Data preparation steps for WWF LPR assessment of grasshoppers and crickets

Hans Van Calster, Els Lommelen

2019-09-18

Received datasets

The following datasets were received:

- Natuurpunt / Natagora: data from waarnemingen.be / observations.be; grasshoppers and crickets; Belgium
- DEMNA: grasshoppers and crickets; Walloon region
- Saltabel: grasshoppers and crickets; Belgium

Data cleaning on separate datasets

Natuurpunt / Natagora

Filtering

- We remove duplicate rows (copied observations, i.e. two observers where the second copied from the first)
- We exlude lifestages: EGG, EXUVIAE and all those relating to LARVAE
- We removed the following variables: authority, soortid_waarnemingen, levensstadium, gedrag_methode, lifestage, behaviour_method
- In case every variable has the same values except a different number of individuals counted, we keep only the maximum number of individuals observed per day.

Harmonisation of species names

All species names are parsed via GBIF and checked against the GBIF taxonomic backbone. Synonyms are replaced by the accepted species name. Note that the Catalogue of Life has 96% overlap with the GBIF taxonomic backbone.

```
## [1] "All column names present"
```

```
## Registered S3 method overwritten by 'crul':
##   method                 from
##   as.character.form_file httr
```

All names could be parsed.

The following names are synonyms in the GBIF taxonomic backbone:

naam_lat	species	rank	synonym
Conocephalus discolor	Conocephalus fuscus	SPECIES	TRUE

The species_id is taken from the field speciesKey (which resolves synonyms).

We keep only species that have matchType EXACT or FUZZY and which are of rank SPECIES.

Harmonisation of spatial reference formats

The dataset contains the following spatial information:

- x_coordinate and y_coordinate: decimaldegree in WGS84 format
- geographic_uncertainty: positional uncertainty in meters (the delivered data were already filtered to include only records where the geographic_uncertainty was less than 1000 meters)

We proceed as follows:

- determine the UTM1 square that intersects with the coordinates
- map the UTM1 square to the corresponding EEA 1 km x 1 km reference grid square
- aggregate the data (multiple observations of the same species, ..., same square, on the same day will be aggregated to the maximum number of indivuals counted)

The mapping of the UTM1 square to the corresponding EEA 1 km x 1 km reference grid square was calculated in advance. To do this, we calculated the centroid of each UTM1 square in ETRS89-LAEA coordinates and determined the reference grid square that intersected with that centroid.

Harmonisation of field names

```
## Observations: 85,131
## Variables: 8
## $ species_id <int> 1716462, 1681247, 1681247, 1722903, 1718308, 171075...
## $ year       <dbl> 2000, 2000, 2000, 2000, 2000, 2000, 2000, 2000, 200...
## $ month      <dbl> 4, 5, 5, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, ...
## $ day        <int> 11, 12, 12, 6, 8, 13, 13, 13, 13, 13, 13, 14, 17, 1...
## $ julian_day <dbl> 102, 133, 133, 158, 160, 165, 165, 165, 165, 165, 1...
## $ site_id    <chr> "1kmE4012N3115", "1kmE3960N3134", "1kmE3961N3135", ...
## $ count      <int> 1, 10, 10, 30, 1, 3, 3, 5, 10, 1, 1, 10, 1, 1, 3, 1...
## $ source     <chr> "waarnemingen.be/observations.be", "waarnemingen.be...
```

Saltabel

Filtering

GBIF data: https://www.gbif.org/dataset/76cc7230-76b6-4763-9caf-22626b29c0a6

Downloaded the DwC-Archive direct from IPT: https://ipt.inbo.be/resource?r=saltabel-occurrences

DOI: https://doi.org/10.15468/1rcpsq

Citatie:

Adriaens T, Decleer K, Devriese H, Lock K, Lambrechts J, San Martin y Gomez G, Piesschaert F, Maes D, Brosens D, Desmet P (2013): Saltabel - Orthoptera in Belgium. v5.3. Research Institute for Nature and Forest (INBO). Dataset/Occurrence. https://doi.org/10.15468/1rcpsq

This dataset contains centroids of UTM5 (5 km) and UTM1 (1 km) and a few more precise data. The data are in WGS84 CRS. The UTM1 and more precise observations can be filtered via the coordinateUncertaintyInMeters field. This field gives the radius of a circle containing the UTM square.

- We removed data at UTM5 resolution
- We removed all data that are not at day resolution and data before 1990 (definitely not enough data for SOM).
- Most records do not contain information on sex and / or lifestage, so excluding these variables.
- We replaced NA values in individualCount by 1 (since the species is recorded, at least one individual is seen).
- In case every variable has the same values except a different number of individuals counted, we keep only the maximum number of individuals observed per day.

Harmonisation of species names

All species names are parsed via GBIF and checked against the GBIF taxonomic backbone. Synonyms are replaced by the accepted species name. Note that the Catalogue of Life has 96% overlap with the GBIF taxonomic backbone.

We changed spelling of one *Gomphocerripus rufus* into *Gomphocerippus rufus* (because GBIF database contains incorrect spelling, the species would not parse using the correct spelling)

```
## [1] "All column names present"
```

Al species names could be parsed.

The following names are synonyms in the GBIF taxonomic backbone:

naam_lat	species	rank	synonym
Chorthippus parallelus	Pseudochorthippus parallelus	SPECIES	TRUE
Metrioptera bicolor	Bicolorana bicolor	SPECIES	TRUE
Chorthippus montanus	Pseudochorthippus montanus	SPECIES	TRUE
Metrioptera roeselii	Roeseliana roeselii	SPECIES	TRUE
Conocephalus discolor	Conocephalus fuscus	SPECIES	TRUE

The species_id is taken from the field speciesKey (which resolves synonyms).

We keep only species that have matchType EXACT or FUZZY and which are of rank SPECIES.

Harmonisation of spatial reference formats

The dataset contains the following spatial information:

- decimalLongitude and decimalLatitude: decimaldegree in WGS84 format
- coordinateUncertaintyInMeters: positional uncertainty in meters

We proceed as follows:

- determine the UTM1 square that intersects with the coordinates
- map the UTM1 square to the corresponding EEA 1 km x 1 km reference grid square
- aggregate the data (multiple observations of the same species, . . . , same square, on the same day will be aggregated to the maximum number of indivuals counted)

The mapping of the UTM1 square to the corresponding EEA 1 km x 1 km reference grid square was calculated in advance. To do this, we calculated the centroid of each UTM1 square in ETRS89-LAEA coordinates and determined the reference grid square that intersected with that centroid.

Harmonisation of field names

```
## Observations: 20,999
## Variables: 8
## $ species_id <int> 1718308, 1681124, 1681247, 1681124, 1681124, 168124...
## $ year       <dbl> 1990, 1990, 1990, 1990, 1990, 1990, 1990, 1990, 199...
## $ month      <dbl> 1, 3, 3, 4, 4, 4, 4, 4, 5, 5, 5, 5, 5, 5, 5, 5, 5, ...
## $ day        <int> 1, 17, 17, 6, 18, 18, 22, 22, 1, 5, 6, 12, 15, 15, ...
## $ julian_day <dbl> 1, 76, 76, 96, 108, 108, 112, 112, 121, 125, 126, 1...
## $ site_id    <chr> "1kmE3870N3088", "1kmE3856N3078", "1kmE3879N3125", ...
## $ count      <dbl> 1, 1, 3, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 3, 1, ...
## $ source     <chr> "Saltabel", "Saltabel", "Saltabel", "Saltabel", "Sa...
```

DEMNA

Filtering

- We removed observations without precise date (month or year resolution):
- We removed "duplicate" rows. In the original database, these are not duplicates, but:
  1. Observations of different individuals of different sexes the same date in the same location
  2. Observations of different individuals of the same species at the same date in the same UTM square but at different locations
- We removed observations before 1980 (definitely too few) and those from 2019 (incomplete year).
- We drop the det field (which contains an identifier for observer) and validation field (we do not filter based on the validation field because common species are never formely validated in the database - pers.comm. Yvan Barbier)
- In case every variable has the same values except a different number of individuals counted, we keep only the maximum number of individuals observed per day.

Harmonisation of species names

All species names are parsed via GBIF and checked against the GBIF taxonomic backbone. Synonyms are replaced by the accepted species name. Note that the Catalogue of Life has 96% overlap with the GBIF taxonomic backbone.

```
## [1] "All column names present"
```

Several species do not have an exact matchtype:

taxprio	rank	matchType
Chorthippus chorthippus parallelus	GENUS	HIGHERRANK
Tetrix sp.	GENUS	HIGHERRANK
Meconema sp.	GENUS	HIGHERRANK
Conocephalus sp.	GENUS	HIGHERRANK
Chorthippus sp.	GENUS	HIGHERRANK
Tettigonia sp.	GENUS	HIGHERRANK
Pholidoptera sp.	GENUS	HIGHERRANK
Myrmeleotettix sp.	GENUS	HIGHERRANK
Oecanthus sp.	GENUS	HIGHERRANK
Conocephalidae sp.	FAMILY	HIGHERRANK
Grande sauterelle verte	NA	NONE
Sauterelle verte	NA	NONE
Chorthippus s/g Glyptobotrus sp.	GENUS	HIGHERRANK
Chorthippus, sous-genre Glyptobotrus sp	GENUS	HIGHERRANK
Chorthippus s/genre Glyptobotrus sp.	GENUS	HIGHERRANK
Platycleis pennipes	GENUS	HIGHERRANK
Omocestus sp.	GENUS	HIGHERRANK
Grillon des bois	NA	NONE
Locusta sp.	NA	NONE

Keeping those that are of rank species and matchType exact or fuzzy:

Some of the remaining species are synonyms.

taxprio	species	synonym
Chorthippus parallelus	Pseudochorthippus parallelus	TRUE
Metrioptera roeselii	Roeseliana roeselii	TRUE
Metrioptera bicolor	Bicolorana bicolor	TRUE
Chorthippus montanus	Pseudochorthippus montanus	TRUE

The species_id is taken from speciesKey (which resolves synonyms).

Harmonisation of spatial reference formats

The DEMNA data contains UTM1 tags. We only need to map the UTM1 tags to the reference grid.

Harmonisation of field names

```
## Observations: 13,038
## Variables: 8
## $ species_id <int> 7792296, 1690432, 1708251, 1708459, 7792296, 779229...
## $ year       <dbl> 2011, 2011, 2011, 2011, 2011, 2011, 2011, 2010, 201...
## $ month      <dbl> 7, 7, 7, 7, 7, 7, 7, 5, 7, 7, 7, 7, 7, 7, 7, 7, 7, ...
## $ day        <int> 11, 29, 29, 11, 11, 29, 11, 30, 11, 11, 11, 11, 11,...
## $ julian_day <dbl> 192, 210, 210, 192, 192, 210, 192, 150, 192, 192, 1...
## $ site_id    <chr> "1kmE3820N3093", "1kmE3820N3093", "1kmE3820N3091", ...
## $ count      <dbl> 100, 1, 1, 10, 100, 100, 5, 2, 1, 4, 100, 100, 5, 5...
## $ source     <chr> "DEMNA", "DEMNA", "DEMNA", "DEMNA", "DEMNA", "DEMNA...
```

Combined datasets

We exclude species that are much too rare for a SOM analysis (less than 200 records in total: this is very liberal; during the actual SOM analysis phase, much more species will likely be removed). A record is here a unique combination of km-square, year, month, day and datasource.

```
## Joining, by = "species_id"
```

species_id	n	species
1699607	197	Sphingonotus caerulans
1708512	196	Chorthippus dorsatus
1708333	179	Chorthippus vagans
1701343	172	Euthystira brachyptera
1681155	124	Tetrix bipunctata
1710986	108	Stenobothrus stigmaticus
1695052	98	Barbitistes serricauda
1684006	57	Ephippiger ephippiger
1690685	31	Decticus verrucivorus
5096614	27	Ruspolia nitidula
1722299	24	Gryllodes sigillatus
9055856	18	Tetrix kraussi
1722140	7	Eumodicogryllus bordigalensis
1703254	3	Calliptamus italicus
1713034	3	Gryllus bimaculatus
1707320	2	Schistocerca gregaria
1713418	2	Locusta migratoria
1702101	1	Anacridium aegyptium
1705891	1	Acanthacris ruficornis
1714140	1	Oedaleus decorus
4423428	1	Yersinella raymondii

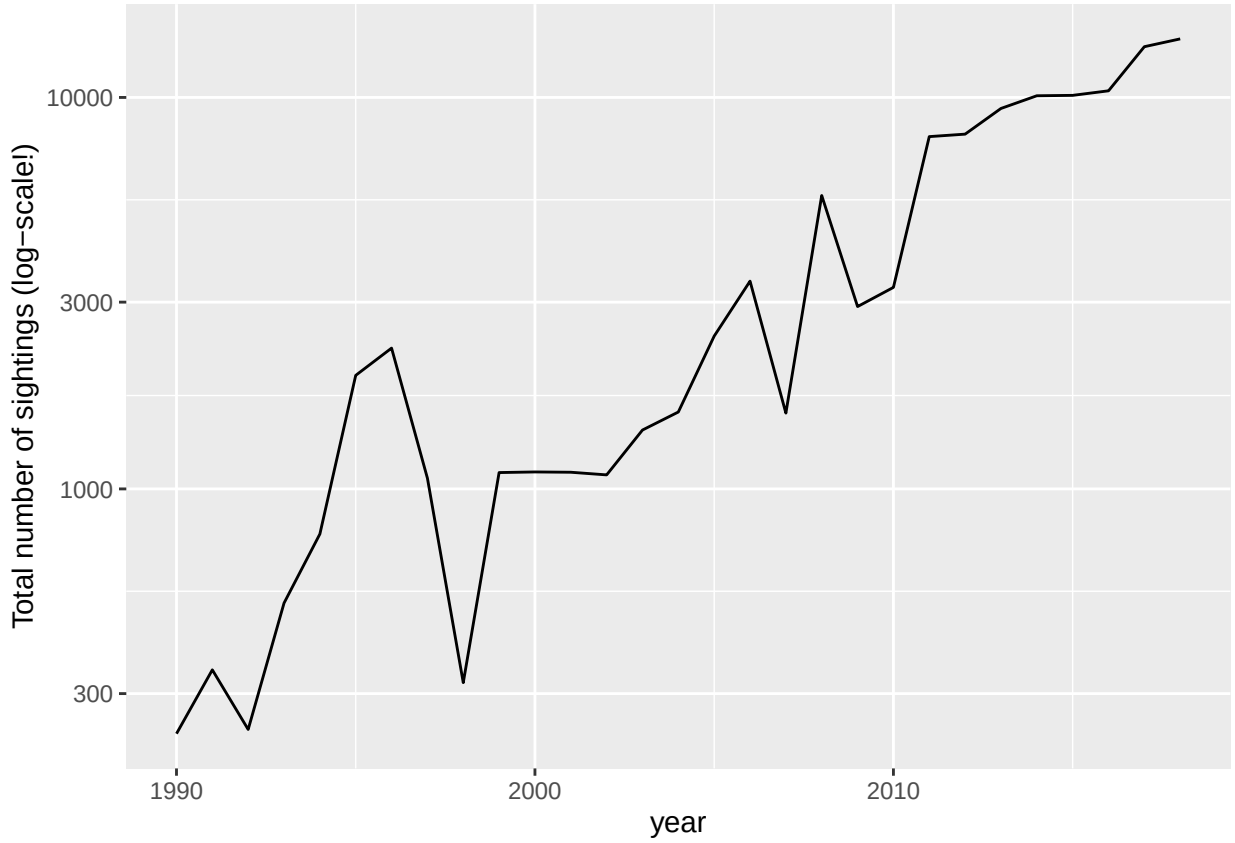
Determination of closure periods

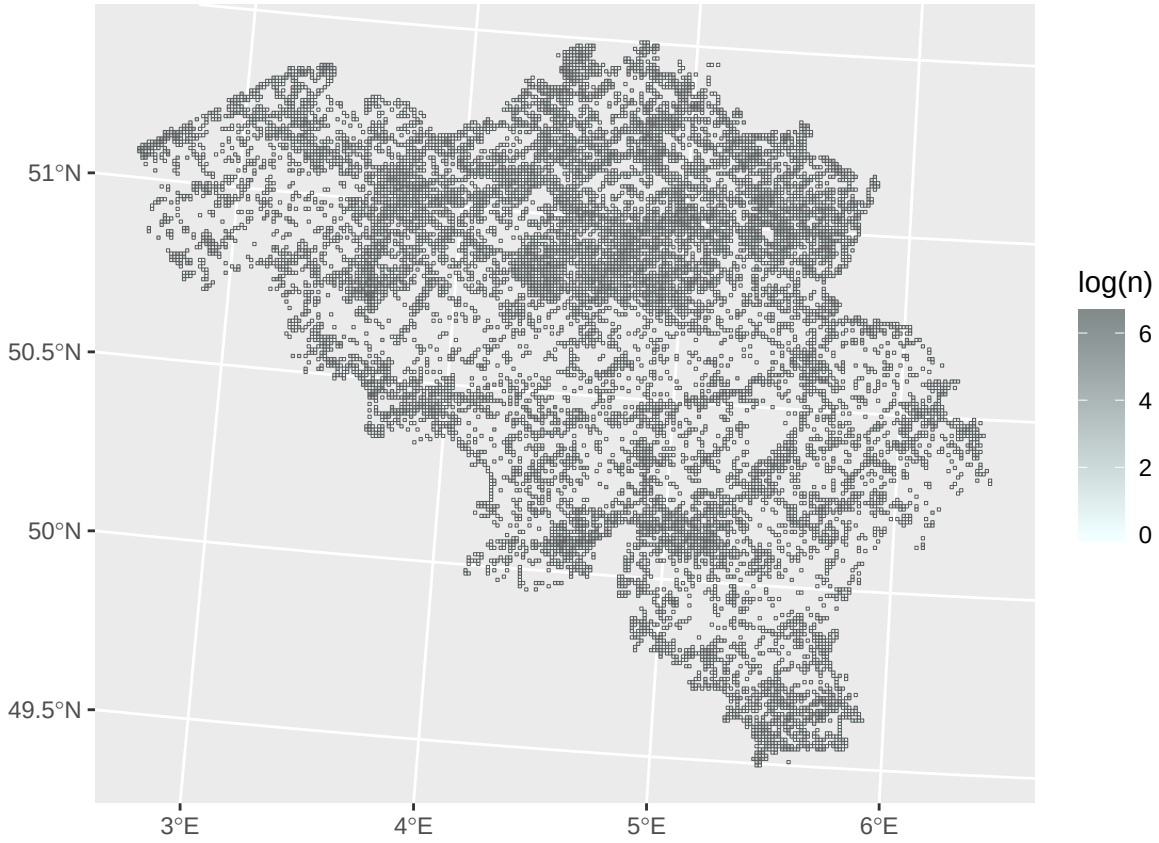
Separation between two flight periods is defined as a concavity in the density curve provided the distance from the concavity is at least 1/3 the maximum density value to the local peak left from the concavity and

1/5th to the local peak right from the concavity. In case of one generation, season_start is 5% percentile and season_end is 95% percentile. In case of more than one generation, the corresponding percentiles of the first generation are used.

We found evidence of two generations in three *Tetrix* species (*T. ceperoi*, *T. subulata*, *T. undulata*).

Total number of sightings by year and by site_id





Final tables

Extra filter to keep only grid cells in grid_file_FINAL.

The following tables were written to csv files (comma separated, and . as decimal mark):

`grasshoppers_observations_finalgrid.csv`, first ten rows:

```
## # A tibble: 117,813 x 8
##    species_id  year month   day julian_day site_id source   count
##         <int> <dbl> <dbl> <int>      <dbl>   <dbl> <chr>    <dbl>
##  1    1681054  1993     5    23        143    2885 Saltabel     2
##  2    1681054  1993     8    13        225    4965 Saltabel     1
##  3    1681054  1995     4    23        113       1 Saltabel     1
##  4    1681054  1995     5    16        136      15 Saltabel     1
##  5    1681054  1995     8    12        224      15 Saltabel     1
##  6    1681054  1995     8    26        238       1 Saltabel     7
##  7    1681054  1995     9    10        253    2097 Saltabel     1
##  8    1681054  1995     9    22        265   25345 Saltabel     1
##  9    1681054  1996     4    14        105    2885 Saltabel     2
## 10    1681054  1996     5    21        142      40 Saltabel    10
## # ... with 117,803 more rows
```

`grasshoppers_species.csv`, first ten rows:

```
## # A tibble: 37 x 6
##    species_id scientific_name species_name_NL species_name_FR season_start
##         <int> <chr>           <chr>           <chr>                  <dbl>
```

```
##  1    1718308 Acheta domesti~ huiskrekel      le grillon dom~        145.
##  2    1688979 Bicolorana bic~ lichtgroene sa~ <NA>                   174
##  3    1708567 Chorthippus al~ kustsprinkhaan  criquet marginé        181
##  4    1708251 Chorthippus bi~ ratelaar        criquet mélodi~        181
##  5    1708459 Chorthippus br~ bruine sprinkh~ criquet duetti~        173
##  6    1708667 Chorthippus mo~ snortikker      criquet des ja~        192
##  7    1700841 Chrysochraon d~ gouden sprinkh~ criquet des cl~        162
##  8    1683331 Conocephalus d~ gewoon spitsko~ le conocéphale~        178.
##  9    1683067 Conocephalus f~ zuidelijk spit~ <NA>                   189
## 10    8356077 Ephippiger diu~ zadelsprinkhaan ephippigère de~        218.
## # ... with 27 more rows, and 1 more variable: season_end <dbl>
```