Using the sampbias R package

#Installing sampbias To install the *sampbias* R-package, you can either download the latest build version from gitHub use the `devtools` package:

```
require(devtools)
install_github(repo = "azizka/sampbias")
library(sampbias)
library(maptools)
```

```
## Checking rgeos availability: TRUE
```

#Input data *Sampbias* calculates spatial bias in a species occurrence data set based on two input files:

1. A table of Species occurrence records
2. A set of geographical gazetteers

The example files for this tutorial are provided in the example_data folder and should be copied to the R working directory, before starting. You can find descriptions and help for all functions by typing `?Functionname`, for instance `?SamplingBias`.

##Species occurrence records Species occurrences can be provided to *sampbias* as an R `data.frame`, which can be loaded from any table format .txt file. The minimum input file needed to run *sampbias* is a table with species occurrences including three columns named "species", "decimallongitude", "decimallatitude". *Sampbias* can also directly work with data downloaded from the Global Biodiversity Information Facility data portal (GBIF). We will use the records of mammals from Borneo provided in the sampbias folder (GBIF, 2016) as example file throughout the tutorial. The file is also included as example data set in the package.

Text files can easily be loaded into R as data.frame using the `read.csv()` function.

```
#loading a text file to R
occ <-read.csv(system.file("extdata", "mammals_borneo.csv", package="sampbias"), sep = "\t")
```

Note: *Sampbias* evaluates the bias by comparing to a random sampling, meaning that the tool is not designed for single species data sets, as distribution of the records might then reflect ecological preferences, but rather for multi-species data sets. In general, the more species and the more records, the more reliable the results will be.

##Geographic gazetteers *Sampbias* evaluates the distribution of the sampled species occurrences in relation to geographic features that might bias sampling effort. These are generally related to accessibility or means of travel. *Sampbias* includes a set of default gazetteers for cities, airports, roads and rivers (Natural Earth, 2016), which can give fair estimates of bias for large and medium scale analyses. These are used by default, if no other gazetteers are provided by the user. However, these defaults include major features only and if available high resolution user-provided gazetteers are preferable.

Any number of gazetteers can be provided to *sampbias* via the `gaz` argument as a list of `SpatialPointsDataFrame` and `SpatialLinesDataFrame` objects. These objects can easily be loaded into R from standard shape files using the `maptools` package:

```
cit <- readShapeSpatial(system.file("extdata", "Borneo_major_cities.shp", package="sampbias"))
```

```
## Warning: readShapeSpatial is deprecated; use rgdal::readOGR or sf::st_read

## Warning: readShapePoints is deprecated; use rgdal::readOGR or sf::st_read
```

```
roa <- readShapeSpatial(system.file("extdata", "Borneo_major_roads.shp", package="sampbias"))
```

```
## Warning: readShapeSpatial is deprecated; use rgdal::readOGR or sf::st_read

## Warning: readShapeLines is deprecated; use rgdal::readOGR or sf::st_read
```

```
gazetteers <- list(cities = cit,
                   roads = roa)
```

See here and `?SpatialPointsDataFrame` or `?SpatialLinesDataFrame` on how to create a SpatialObjects from tables of coordinates.

#Running an analysis A sampbias analyses is run in one line of code via the `SamplingBias` function:

```
bias.out <- SamplingBias(x = occ, gaz = gazetteers)
```

```
## Adjusting to terrestrial surface... Done.
## Creating occurrence raster... Done
## Creating species raster... Done
## Calculating distance raster...

## Warning in DisRast(gaz = gaz, ras = dum.ras, buffer = buffer, ncores =
## ncores): Evening buffer. Buffer set to 2

##  Done
## Extracting values...

## Warning in .DisVect(x = dat.pts, dist = dis.ras, convexhull = convexhull, :
## Low resolution, increase resolution or switch to absolute distance
## calculation

##  Done
## Calculating likelihood...
## rescale factor: 1750000 350000
## [1] 99.99993
##
## rescale factor: 750000 150000
## [1] 99.99993
##  Done
## Preparing output... Done
```

In addition to the input from above, `SamplingBias` offers a set of options to customize analyses, of which the most important ones are shown in Table 1. See `?SamplingBias` for a detailed description of all options.

Option	Description
res	the raster resolution for the distance calculation to the geographic features and the data visualization in decimal degrees. The default is to one degree, but higher resolution will be desirable for most analyses. Res together with the extent of the input data determine computation time and memory requirements.
convexhull	logical. If TRUE, the empirical distribution (and the output maps) is restricted to cells within a convex hull polygon around `x`. If FALSE a rectangle around `x` is used. Default = FALSE.
terrestrial	logical. If TRUE, the empirical distribution (and the output maps) are restricted to terrestrial areas. Default = TRUE.
biasdist	numerical. A vector indicating the distance at which the average bias should be calculated for the output table (in meters). Can also be a single number. Default = c(0, 10000).

#Output The output of `SamplingBias` is a list of different R objects.

Object	Description
summa	A list of summary statistics for the sampbias analyses, including the total number of occurrence points in x, the total number of species in x, the extent of the output rasters as well as the settings for res, binsize, and convexhull used in the analyses.
occurrences	a raster indicating occurrence records per grid cell, with resolution res.
species	a raster with indicating the number of species per grid cell, with resolution res.
biasmaps	a list of rasters, with the same length as gaz. Each element is the spatial projection of the bias effect for a sources of bias in gaz. The last raster in the list is the average over all bias sources.
biastable	a data.frame, with the estimated bias effect for each bias source in gaz, at the distances specified by biasdist.

The bias effect is the fraction of samples expected to be missed at a given distance relative to distance 0. A high bias effect means a high fraction of missed records. The fractions are comparable between sources of bias and different data sets and the bias at any given distance can thus inform on the relative severity of a biasing source.

The package includes summary and plot methods for an easy exploration of the results:

```
summary(bias.out)
```

```
##             Length Class       Mode
## summa         6    -none-      list
## occurrences 110    RasterLayer S4
## species     110    RasterLayer S4
## biasmaps      3    -none-      list
## biastable     2    data.frame  list
```

```
#plot(bias.out)
```

The plot function creates a set of publication level plots using the `ggplot2` package, including the number of occurrence per gridcell, the number of species per gridcell, a visualization of the bias for sources provided via the gazetteers in `gaz` and the average bias over all sources:

```
#plot(bias.out, gaz = gazetteers)
```