LPR BE: Steps for data preparation for SOM

(Modified from Hans Van Calster 15 february 2019)

## Data preparation steps

* 1. **Data Filtering:** Per dataset

various checks for, e.g.

* + invalid values
  + missing data
  + incomplete data (incl. data where location not precise enough to allocate to a 1x1 site)
  + Add information on covariates when they are judged relevant (e.g: length of species list, observer skill level,..)
  + exclusion of uncertain identifications
  + exclusion species that require targeted search efforts (i.e. when non-detection cannot be determined from data)
  + standardizing date-time formats into: 3 fields with numbers only

|  |  |  |
| --- | --- | --- |
| DD | MM | YYYY |

* 1. **Data Preparation:**

1.2.1 Per species group:

* + Harmonisation of species names
  + Harmonization of spatial reference formats
  + Harmonization of field names/headers
  + Addition of data source name
  + Filtering for duplicates
  + Determination of closure period for each species (= period in the year during which the site is assumed to be either occupied or unoccupied, but does not become permanently abandoned or colonized – for insects, reptiles and amphibians only).
  + Exclude observation that fall outside the closure period for a species

1. 2. 2 Once for all:
   * quantization of the data to 1 km2 grid squares

given that this needs to be done for multiple data sources, this requires possibly procedures for deciding how to assign a transect, a GPS coordinate with uncertainty, an unaligned grid square, a polygon, …, to (possibly more than one) (a) grid square(s)

* 1. **Species Selection**
* Compilation of day lists for opportunistic data (= all species from a species group observed at a particular 1 km x 1 km site on a particular day by the **any** observer)
* Defining study period (start year - end year) for each species group (or species)
  + exclude observations that fall outside the study period
    - Apply exclusion/inclusion criteria per species

*for instance, if a species only started appearing from, say, 2004, the period from 1990 - 2003 should be discarded for that species*

Determine suitable 1 km x 1 km sites for each species = the set of sites where the species is observed at least once during the study period (or 1 or 2 species observed for each decade to avoid unstable trend estimates)

* + excluded very rare species: defined as species inhabiting less than 35 1 km x 1 km sites in the study period (we feel that this should probably be more stringent)
* Deduce non-detection (0) for each study species
  + = all cases in which a species was not on a day-list at a site in a year
  + Any observation of the species under consideration within the closure period was taken as a 1 (detection), whereas we rated 0 (nondetection) if any other species but not the species under consideration had been reported by an observer at a particular 1 km x 1 km site and on a particular date within the closure period. This procedure was repeated for all species in order to obtain detection histories for each species.
* Gather information on covariates to account for differences in detection between visits: detection probability was modelled as a function of:

**(Potential Steps for Later**)

* + Exclude migratory species
  + Exclude species with large home ranges
  + Exclude species that require targeted search efforts (because non-detection cannot be determined from the data- if not done by data provider)

Perhaps apply criteria for inclusion/exclusion for species groups:

* + for a species group there should be at least 100 sites that were also visited in the previous year and
  + a minimum of 200 sites at a given year and
  + a mean number of at least 2.5 visits per site (in Strien, Swaay, and Termaat (2013) there are about 7 replicated visits available for the monitoring data and about 3 replicated visits for the opportunistic data)
  + Exclude very rare species defined as species inhabiting less than 35 1 km x 1 km sites in the study period (we feel that this should probably be more stringent)

If judged relevant: include co-variate ‘length of species list’ (only for opportunistic data - expectation: longer species lists, higher chance of detection) by categorizing them into:

* Single record lists (only one species on the day-list) (about 50% of all lists)
* Short day-lists (2-3 species day-lists) (about 25% of all lists)
* Comprehensive day-lists (> 3 species day-lists) (about 25% of all lists)