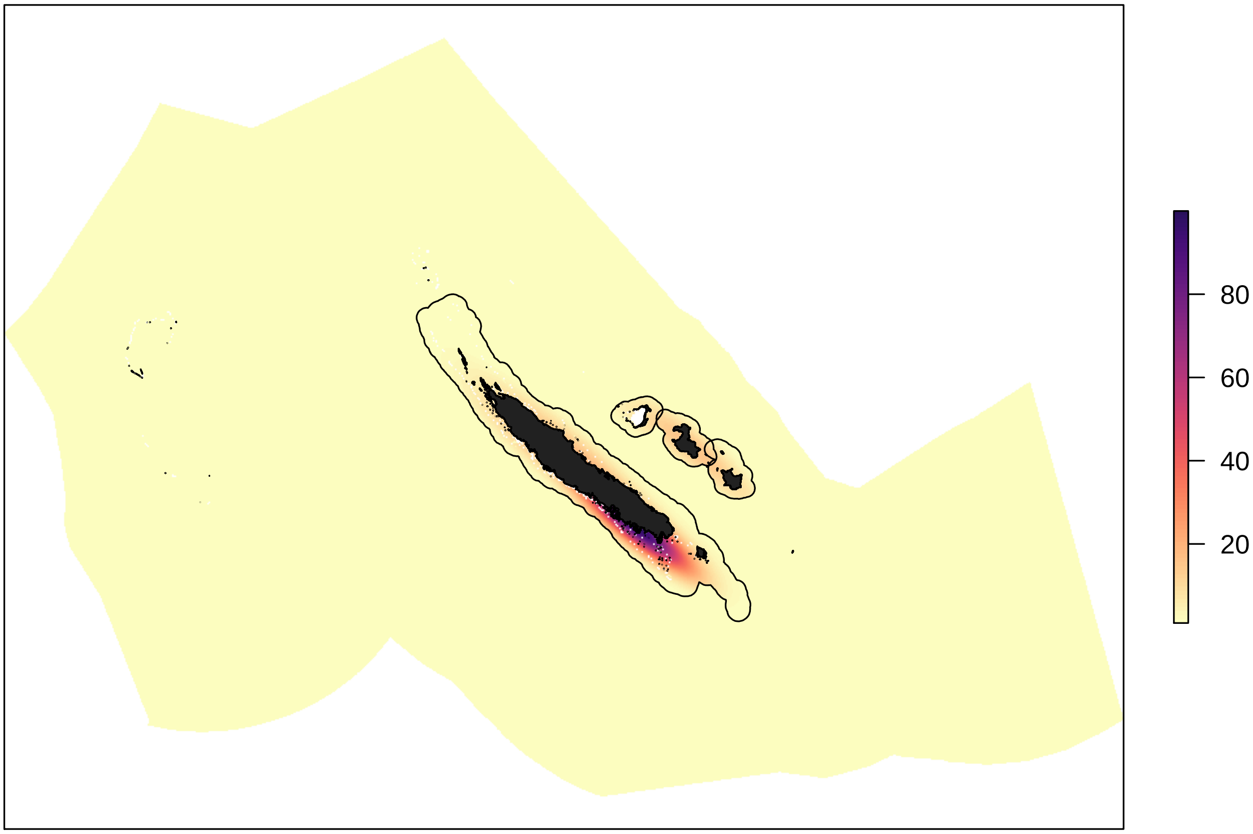
```
grid=SpatialPixelsDataFrame(grid, data=grid@data, tolerance=0.01)
grid=raster(grid)

# generate a full grid of 1 to make the background of the eez outside the lagoon_buf
grid_1 <- grid ; grid_1[] <- 1 # create grid of ones
grid_lagoon <- mask(grid, lagoon_buf_merc, inverse=F) # crop grid with kernel to lagoon

# overlay the grid_lagoon on top of grid_1
access_ras <- cover(grid_lagoon, grid_1)

# mask to NA outside eez and on land/reef
pop_ras <- mask(access_ras, ocean_shp_merc, inverse=F)
names(pop_ras) <- "weight.pop"
```

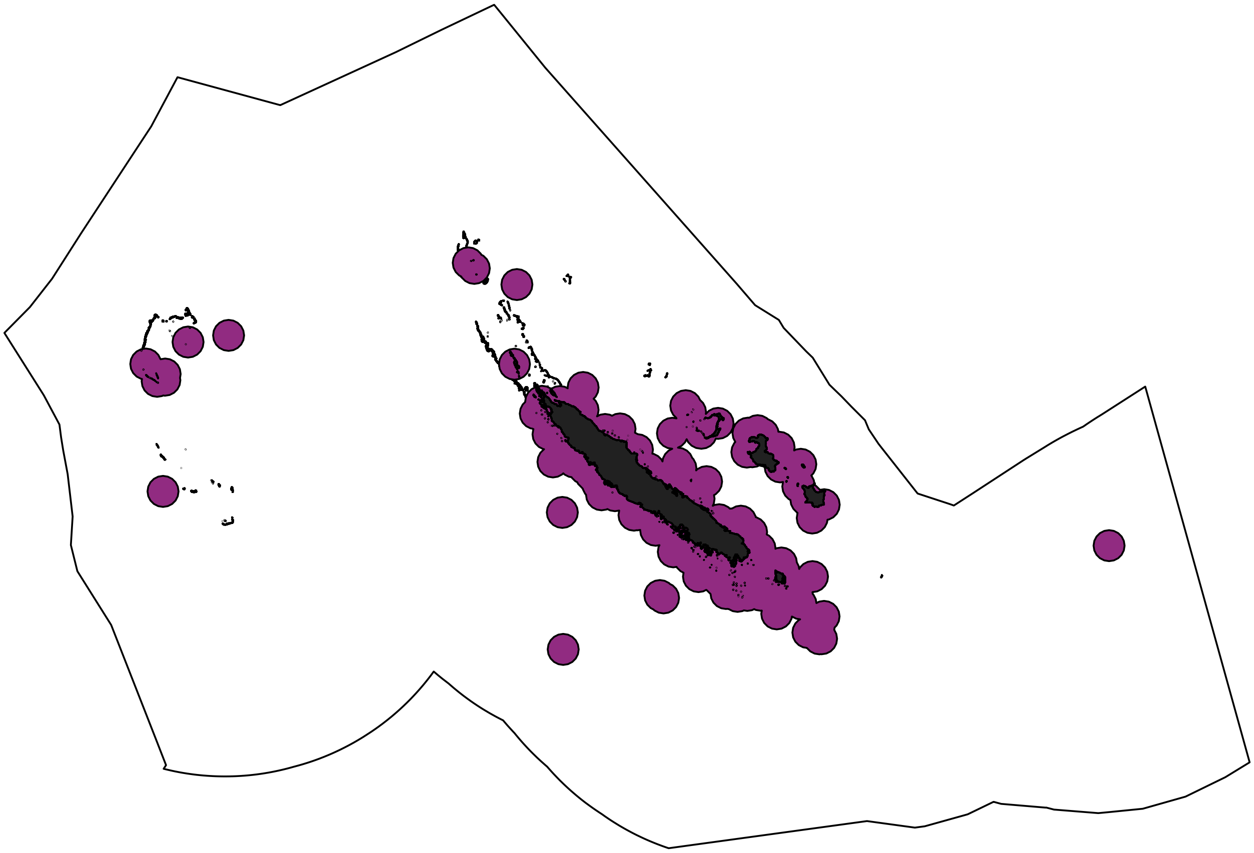
This results in a raster representing human occupation (colour scale from pink to purple). This is the POP background used to generate control points for modeling humpabck whale habitat preferences.



2- Account for observations: the TARGET background

```
# load the full dataset containing all crowdsourced sightings of marine mammals in NC
sig_sp <- sig_df_target
# convert to a spatial format and project
coordinates(sig_sp) =~ sig_X + sig_Y
proj4string(sig_sp) <- CRS("+proj=longlat +ellps=WGS84 +datum=WGS84 +no_defs")
sig_sp <- spTransform(sig_sp, CRS("+proj=merc +lon_0=165 +lat_1=-21.5 +k=1
                          +x_0=0 +y_0=0 +ellps=WGS84 +datum=WGS84 +units=m +no_defs"))
# create 25 km-radius circle buffers around sightings
sig_buffer_sp <- rgeos::gBuffer(sig_sp, byid=F, width=25000, capStyle="ROUND",
                                joinStyle="ROUND")

# based on access_ras raster, create target_ras
# all pixels outside the buffers are set to 0, the ones inside are set to 100
target_ras <- access_ras
target_ras[!is.na(target_ras)] <- 100
target_ras <- mask(target_ras, sig_buffer_sp)
names(target_ras) <- "weight.target"
```

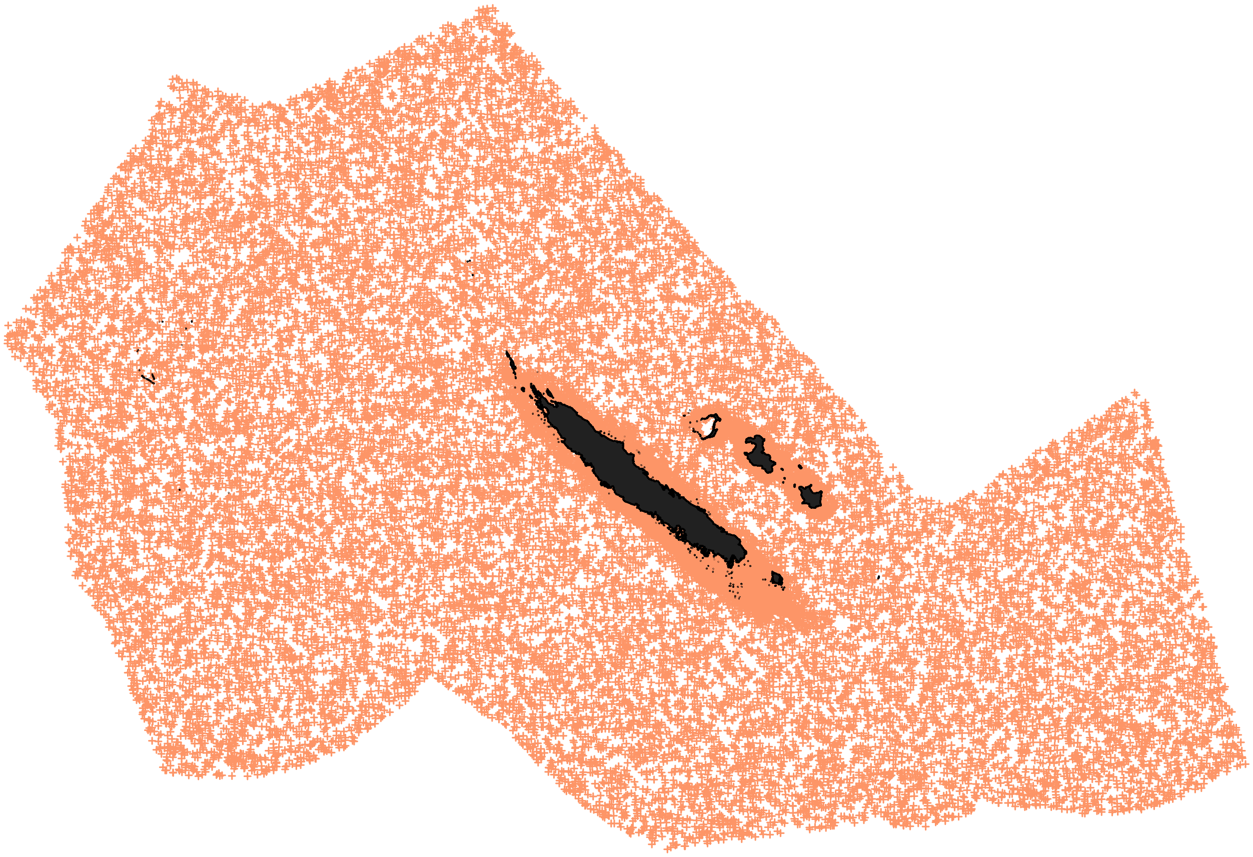


4- Sampling control points

POP SAMPLING

Control points are sampled within the POP background: the sampling is randomized but the probability of sampling a position within the background is proportional to the **weight.pop**, this is the human density index calculated above. In summary, the more populated/used an area is, the more densily it is sampled in the control point dataset.

```
universe_df_pop <- data.frame(rasterToPoints(pop_ras, spatial=F))
control_df_pop <- ddply(sig_df, .(year,month), function(d) {
  # The probability of sampling a position is proportional to the weight.pop
  s <- sample(seq(1,nrow(universe_df_pop),1), 605, prob=universe_df_pop$weight.pop)
  abs <- universe_df_pop[s, c("x","y")]
  abs$date_approx <- min(d$sig_dat)
  return(abs)
})
```



TARGET SAMPLING

Control points are only sampled inside the buffers surrounding the sightings of marine mammals recorded in New Caledonia.

```
universe_df_target <- data.frame(rasterToPoints(target_ras, spatial=F))
control_df_target <- ddply(sig_df, .(year,month), function(d) {
  s <- sample(seq(1,nrow(universe_df_target),1),39,prob=rep(1, nrow(universe_df_target)))
  abs <- universe_df_target[s, c("x","y")]
  abs$date_approx <- min(d$sig_dat)
  return(abs)
})
```