# Preliminary Research and Question Development

## Environmental Niche Modelling Methods

Bioclim

* Climate Envelope Method
* Occurrence points and environment space – draws a box around it
* Look at environmental gradients of various variables and see what conditions a species appears in most. Assume that the highest quality habitat are the most populated areas
* Does not extrapolate much. Quite 2 dimensional
* Finds combinations of environments that do not really exist

Domain

* Climate Envelope Method
* Environment space distance to other data points using a modified Manhattan distance rather than Euclidian
* A math equations and formulas

  Description automatically generated with medium confidence
* The domain score for this combination environment is proportional to the shortest of those distances

GLM

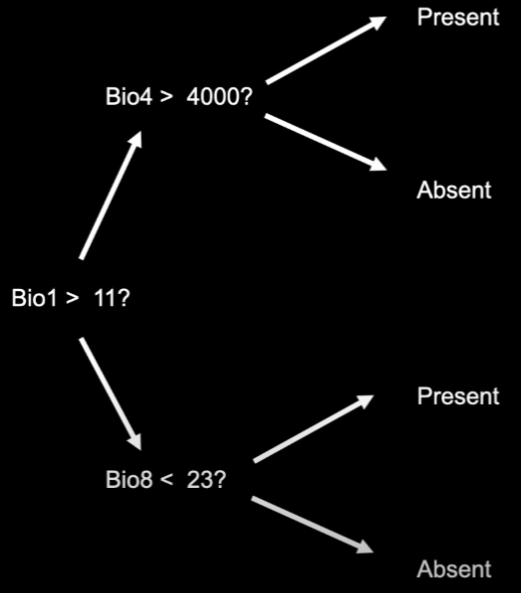
* Extremely flexible and complex
* Generalized Linear Modelling
* Uses background, pseudoabsence, or absence points as well as occurrence data
* Fits models using maximum likelihood
* Very flexible but requires a lot of hands-on work (you have to know what to do, needs great understanding)
* Binomial GLM regression on presence/absence data
* A graph of a function

  Description automatically generated

GAM

* Generalised Additive Model
* Looks quite a bit like GLMs
* The primary difference is that they allow us to fit a species’ response using a nonlinear smoother function of the environmental variables
* Has more localized fits to the equation
* Allows you to control how complex the function will be
* We can combine smoothed predictors with unsmoothed predictors

Random Forests

* Models are based on decision trees
* Decision trees take a set of outcomes and a set of predictors and produce a set of binary decisions intended to predict those outcomes
* Decision trees: Start by choosing the predictor and threshold value that best separates our data into classes (presence vs. absence or pseudoabsense)
  + As only presence data can truly be available, Pseudoabsense is artificial absence data, sometimes called background data
* 
* Tend to fit training data extremely well
* Can often be terrible at predicting other data
* Random forests extends decision trees to compensate for limitations
* For each replicate, start by randomly resampling your occurrences with replacement
  + Helps control overfitting to training data
* At each node, sample a random subset of your predictors
  + Helps control for correlations between trees based on one or a few strong predictors

## Model Selection

All models are wrong, but some are useful. Some are wrong and useful.

What does a good model do?

* Predict new occurrences?
* Predict relative suitability of habitats?
* Feature Selection?
* ect

What do I want my model to do?

* For platypus: I want my model to predict relative suitability of habitats
* For Rakali: I want my model to predict new occurrences

We use discrimination accuracy.

* (AUC, TSS, Kappa, Training and test data)

Or you could use calibration

## Possible Questions or Aims

Using ENM (Bioclim) to predict the presence of species around Australia/NSW

Using Gov water datasets to identify rivers and waterways they could be present in, but are currently not

Create a function, allowing somebody to insert a location, and it gives the closest observed and likely spots to see platypus.

Platypus or use a less locomotive species?

|  |  |  |  |
| --- | --- | --- | --- |
|  | Pros | Cons |  |
| Platypus | - Animal of personal interest  - Lots of datapoints  - Public interest | - Mobile species (? Research more)  - Likely spotted if it is already there  - Does ENM for platypus exist already? | - Due to high social presence of animal, it is likely that any potential sightings have been recorded. Not much new information  - Platypus are still rare, and there would likely be public interest in the topic |
| Trees | - Less Mobile, thus in theory, spatial distribution is more closely tied to  - | - Does not interest me as much, no use of water datasets, may make the project too simple |  |
| Wollemi Pines | - May be very useful if can be done accurately due to the rarity of the species | - No datapoints to train from (Cant do Bioclim or Domain). Very closely guarded for environmental safekeeping.  - Ethics if successful. Should that information be out there?  - | - By using Random Forests method -> the use of decision trees, it could be done, or by finding a species with the same fundamental niche. Both would be very difficult |
| Rakali (Native “Water Rat” | - More neglected than platypus  - May be more useful information if less occurrence is out there  - Water dwelling creature | - Less public interest | - Gives some love to |
| Invasive Species | - In a way, ENM models such as Bioclim or Domain is perfect for predicting where an incursion of an invasive species would be most harmful | - Harder to predict all variables, especially as invasive species live or die depending on their interactions with the native ecosystem.  - Would likely require decision trees. Very difficult  - No point with the location function if it is not yet present | - If I were interested in shifting the topic of the project to biosecurity, this could be interesting |

Trees?

Wollemi Pines?

* No data available

MOSES Tests?

* If I could find a creature with the same fundamental niche as platypus maybe… (Rakali?) but I doubt it
* Could be used for Wollemi pines?

Water Datasets