Hawaii Drosophila Data Cleaning

Data version

The raw data were received 30 June 2021.

De-duplication

The raw data started with 4951 rows. As a result of de-duplication a total of 310 rows were removed. We used the following columns as criteria to check for duplicates (i.e. if a record had equal values for all these columns it was deemed a duplicate):

```
## [1] "SPECIES"  "ISLAND"   "POINT_Y"  "POINT_X"  "LOCALITY" "DATE"
```

Cleaning geographical data

Some island names were inconsistent. The original island names were

```
## [1] "Maui"       "Lanai"      "Oahu"       "Hawaii"     "Kauai"
## [6] "Molokai"    "Big Island" "Hawaa"      "Hawaii "
```

The updated names are

```
## [1] "Maui"    "Lanai"   "Oahu"    "Hawaii"  "Kauai"   "Molokai"
```

Some records had low spatial accuracy. Removing those records further eliminated 522 rows.

Furthermore, we checked that all records fall within the bounds of the islands they were reported from (e.g. a record from Hawai‘i Island does indeed fall within the boundary of Hawai‘i Island). We found 1 record falling outside the island polygons.

These are the records falling outside the island polygons:

REFERENCE	SPECIES	ISLAND	POINT_Y	POINT_X	RESERVE	TYPE	LOCALITY
BPBM	obscuripes	Oahu	20.71789	-156.1413	Haleakala	NP	Paliku

These records falling outside the island polygons will be removed unless they can be corrected.

Cleaning up collection dates

Dates were in multiple formats which have been standardized to `YYYY-MM-DD` format. We checked for missing dates and found 0 missing dates.

Final dataset

The final dataset contains 4118 records and is saved as an R object of class `SpatialPointsDataFrame` from the *sp* package (Pebesma and Bivand 2005), and has geographic coordinate reference system `+proj=longlat +datum=WGS84 +no_defs`.

References

Pebesma, Edzer J., and Roger S. Bivand. 2005. "Classes and Methods for Spatial Data in R." *R News* 5 (2): 9–13. https://CRAN.R-project.org/doc/Rnews/.