Determine cutoff threadholds for classifying presences in ensembles

steppe

3/17/2022

```
library(tidyverse) # data tidying
library(sf) # spatial data compliant with tidyverse
library(raster) # raster data
library(here)
library(rgbif)

source(here::here('scripts/sdm_functions.R'))
set.seed(12)
```

We will use the ecoregion bound to query GBIF for more presence records for testing ensembles for prediction thresholds. We will also used the records which were removed from the intial BIEN data due to high levels of spatial auto-correlation.

We will also import the AIM dataset to feed in more true absences.

We will scrape GBIF for some new records to test our predictions against.

Process united data set to assess validity for binomial logistic regression

obtain BLM absences to generate our full test set for regression

```
## # A tibble: 668 x 3
## # Groups:   binomial [334]
##    binomial             occurrence     n
##    <chr>                     <dbl> <int>
##  1 Acer_glabrum                  0   759
##  2 Acer_glabrum                  1   759
##  3 Acer_negundo                  0   410
##  4 Acer_negundo                  1   410
##  5 Achillea_millefolium          0   892
##  6 Achillea_millefolium          1   892
##  7 Aconitum_columbianum          0   843
##  8 Aconitum_columbianum          1   843
##  9 Actaea_rubra                  0   440
## 10 Actaea_rubra                  1   440
## # ... with 658 more rows
```

```
stacks <- names(predictions_stack)
stacks <- gsub("_glm.*", "", stacks)
species <- testing_set %>% distinct(binomial) %>% pull()

# all species with a prediction and all species
# with enough records to evaluate thresholds
positions <-intersect(stacks,species)
```

```
# subset the evaluation data
testing_set <-testing_set %>%
  filter(binomial %in% positions) %>%
  st_transform(predictions_stack@crs@projargs)
testing_set_L <- split(testing_set, ~binomial)
# to subset the raster stack to the species records
to_sub <- match(positions, stacks)
predictions_stack <- predictions_stack[[to_sub]]

# we now run a one to one extraction.
values_data <-  vector(mode = "list", length = length(testing_set_L))
names(values_data) <- positions
for (i in 1:length(testing_set_L)){
  values_data[[i]] <- raster::extract(predictions_stack[[i]], testing_set_L[[i]],
                                      df = T)
}

# names(values_data) <- positions
values_data <- values_data %>%
  map(~ rename(., Value = 2)) %>%
  bind_rows(., .id = "binomial")

testing_data <- testing_set %>%
  st_drop_geometry( ) %>%
  dplyr::select(occurrence) %>%
  cbind(., values_data) %>%
  dplyr::select(-ID)

write.csv(testing_data, paste0(here() , '/data/processed/logistic_regression_theshold_data.csv'), row.na

rm(stacks, species, positions, to_sub, i, testing_set_L)
```

```
rm(AIM, ecoregion_bound, predictions_stack, presences, values_data, testing_set)
```

```
## Warning in rm(AIM, ecoregion_bound, predictions_stack, presences, values_data, :
## object 'values_data' not found
```

Run models

import and classify data, some values are slightly below and very slightly above 0 and 1, reclassify them here.

split data set

```
## Waiting for profiling to be done...
```

term	CI 2.5	estimate	CI 97.5	std.error	statistic	p.value
(Intercept)	-3.247110	-3.196356	-3.146100	0.0257676	-124.0457	< 0.001
Value	5.753251	5.830559	5.908564	0.0396204	147.1604	< 0.001

We can present results using a classification table, more importantly this table will be used to calculate a variety of metrics for evaluating the ability of the model to accurately classify observations.

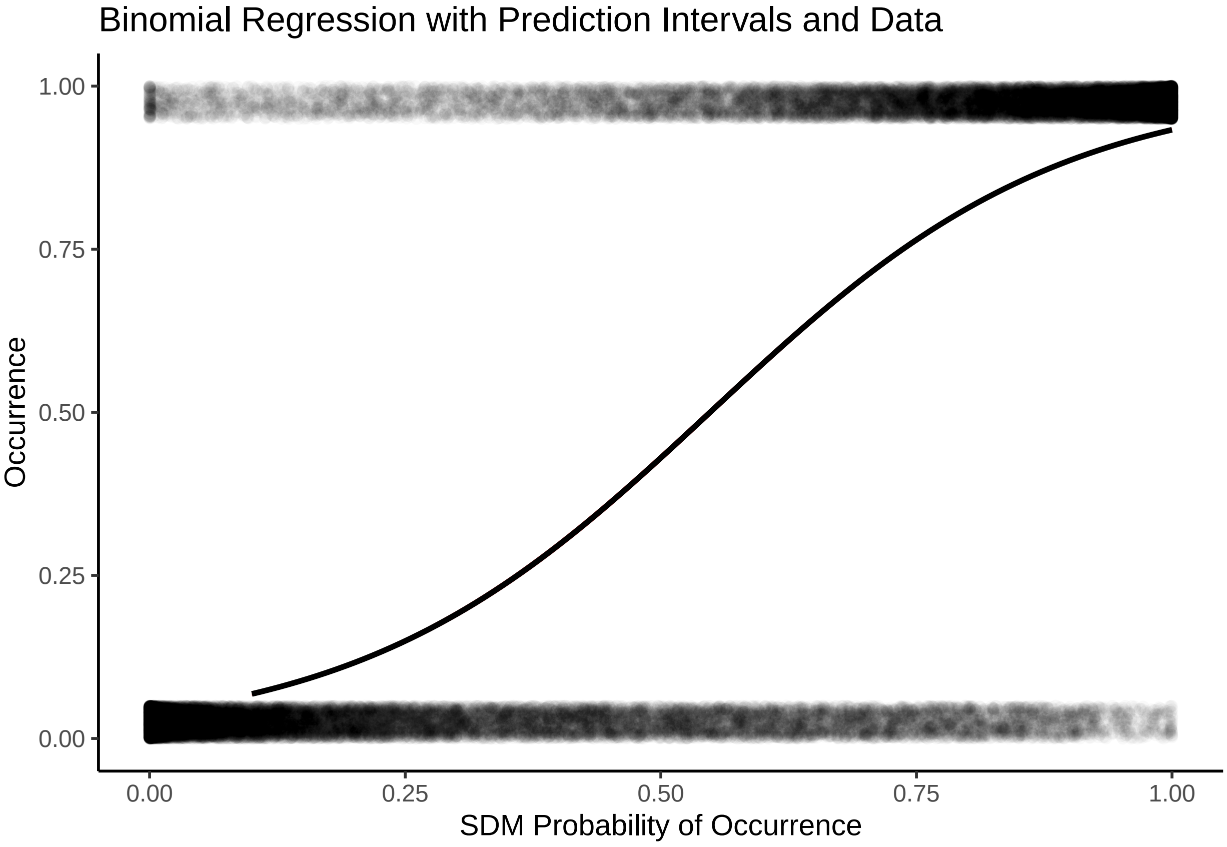
	Absence_train	Presence_train	Absence_test	Presence_test
Absence	25620	3838	11130	1653
Presence	6614	28248	2758	12024

```
## Setting levels: control = 0, case = 1
```

```
## Setting direction: controls < cases
```

Metric	Value
Accuracy (Training)	83.75
Accuracy (Test)	84.00
Recall	81.03
True Neg. Rate	86.97
Precision	88.04
F-Score	0.84
AUC	0.92
Concordance	0.92
Discordance	0.08
Tied	0.00

```
## Warning in rm(accuracy_train, accuracy_test, concordance, Recall, TNR,
## Precision, : object 'pair' not found
```


Binomial Regression with Prediction Intervals and Data
Occurrence
SDM Probability of Occurrence