Applied Remote Sensing Training (ARSET) Program

Species Distribution Modeling with Google Earth Engine

Homework Questions

Question 1

Which statement best describes a Species Distribution Model (SDM)?

Answers: (bold correct)

a. A model that predicts future climate
b. A model that estimates where environmental conditions are suitable for a species
c. A model that classifies satellite imagery into land cover types
d. A model that directly measures biodiversity

Feedback:
A Species Distribution Model uses environmental factors to predict where a species may be present based on information about the conditions in which it was observed previously.

Question 2

Which of the following are commonly used as predictor variables in an SDM?

A. Elevation
B. Annual precipitation
C. Land Cover
D. Species Names

Answers: (bold correct)

a. A and D
b. B and C
c. A, B, and C
d. A, B, and D

Feedback
Since the SDM is focused on finding suitable conditions based on environmental variables, the species name is not a predictor. The model doesn't have any information about the species other than the relationship between previous observations and the predictor variables.

Question 3

To access analysis-ready predictor variables from the Google Earth Engine Data Catalog for species distribution modeling, which of the following best describes the first step?

Answers: (bold correct)

a. Run a machine learning algorithm on raw satellite data
b. Search the GEE Data Catalog for preprocessed datasets with appropriate spatial resolution and temporal coverage
c. Download raw data from USGS and process it in Python
d. Contact GEE support for a custom data request

Feedback:
The GEE Data Catalog contains a number of preprocessed analysis-ready datasets which can be filtered to find options which fit your time, place, and species of interest. Searching the Catalog for ecologically relevant predictor variables is a great first step in setting up a species distribution model run in GEE.

Question 4

What is the correct line of code to create an Earth Engine object?

Answers: (bold correct)

a. var number = 7
b. Num = 89
c. Number = ee.Number(71)
d. var number = ee.Number(71)

Feedback:
When using Google computing infrastructure to run operations, rather than the processor of your own computer, we need to create objects that can be sent in a form that is understood by Google servers. These are the Earth Engine objects. These are written in the form of ee.thing with thing being a data type, also known as containers. Also, note that you need to use var to define and save an object.

Question 5

When evaluating a model's performance using a threshold-independent metric like Area Under the ROC Curve (AUC-ROC), what does an AUC value of 0.5 indicate?

Answers: (bold correct)

a. Perfect discrimination between presence and absence/pseudo-absence locations
b. The model has successfully extrapolated data to a new geographic area
c. Performance is worse than random
d. Discrimination is no better than random performance

Feedback:
A value of 0.5 is equivalent to random chance, whereas values closer to 1 indicate better performance. Remember that AUC stands for Area Under the Curve and this integrated area across all threshold values is highest when the true positive rate approaches 1 and the false positive rate approaches 0. In other words, the AUC is a measure of how good the model is at dividing two classes and at a value of 0.5 that would be no better than chance.

Question 6

When using machine learning algorithms like Random Forests that require presence and absence data, but you only have presence data available, what is synthetically generated to act as "0s" in the model?

Answers: (bold correct)

a. Background points
b. Spatially thinned points
c. Pseudo-absences

d. Continuous probabilities

Feedback:
It is often the case that we only have species presence since it is far more remarkable to see a target species and therefore to record positive (presence) observations than to record negative observations (true absence). We can still use this presence-only data in machine learning by generating pseudo-absences which are sampled from the study area to capture the range of environmental conditions present in the study area. It is worth noting that these pseudo-absences are not predictions of true absence and thus the sampling method(s) used to create the pseudo-absences should follow a deliberate strategy with respect to the environment and species of interest.

Question 7

According to the presentation, what are the two prominent applications of Species Distribution Models (SDMs)?

Answers: (bold correct)

a. Eradicating invasive species and breeding endangered animals
b. Calculating vegetation indices and mapping raw satellite data
c. **Understanding the drivers of species distributions and predicting habitat suitability across time and space**
d. Identifying false negatives and estimating temporal resolutions

Feedback:
Species distribution models give us information about the drivers of species distribution and predictions of where a species would be expected to occur based on the environmental conditions across time and space.

Question 8

In Google Earth Engine (GEE), how is raster data (such as a digital elevation model or average annual temperature) usually organized and referred to within the platform's coding environment?

Answers: (bold correct)

a. **Images or image collections**
b. Vector arrays and attributes
c. Object grids and schemas
d. Polygons and feature collections

Feedback:
Google Earth Engine organizes raster data into images and image collections which are referenced in the code as ee.Image and ee.ImageCollection respectively.

Question 9

Why is "strategic spatial thinning" used on species occurrence data (such as opportunistic data from GBIF)?

Answers: (bold correct)

a. To increase the total number of presence records for a better model fit
b. To convert true absences into pseudo-absences
c. To replace missing environmental predictor variables with synthetic data
d. **To eliminate spatial clustering due to sampling bias (e.g., species recorded multiple times near roads)**

Feedback:
Strategic spatial thinning of occurrence data is used to reduce the effects of sampling bias in the species presence observations.

Question 10

From the three different strategies presented to delineate the area to create pseudo-absences or background points, which one would be more appropriate when using a Random Forest classifier on a terrestrial species?

Answers: (bold correct)

a. Option 1: create “absence points” across the entire area of interest but only after applying a water mask to remove the oceans and continental water
b. Option 2: create “absence points” only at a defined distance from any presence point
c. Option 3: use an environmental profiling technique to create “absence points” on dissimilar areas within the defined area of interest

d. **Option 2 and 3 are valid.**

Feedback:
Using the entire area of interest to create "absence points" is only recommended when using Maxent with the goal of providing the "average" environmental conditions to contrast against the known presence locations.

Question 11

A researcher wants to understand the ecological effect that vegetation productivity has on the occurrence of zebras. It downloads available data from GBIF and runs an SDM with a Random Forest classifier using NDVI as a predictor variable (among other environmental variables). The variable importance results show that NDVI is on the top two most important variables in identifying zebra suitable habitat. Is this an appropriate manner of answering the research question?

Answers: (bold correct)

a. Yes, it is a sound approach
b. **No, the researcher should be using a parametric approach (e.g., GLM) and other types of data (e.g., movement data or sampling with transects)**

Feedback:
The variable importance quantifies how effectively each environmental layer was able to sort presence points away from the absence points. However, machine learning techniques such as Random Forest should be used for making predictions, not inference. The variable importance result does not provide the direction of the effect, so it does not directly provide information on how increases or decreases of a variable relate to increasing or decreasing habitat suitability. Importantly, when using presence only data and creating "pseudo-absences" we lack real information on areas species avoid to make inferences.