Multinomial model to classify shrew species

David Simons

2022-07-06

Motivation

Using a multinomial machine learning approach it appeared that classification to shrew species may be possible based on simulated distributions of morphological features obtained from the literature (primarily Kingdon). It potentially suggested that it would be possible to classify these species based on morphological approaches only, reducing the need for molecular classification.

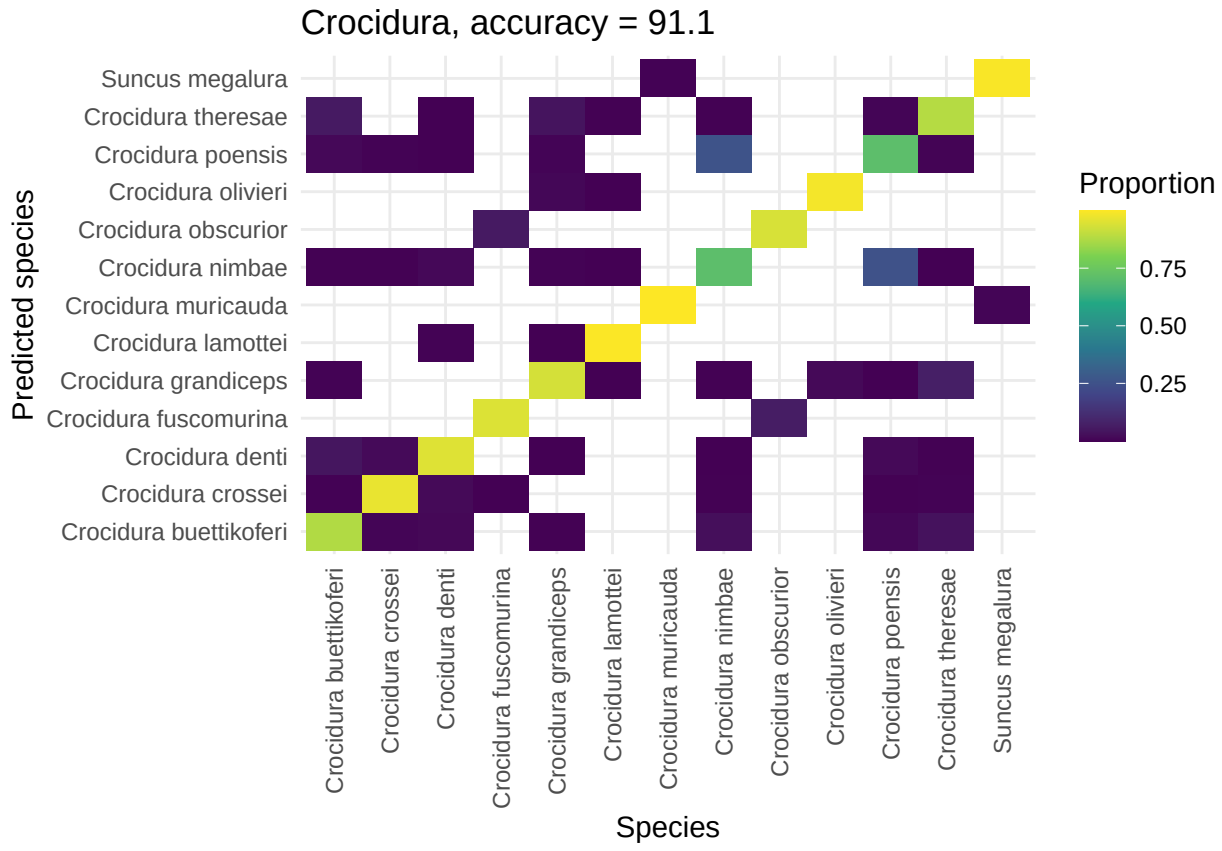


Figure 1: Model predictions using a training and testing subset of the simulated distributions for weight, head-body length, tail length, tail/head-body ratio and hind foot length.

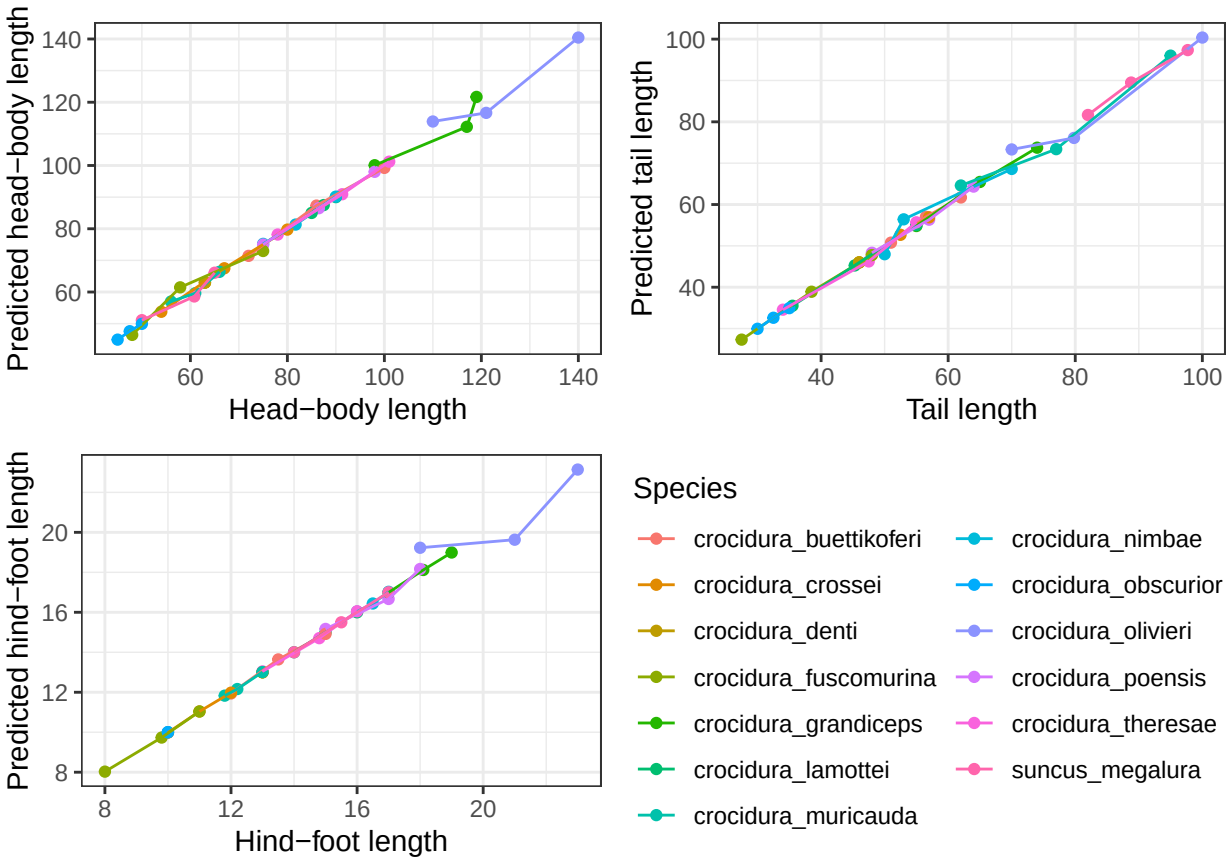
Methods

Training data

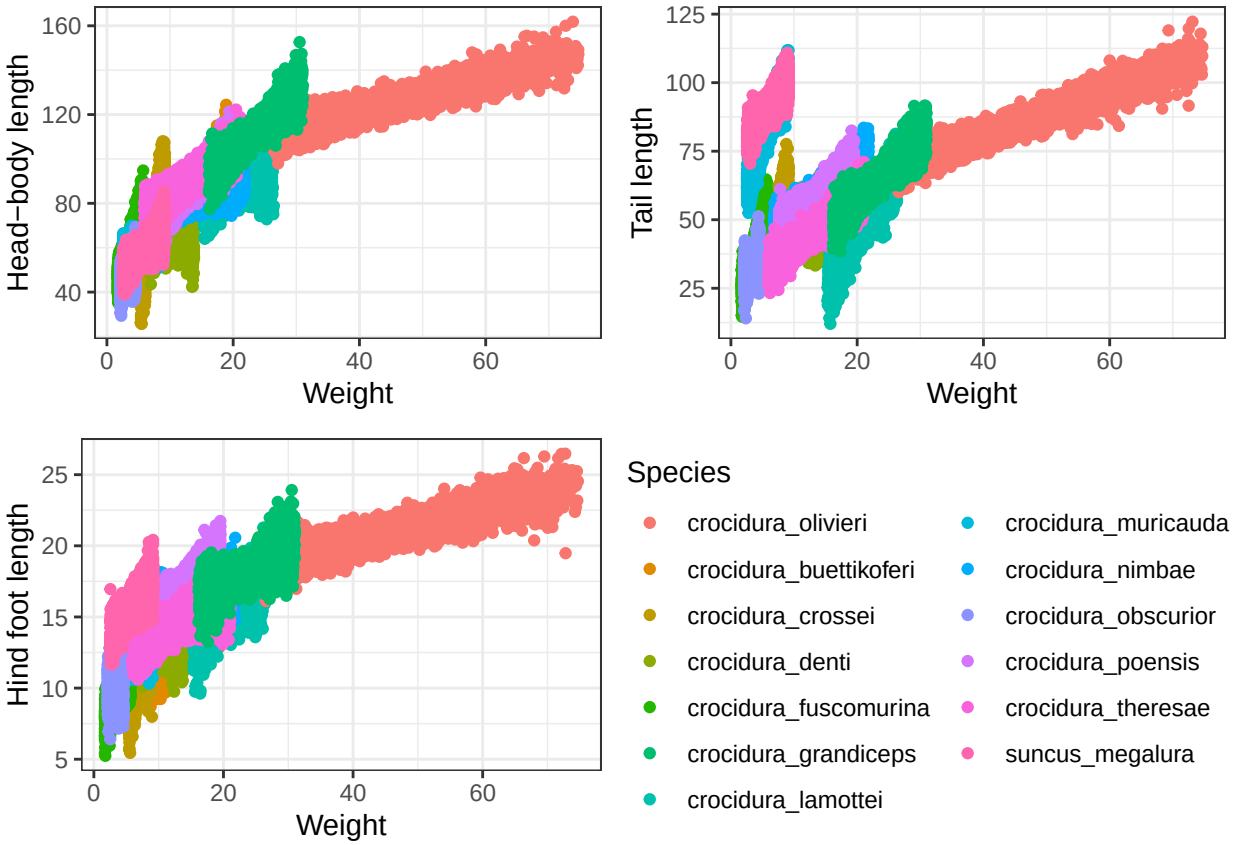
Measurements of shrew weights, head-body lengths, tail lengths, hind foot lengths and ear lengths were obtained from the literature. Where available means, minimums and maximums for each metric were extracted. Where means were not available the midpoint of the minimum and maximum value were used, for species where sex disaggregated values were provided weighted averages were used to calculate the species means. Standard deviations were estimated using the minimum and maximum values (SD = (max - min)/4).

To produce a sufficient number of records to train the multinomial model records were simulated from these values. The weight of a shrew species was assumed to follow a normal distribution with the midpoint matching the literature reported mean and 95% of values falling within two standard deviations of the mean. A truncated normal distribution was used to simulate the weights of 5,000 individuals of each species with the lower bound set at 85% of the literature reported minimum and the upper bound set at 115% of the literature reported maximum.

It was expected that there is a strong correlation between an individual shrews weight and other morphological parameters including head-body length, tail length and hind-foot length. To account for this a Generalised Linear Model was produced for each of these parameters (i.e. head-body length ~ weight * species). These models were used to predict these morphological parameters based on the simulated weights of each individual using the predicted value and standard deviation. Reasonable concordance between literature reported values and model derived values were produced for each species.



The morphological parameters in the simulated dataset are shown below.



These data are used to train the Multinomial Log-linear model via a neural network. The formula is of the structure:

$$multinom(\text{Species} \sim \text{Weight} + \text{Head-body length} + \text{Tail length} + \text{Hind foot length}) \quad (1)$$

The accuracy of this model on the 30% of withheld simulated data is 91.1%. The variable importance for the model, produced using the `VarImp` function is:

```
##              Overall
## weight    8366.9334
## head_body  172.4048
## tail       648.5845
## hind_foot  964.5629
```

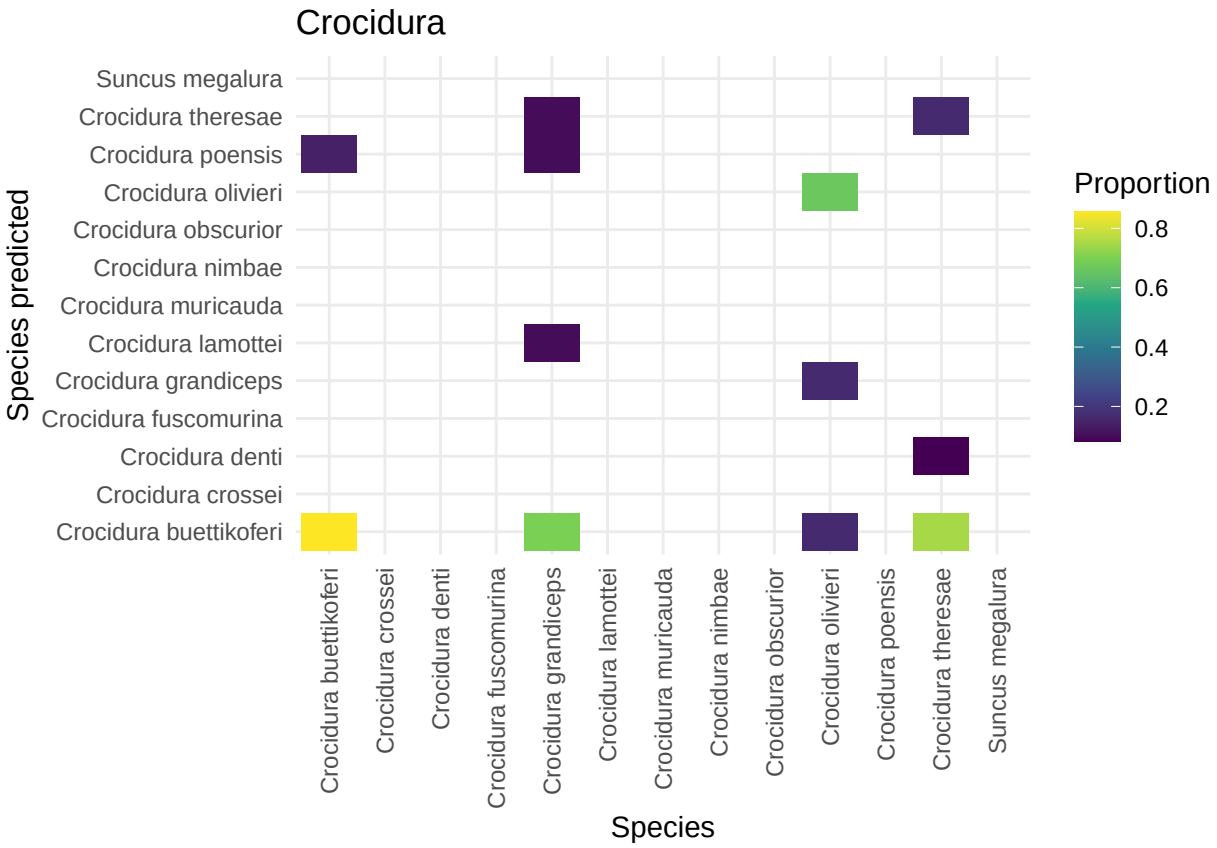
This suggests that classification is primarily based on weight with a lower contribution from hind foot length, tail length and head-body length.

Results

Out of sample predictions

To assess the ability of the model to classify observed - out of sample - data, 41 records of shrews that have been sequenced with their morphological measurements were used. 16 of the 41 records were correctly attributed to their species (40%). The model correctly classified 6 *Crocidura buettikoferi* (86%), 8 *Crocidura olivieri* (67%), and 2 *Crocidura theresae* (17%). No *Crocidura grandiceps* were successfully classified with the majority of these being incorrectly classified as *Crocidura buettikoferi* (70%). Similarly, for *Crocidura*

theresea the majority of individuals were misclassified as *Crocidura buettikoferi* (75%). The confusion matrix for these out of sample records are shown below.



Sensitivity analysis

In sensitivity analysis, restricting the model training data to species identified in the out of sample dataset there was minimal improvement in correct classification (17 compared to 16) with a single previously incorrectly classified *Crocidura buettikoferi* correctly classified.

Discussion

The expected ability to classify shrews based on their morphology was not found when using out of sample data. There are several potential reasons for this.

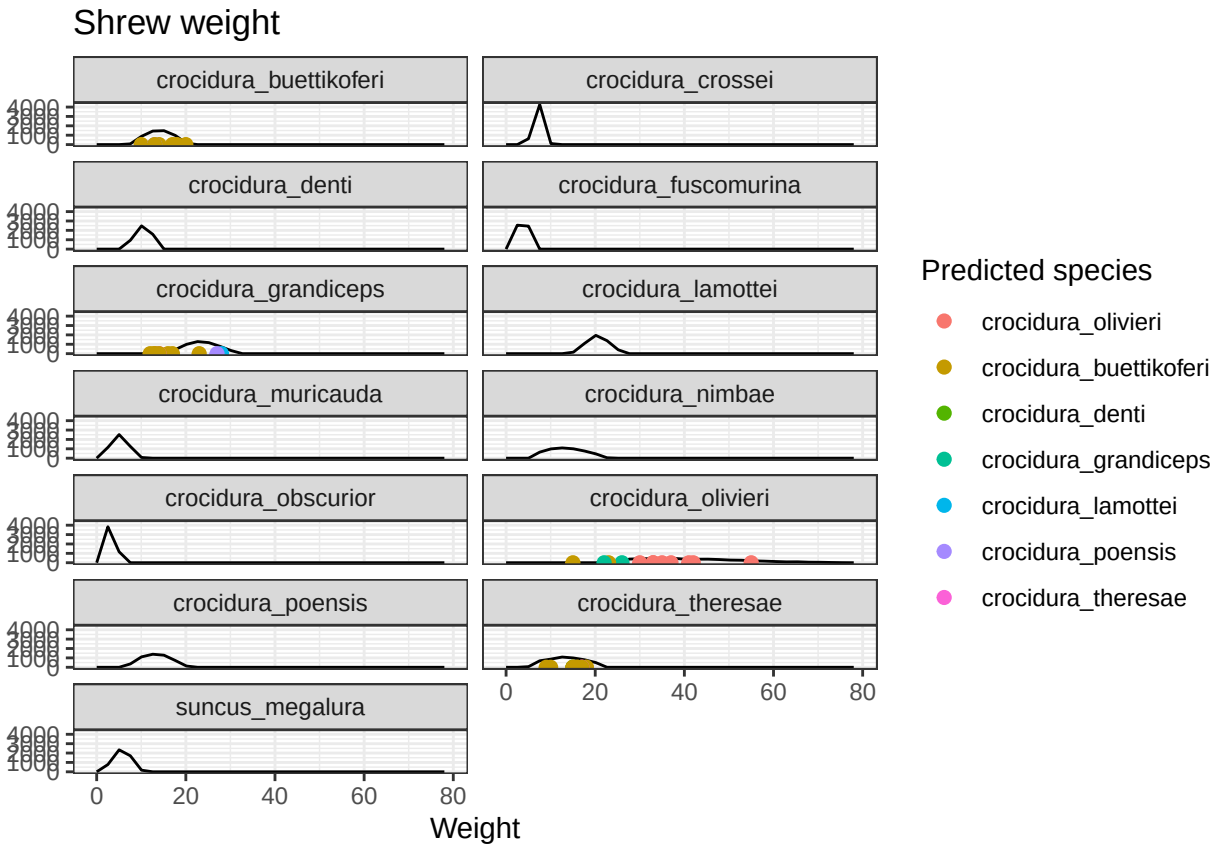
First, the literature had limited number of measurements for each species (< 20 individuals), this necessitated the simulated approach which may not be suitable if the measurements obtained are not truly reflective of the true distribution of measurements in natural populations of these species.

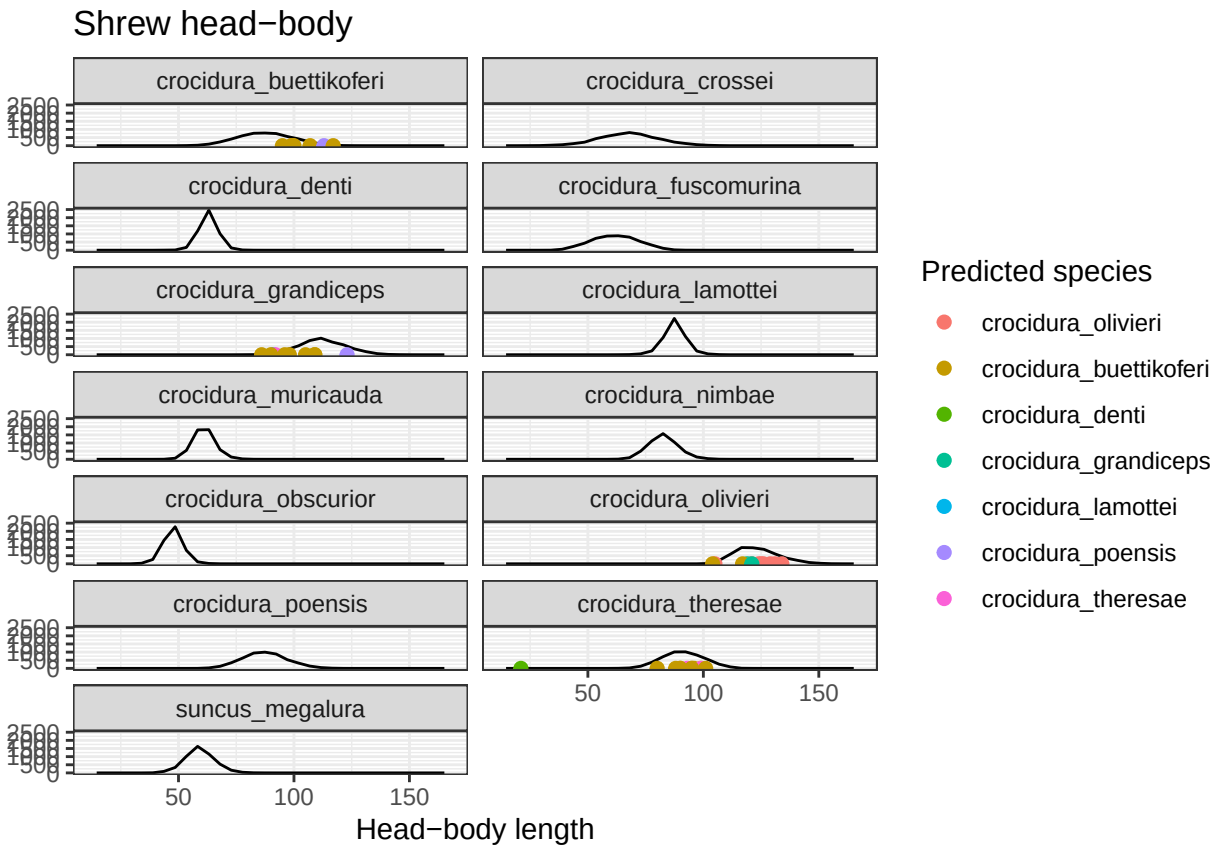
Second, there is some evidence that shrew morphology differs across their ranges which may be important if the measurements came from specific sub-populations not representative of the true morphology.

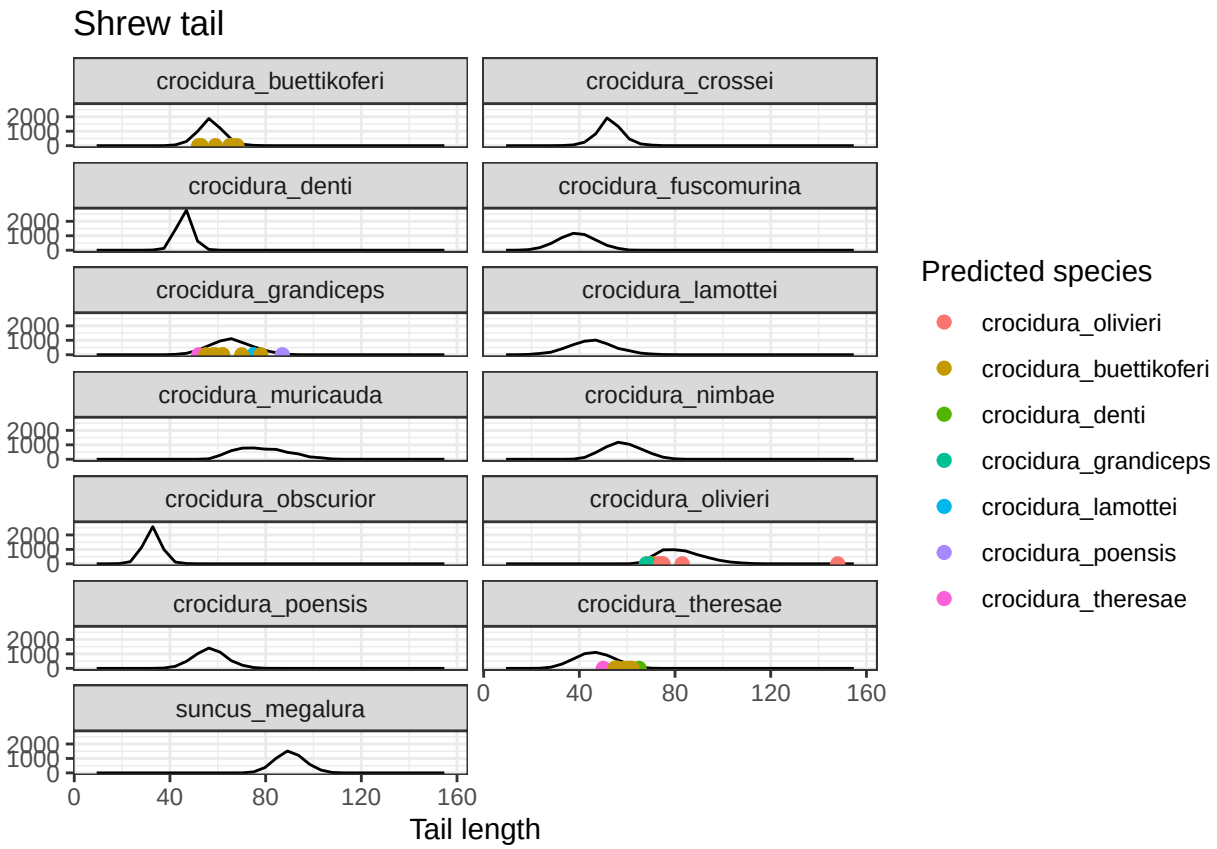
To explore where the model may be falling down I have looked at where the observed values fall within the simulated values for each metric. For the weights, most individuals fall within the simulated values, however for *Crocidura grandiceps* observed weights possibly fall below the simulated values. Head-body lengths for observed *Crocidura olivieri* and *Crocidura theresae* generally fall within the simulated values however, for *Crocidura buettikoferi* and *Crocidura grandiceps* the simulated distributions do not align well with the observed values. Tail length is generally well matched between observed and simulated data, perhaps

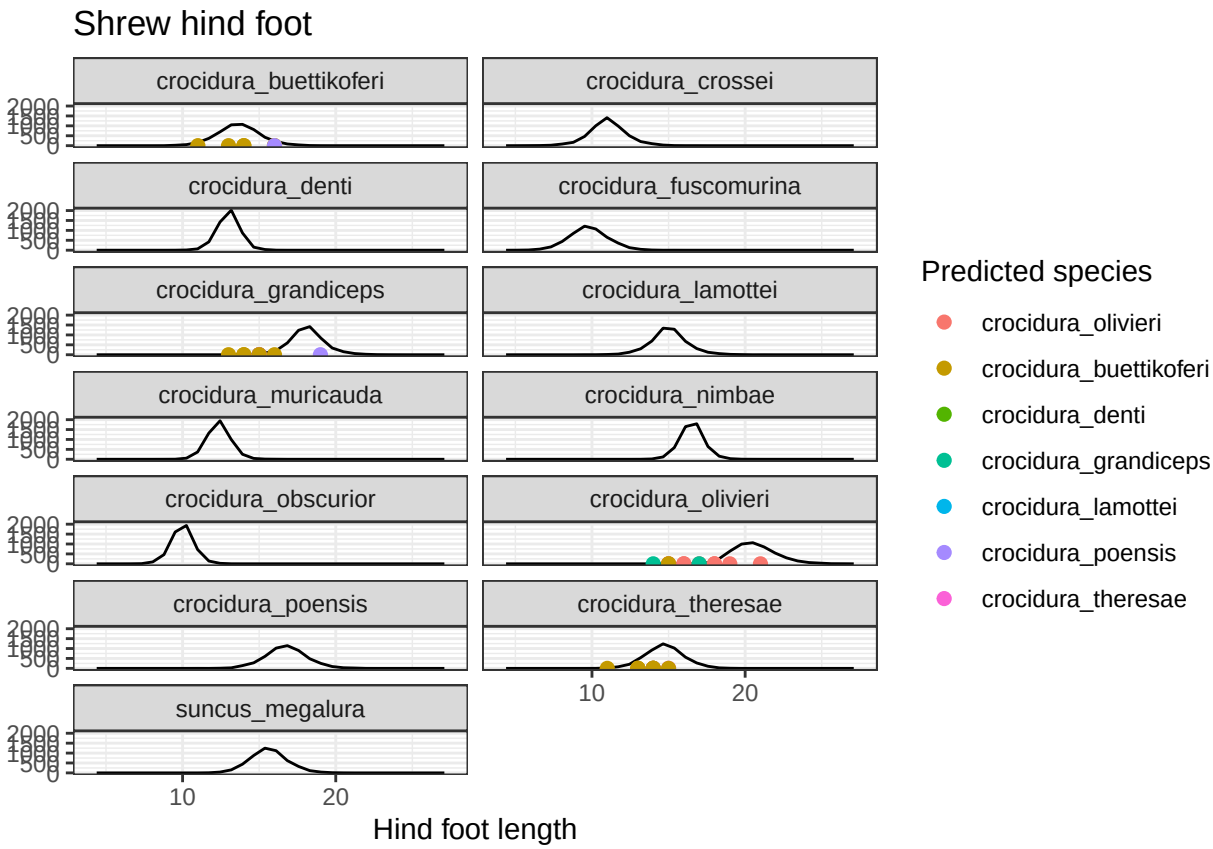
less well for *Crocidura theresae* and a single *Crocidura olivieri* which has a particularly long tail. Hind foot is perhaps the least well matched particularly for *Crocidura grandiceps* and *Crocidura olivieri*.

The below graphs show the where the out of sample measurements fall within the simulated morphology for each species with the points colour coded to the classification produced by the model.




Shrew head−body
crocidura_buettikoferi
crocidura_crossei
crocidura_denti
crocidura_fuscomurina
crocidura_grandiceps
crocidura_lamottei
crocidura_muricauda
crocidura_nimbae
crocidura_obscurior
crocidura_olivieri
crocidura_poensis
crocidura_theresae
suncus_megalura
50
100
150
Head−body length
Predicted species
crocidura_olivieri
crocidura_buettikoferi
crocidura_denti
crocidura_grandiceps
crocidura_lamottei
crocidura_poensis
crocidura_theresae


Shrew tail
crocidura_buettikoferi
crocidura_crossei
crocidura_denti
crocidura_fuscomurina
crocidura_grandiceps
crocidura_lamottei
crocidura_muricauda
crocidura_nimbae
crocidura_obscurior
crocidura_olivieri
crocidura_poensis
crocidura_theresae
suncus_megalura
2000
1000
0
0
40
80
120
160
Tail length
Predicted species
crocidura_olivieri
crocidura_buettikoferi
crocidura_denti
crocidura_grandiceps
crocidura_lamottei
crocidura_poensis
crocidura_theresae


Shrew hind foot
crocidura_buettikoferi
crocidura_crossei
crocidura_denti
crocidura_fuscomurina
crocidura_grandiceps
crocidura_lamottei
crocidura_muricauda
crocidura_nimbae
crocidura_obscurior
crocidura_olivieri
crocidura_poensis
crocidura_theresae
suncus_megalura
2000
1500
1000
500
0
10
20
Hind foot length
Predicted species
crocidura_olivieri
crocidura_buettikoferi
crocidura_denti
crocidura_grandiceps
crocidura_lamottei
crocidura_poensis
crocidura_theresae