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 311 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"
```

It should be noted that de-duplication happened after cleaning all species and geographic names as detailed below.

Cleaning species names

The following species names were corrected (i.e. changed from `old_name` to `new_name`):

old_name	new_name
silvestrs	silvestris

Cleaning geographical data

Some island names were inconsistent. The original island names were

```
## [1] "Maui"       "Lanai"      "Oahu"       "Hawaii"     "Kauai"      "Molokai"    "Big Island"
## [8] "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 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`.

The final dataset contains 4117 records. Below we summarize changes between the raw data and filtered data.

Geographic localities

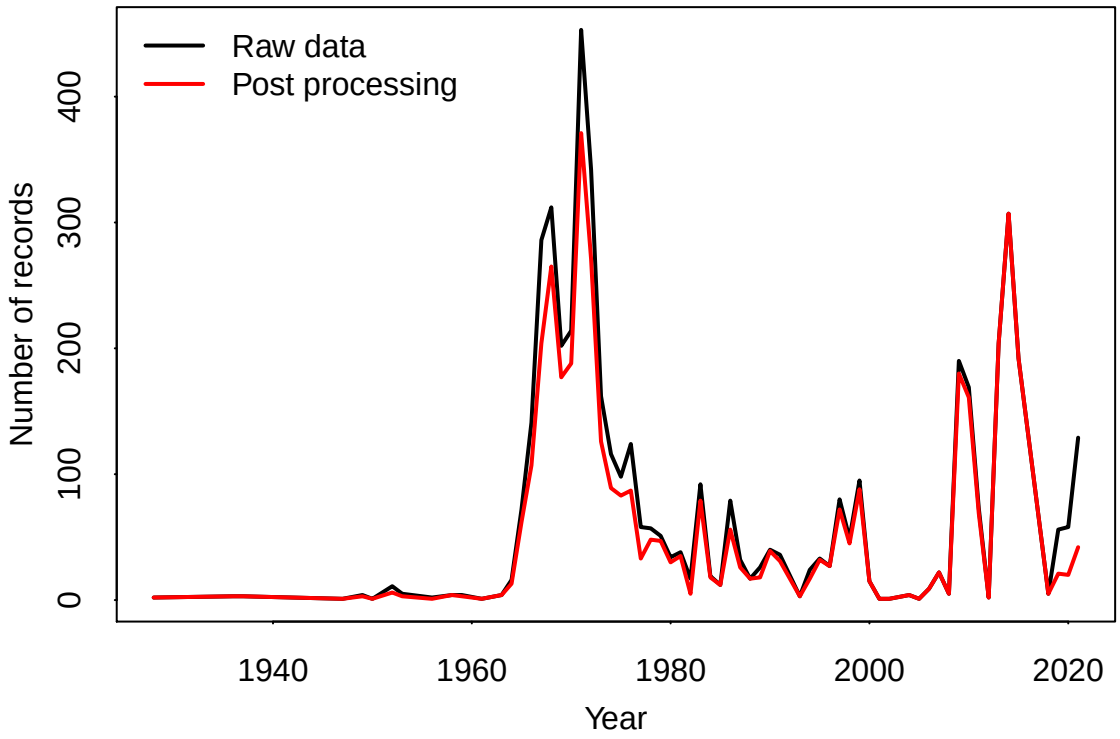
The following localities were lost after filtering:

ISLAND	LOCALITY
Hawaii	Hualalai Ranch
Hawaii	Laupahoehoe Forest Reserve
Hawaii	Papa
Hawaii	Puu Makaala
Hawaii	Kaiholena Ridge
Hawaii	Kapua (sect.), Hoopuloa (quad.)
Hawaii	Papa, South Kona
Hawaii	Puu Waawaa
Hawaii	Upper Olaa Forest
Hawaii	Honaunau Forest Reserve
Hawaii	Keaau Forest
Hawaii	Hualalai Ranch NW Rift Zone
Hawaii	Honomalino
Hawaii	Upper Olaa
Hawaii	Kilauea Forest Reserve
Hawaii	Puna Forest Reserve
Hawaii	Puuwaawaa
Hawaii	Kau Forest Reserve
Hawaii	Keauhou Ranch
Hawaii	Kilauea Forest
Hawaii	Puu Oo Volcano Trail
Hawaii	Humuula
Hawaii	Humuula Saddle
Hawaii	Greenwell Ranch
Hawaii	Pig Hunter's Trail
Hawaii	Waipio Valley
Hawaii	Honaunau
Hawaii	Saddle Road
Hawaii	Mt. Hualalai
Hawaii	Alakahi Stream
Hawaii	Kapua (land section), Hoopuloa (quadrant)
Hawaii	Napau Crater
Hawaii	Holualoa
Hawaii	Wailuku River
Hawaii	Kulani
Hawaii	S. Kohala
Kauai	Kokee
Kauai	Halemanu Stream

ISLAND	LOCALITY
Kauai	Halemanu
Kauai	Halemanu Valley
Lanai	Lanai City
Maui	ridge so. of Iao Valley
Maui	Hana Forest Reserve
Maui	Honomanu
Maui	Makawao
Maui	ridge above Kipahulu Valley
Maui	Olinda
Oahu	Halawa
Oahu	Halawa Ridge Trail
Oahu	Waimano Trail
Oahu	Mount Kaala
Oahu	Ohikilolo Ridge, Makua Valley
Oahu	Waianae
Oahu	Paliku
Oahu	Kawailoa Trail

Samples per year

The below plot shows the differences between sample sizes per year



Samples per species

The below table shows the differences between sample sizes per species. This table should also be checked manually for misspelled names.

species	raw data	post processing
adiastola	122	110
affinidisjuncta	33	31
aglaia	11	10
alsophila	5	2
ambochila	85	80
anomalipes	49	23
assita	19	10
atrimentum	5	4
balioptera	23	20
basisetae	56	35
bostrycha	25	25
cilaticrus	1	1
ciliaticrus	30	24
cilifera	33	29
clavisetae	71	69
claytonae	13	10
conspicua	24	16
craddockae	76	56
crucigera	273	245
cyrtoloma	97	86
differens	18	16
digressa	30	25
discreta	55	50
disjuncta	53	41
distinguenda	24	24
divaricata	34	33
engyochracea	46	42
fasciculisetae	76	56
flexipes	36	34
formella	16	9
glabriapex	32	19
gradata	85	83
grimshawi	144	103
gymnobasis	4	3
gymnophallus	13	12
hamifera	41	40
hanaulae	14	12
hawaiiensis	73	50
heedi	36	33
hemipeza	45	45
heteroneura	89	71
hexachaetae	51	49
hirtipalpus	12	11
inedita	147	139
ingens	19	18
kikiko	5	4
kinoole	3	3
lanaiensis	23	21
lasiopoda	23	23
limitata	43	36
lineosetae	13	13
liophallus	16	15

species	raw data	post processing
macrothrix	34	32
melanocephala	75	71
micromyia	17	16
moli	3	3
montgomeryi	78	78
mulli	7	7
murphyi	110	74
musaphilia	10	5
neoclavisetae	4	4
neogrimshawi	12	9
neoperkinsi	24	21
neopicta	61	52
nigribasis	41	40
nr. alsophila	1	1
nr. truncipenna	4	4
nukea	1	1
oahuensis	54	49
obatai	28	27
obscuripes	12	7
ocellata	3	3
ochracea	70	46
ochrobasis	44	38
odontophallus	23	23
opuhe	1	1
oreas	6	4
ornata	19	14
orphnopeza	54	50
orthofascia	24	17
paenehamifera	13	12
paucicilia	18	18
paucipuncta	38	30
peniculipedis	11	11
picticornis	65	35
pihulu	9	6
pilatisetae	1	1
pilimana	76	74
pilipa	2	2
pisonia	8	3
planitibia	80	78
primaeva	40	27
prolaticilia	63	43
prostopalpis	5	5
psilophallus	6	6
psilotarsalis	4	2
pullipes	34	22
punalua	176	165
quasianomalipes	60	36
recticilia	41	32
reynoldsiae	16	12
sejuncta	15	14
setosifrons	30	25
setosimentum	191	151

species	raw data	post processing
sharpi	4	1
silvarentis	54	38
silvestris	186	138
sobrina	14	13
sodomae	13	13
spaniothrix	10	10
spectabilis	33	33
sproati	172	116
substenoptera	63	57
tarphytrichia	10	10
touchardiae	7	3
toxochaeta	5	3
truncipenna	58	55
turbata	38	37
uniseriata	5	5
varipennis	10	10
vesciseta	8	7
villitibia	9	9
villosipedis	57	33

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/.