Validating SDMs

Bob O'Hara

Abstract

Introduction

- SDMs often used. Validation, i.e. how good are they, is an issue. (A. Lee-Yaw et al. (2022)).
- Current validation often on same data, which can lead to over-fitting. Also, metrics often wrong: because it is the same data, will replicate biases. Thus TSS should be FSS in this context.
- A better approach is to validate on external data. Could calculate AUC, TSS etc. but teh validation daat will be imperfect too. Here we take a slightly different approach that can be adapted to different data types.

Smith and Levine (2025) showed that Yackulic et al. (2013) were right: you shouldn't think at all species are equal. But…

The same data was used by Valavi et al. (2022) to compare the perform,ance of different models. Here our focus is on the comparisons can and should be made, and what the results tell us.

Theory

Most data is modelled as presence-only, so we will follow that here but indicate the adjustments needed if the data are of a different type. For presence-only data the underlying theory can be derived from point processes (e.g. Fithian and Hastie (2013), Renner and Warton (2013), Warton and Shepherd (2010), Aarts, Fieberg, and Matthiopoulos (2012)). The actual distribution is viewed as an intensity surface, with a higher intensity where a species is more likely to be present. The data is made up of locations of observations of the species of interest, along with locations in the same region where the species was not observed. These latter points are usually called "pseudo-absences" in the SDM world. In a point process approach they are seen as integration points - to calculate the likelihood for the presences we should integrate the intensity over the full space.

Of concern here is the model for the intensity surface. It is assumed that it is affected by several environmental covariates in an additive way. Specifically, of $\lambda(\mathbf{s})$ is the intensity at point $\mathbf{s}$, the assumption is that $\log \lambda(\mathbf{s}) = \alpha + \sum_i \beta_i X_i(\mathbf{s}) = \alpha + \eta(\mathbf{s})$ where $X_i(\mathbf{s})$ is a *feature*, i.e. a function of the environmental covariates. For simplicity we can think of this as an environmental covariate, but it could also be a non-linear term, such as a quadratic or the product of two environmental covariates (i.e. an interaction). This same approach is used in GLMs, MaxEnt, GAMs and many other methods. The β_is are standard regression coefficients, and α is an intercept. This is not normally of interest for presence-only data, as it is determined by the amount of data, i.e. the sampling effort. Thus the focus is on $\eta(\mathbf{s})$.

If the model is correct, then the number of individuals of a species in an area A would follow a Poisson distribution with mean $A\lambda(\mathbf{s})$, where A is a constant that will depend on sampling effort. The probability of an absence is then the probability of zero individuals, i.e. $e^{-A\lambda(\mathbf{s})}$. This leads to a model for presence/absence where $P(Z = 1) = 1 - e^{e^{-\alpha_c - \eta(\mathbf{s})}}$, which is equivalent to a GLM with a binomial response and a cloglog link function $(\log(-\log(1 - P(Z = 1))) = \alpha_c + \eta(\mathbf{s}))$. An alternative is to use a logistic regression (Elith et al. (2011)), i.e. $\log P(Z = 1)/(1 - P(Z = 1)) = \alpha_l + \eta(\mathbf{s})$, with a derivation based on averaging over possible distributions of the covariates and response (Phillips and Dudík (2008)).

The purpose of laying this out is to suggest an approach to model validation. If we fit a model to presence-only data, we get estimates for $\eta(\mathbf{s}) = \sum_i \beta_i X_i(\mathbf{s})$. We can then use this to calculate predicted probabilities (up to an intercept) for a new presence/absence data set. We can then use these predictions in a GLM with a cloglog or logit link. If the model is correct, the slope should equal 1.

Although we are not primarily interested in the intercept, except to note that it should not equal 0 (as pointed out by both Yackulic et al. (2013) and Smith and Levine (2025)), we might expect that if we look at the estimates for the fitting and validation models across species from the same data sets, they should be correlated. i.e. a species that is more common in the presence only data set may also be more common in the presence-absence data, because it may be more common or be easier to spot.

Why, though, might a regression coefficient not equal 1? Aside from random chance, it may be because the model is wrong. Indeed, as all models are wrong this is likely. But it may also be because, even if the fitted model is correct, it is fitted to finite data, so there is uncertainty in the parameter estimates. This uncertainty will feed through to validation model, where $\eta(\mathbf{s})$ has been estimates with error. The overall effect of this is not just to increase uncertainty, but to bias the estimate of the slope towards zero (Carroll et al. (1984)). This could be mitigated by incorporating this uncertainty into the second model (REF).

Methods

Data

The

From Elith et al. (2020). For New Zealand we merged classes 2 (high) and 3 (not documented) of toxicats, and for NSW we merged level 8 of vegsys (pine plantation) with Dry Open Forest. In both cases, these are rare classes that are not in every data set.

Modelling

- fit MaxEnt to PO data
- predict on PA data
- fit prediction to PA data
- fit "MaxEnt" to PA data (i.e. change weights to be 1)

If good fit, should get beta=1, and correlation between predictions should be high.

Compare AUC & TSS with beta

Compare to simulations, where model is true.

Results

Summary of Predictions on PA data

The estimated slopes and intercepts of the GLM linking the predictions from the MaxEnt model to the PA data are shown in Figure 1. Most of the slopes (NA%) are between 0 and 1, as we might expect, but NA% were negative, and NA% had slopes above 1, with NA% not having 1 in the confidence interval

As would be expected, the estimated intercepts varied between species, i.e. species had differing observed prevalences. This was not correlated between data sets: the largest largest correlation is 0.32, for the NZ data.

The MaxEnt model needs an intercept, but this is determined by the number of presence and pseudo-absence points, so it is not generally helpful. However, we might hope that the relative values of intercepts across species represent relative measures of prevalence. We can examine this be asking whether the correlations between the intercepts from the PA data are similar to those in the PO data. Unfortunately they are not: the largest correlation is 0.32, for the NZ data.
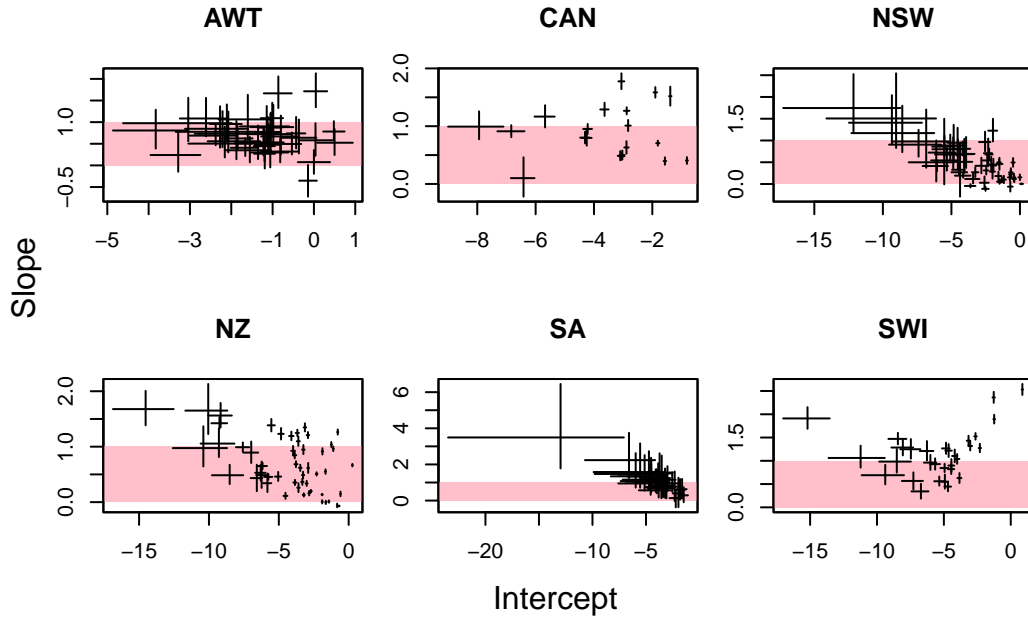
Figure 1: Estimated intercept and slope for calibration of predictions to presence absence data, from Maxent model with all features.

Coefficients above 1?

We can see the effect of adding extra variation in the PO data model in **?@fig-CoefOver1**, for the Swiss species 04. As the extra variation increases, i.e. the PO data becomes worse, the slope tends to increase.

```
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
Warning: UNRELIABLE VALUE: One of the 'future.apply' iterations
('future_sapply-1') unexpectedly generated random numbers without declaring so.
There is a risk that those random numbers are not statistically sound and the
overall results might be invalid. To fix this, specify 'future.seed=TRUE'. This
ensures that proper, parallel-safe random numbers are produced via a parallel
```

RNG method. To disable this check, use 'future.seed = NULL', or set option
'future.rng.onMisuse' to "ignore". [future 'future_sapply-1'
(cbd2217c174b7bf544881e02d5898d6a-109); on
cbd2217c174b7bf544881e02d5898d6a@NTNU-C02DD00KML7J<9793>]

Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

Warning: UNRELIABLE VALUE: One of the 'future.apply' iterations
('future_sapply-2') unexpectedly generated random numbers without declaring so.
There is a risk that those random numbers are not statistically sound and the
overall results might be invalid. To fix this, specify 'future.seed=TRUE'. This
ensures that proper, parallel-safe random numbers are produced via a parallel
RNG method. To disable this check, use 'future.seed = NULL', or set option
'future.rng.onMisuse' to "ignore". [future 'future_sapply-2'
(cbd2217c174b7bf544881e02d5898d6a-110); on
cbd2217c174b7bf544881e02d5898d6a@NTNU-C02DD00KML7J<9793>]

Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

Warning: UNRELIABLE VALUE: One of the 'future.apply' iterations
('future_sapply-3') unexpectedly generated random numbers without declaring so.
There is a risk that those random numbers are not statistically sound and the
overall results might be invalid. To fix this, specify 'future.seed=TRUE'. This
ensures that proper, parallel-safe random numbers are produced via a parallel
RNG method. To disable this check, use 'future.seed = NULL', or set option
'future.rng.onMisuse' to "ignore". [future 'future_sapply-3'
(cbd2217c174b7bf544881e02d5898d6a-111); on
cbd2217c174b7bf544881e02d5898d6a@NTNU-C02DD00KML7J<9793>]

```
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

Warning: UNRELIABLE VALUE: One of the 'future.apply' iterations
('future_sapply-4') unexpectedly generated random numbers without declaring so.
There is a risk that those random numbers are not statistically sound and the
overall results might be invalid. To fix this, specify 'future.seed=TRUE'. This
ensures that proper, parallel-safe random numbers are produced via a parallel
RNG method. To disable this check, use 'future.seed = NULL', or set option
'future.rng.onMisuse' to "ignore". [future 'future_sapply-4'
(cbd2217c174b7bf544881e02d5898d6a-112); on
cbd2217c174b7bf544881e02d5898d6a@NTNU-C02DD00KML7J<9793>]

Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

Warning: UNRELIABLE VALUE: One of the 'future.apply' iterations
('future_sapply-5') unexpectedly generated random numbers without declaring so.
There is a risk that those random numbers are not statistically sound and the
overall results might be invalid. To fix this, specify 'future.seed=TRUE'. This
ensures that proper, parallel-safe random numbers are produced via a parallel
RNG method. To disable this check, use 'future.seed = NULL', or set option
'future.rng.onMisuse' to "ignore". [future 'future_sapply-5'
(cbd2217c174b7bf544881e02d5898d6a-113); on
cbd2217c174b7bf544881e02d5898d6a@NTNU-C02DD00KML7J<9793>]

Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

Warning: UNRELIABLE VALUE: One of the 'future.apply' iterations
('future_sapply-6') unexpectedly generated random numbers without declaring so.
There is a risk that those random numbers are not statistically sound and the
overall results might be invalid. To fix this, specify 'future.seed=TRUE'. This
ensures that proper, parallel-safe random numbers are produced via a parallel
RNG method. To disable this check, use 'future.seed = NULL', or set option
'future.rng.onMisuse' to "ignore". [future 'future_sapply-6'
(cbd2217c174b7bf544881e02d5898d6a-114); on
cbd2217c174b7bf544881e02d5898d6a@NTNU-C02DD00KML7J<9793>]
```
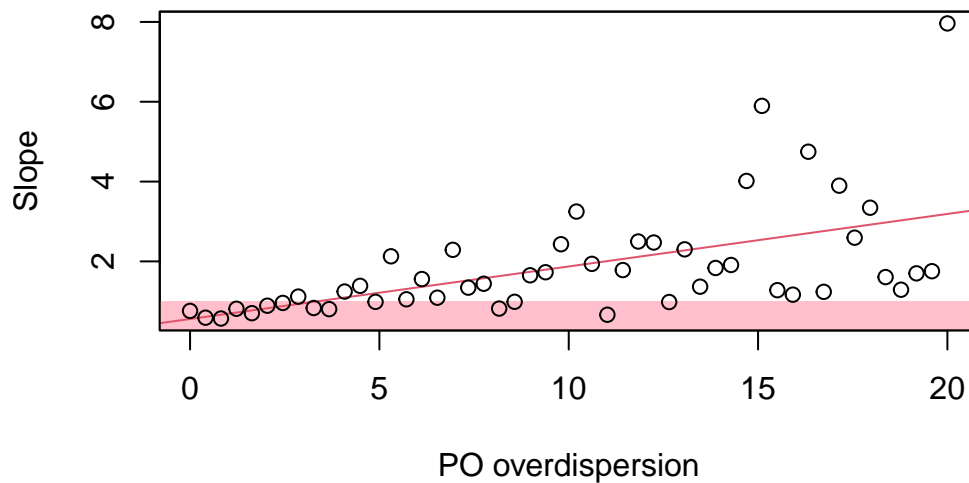

Figure 2: Estimated slopes for calibration of predictions to simulated presence absence data, from Maxent model with linear features fitted to simulated PO data with overdispresion in the linear predictor.

Comparison with Traditional Indicators

In general there is a positive correlation with AUC (Figure 3), but smaller correlations with TSS. As with AUC, the slope can tell us which species have bad models, i.e. where treating a predicted presence as an absence is more likely to be correct. Overall the data, 35% of the species have an AUC below 0.7, and 4% have an AUC below 0.5.
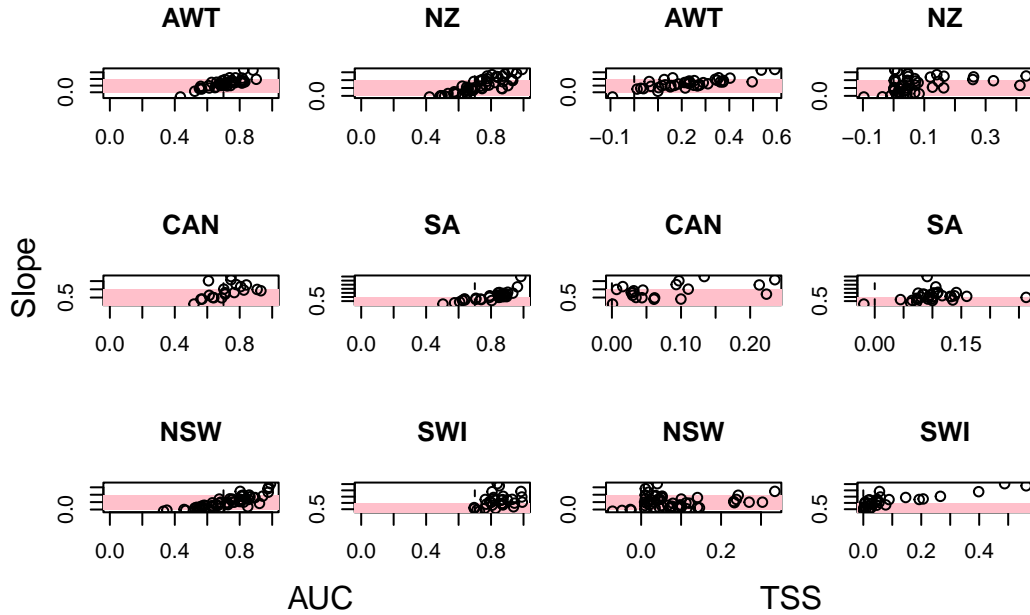
Figure 3: Estimated intercept and slope for calibration of predictions to presence absence data, from Maxent model with all features.

Correlations between models fitted to PA and PO data

Most of the correlations between the models fitted to the PA and PO data were weakly positive, with 1% being negative, and 44% having a correlation below 0.6 (Figure 4). This contrasts with the simulated data where only 86% of the simulated corresations were below 0.6.

Plots of predictions from the MaxEnt model on the validation data model plotted against predictions for a model with MaxEnt features fitted to the presence-absence data.

Species nz21 & swi12 have negative correlations in the simulated data, so let's look at them...

Correlations with bias correction

Discussion

Traditional assessment of SDMs has relied on calculating standard statistics, such as AUC or FSS, through internal or external validation. Here we have shown that it is fruitful to examine the validation more deeply. In particular, we show that (a) we are able to calibrate the fitted model to the validation model, (b) the fitted models often fail, producing predictions that are either counter-intuitive or wrong, and (c) that through plotting predictions we can see how bad our models are.
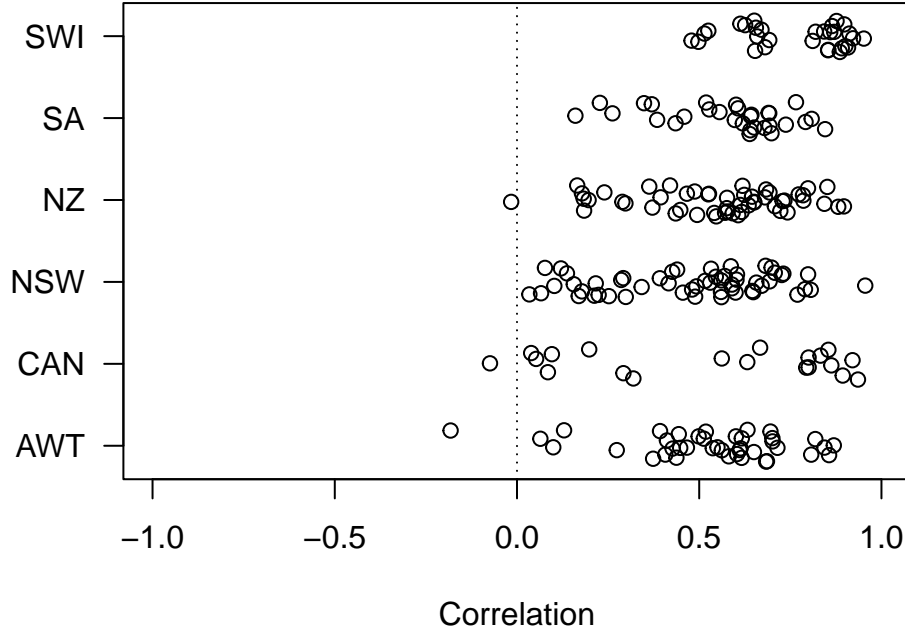
8

Figure 4: Correlations between predictions of Maxent models fitted to presence-only data and presence-absence data, predictions made on presence-absence locations

The simulations show that when the model is correct, the correlations between the models fitted to the PA and PO data are not perfect, but are (unsurprisingly) higher than with the real data. This suggests

Overall, our results are best explained by the confluence of two statistical aphorisms: Box's famous "All models are wrong, but some are useful" (Box (1979)), and Abelson's first law of statisics: chance is lumpy (Abelson (1995)). Because of the lumpiness of chance, the different covariates are correlated, as is common with spatial data (Dormann (2007)). This means that, because the model is wrong, it is easy for it to select the wrong covariates to explain the data: we can see this for nz21 and swi12 where, even when the true model is one of the candidate models, fits to the two diffeent data sets chose different covariates.

Selecting the wrong covariate is not a problem for prediction, as long as the correlation with the "true" covariate is stable across all possible data (REF). But if this changes, the models can be wildly different.
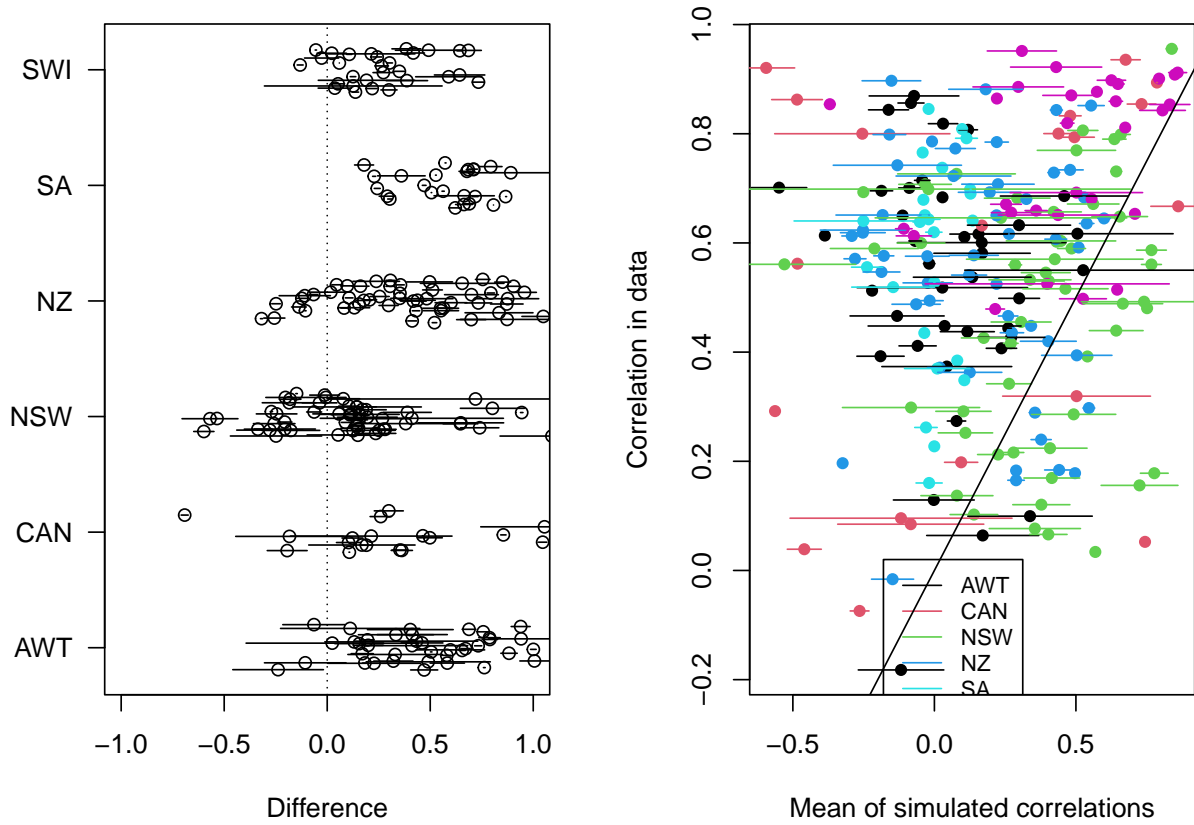
Figure 5: Differences between correlations between predictions of Maxent models fitted to presence-only data and presence-absence data, predictions made on presence-absence locations, from actual data and simulated data. Left: differences, right: Correlations plotted agaionst each other
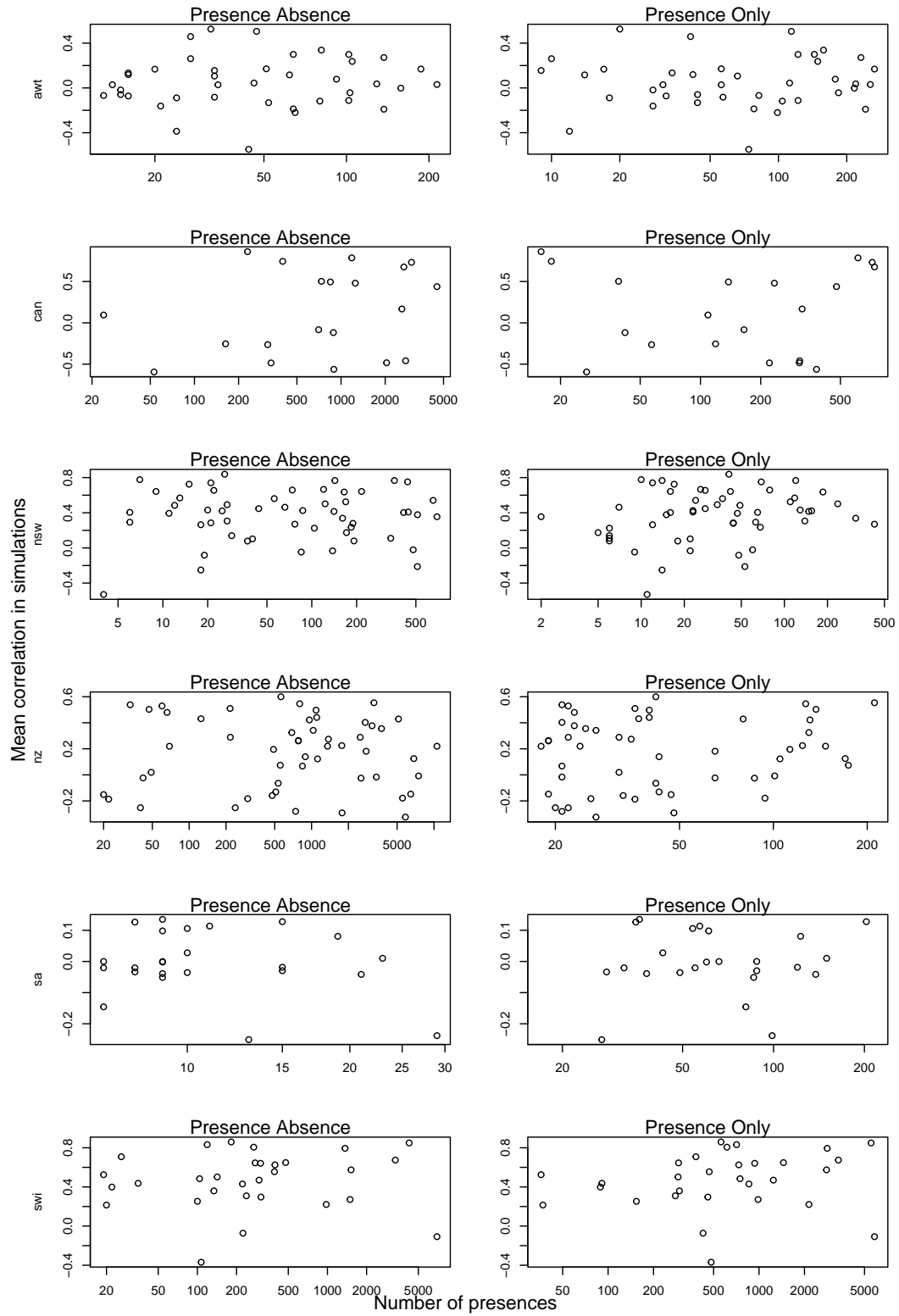
Figure 6: Effects of sample size (i.e. number of presences) on correlations between predictions of Maxent models fitted to presence-only data and presence-absence data, predictions made on presence-absence locations11
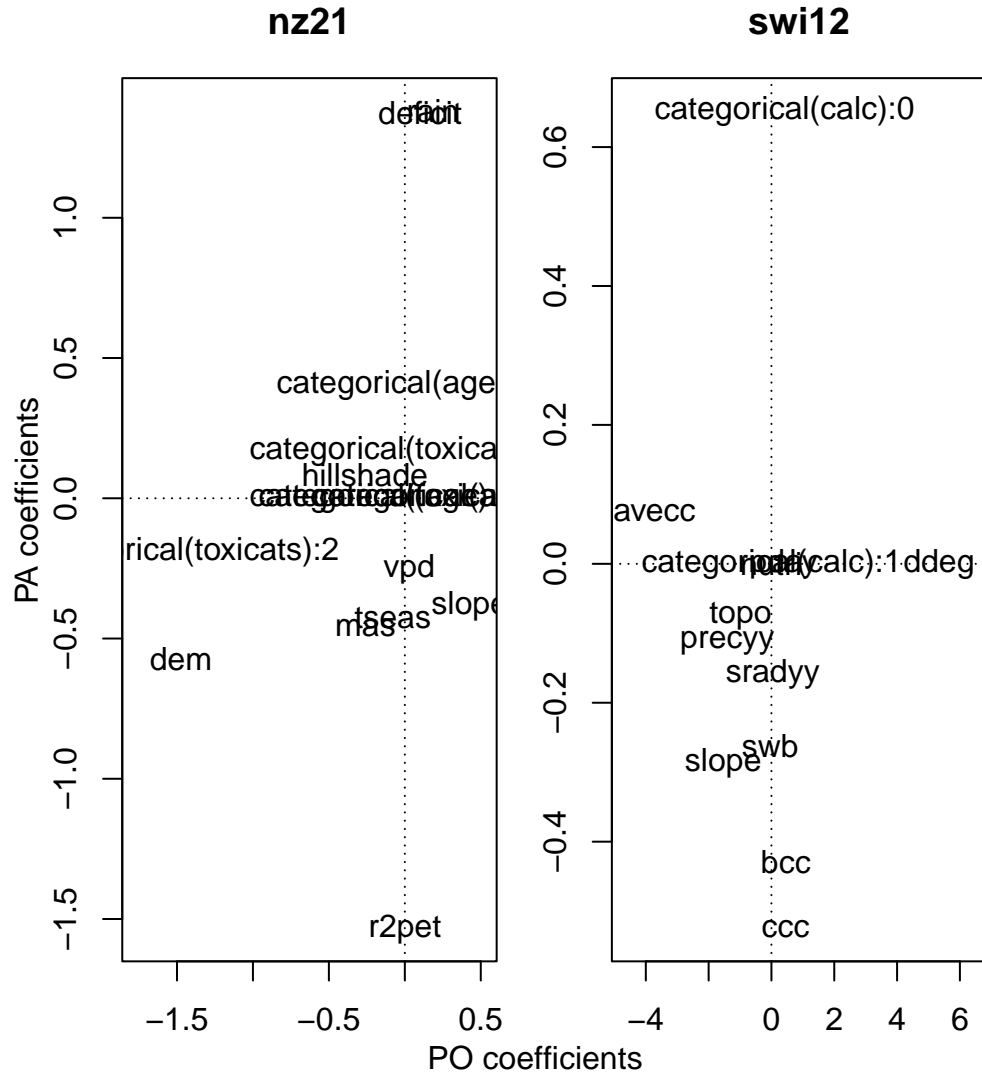
Figure 7: Estimates of parameters from fits of MaxEnt models to presence-only and presence absence data for species nz21 and swi12.
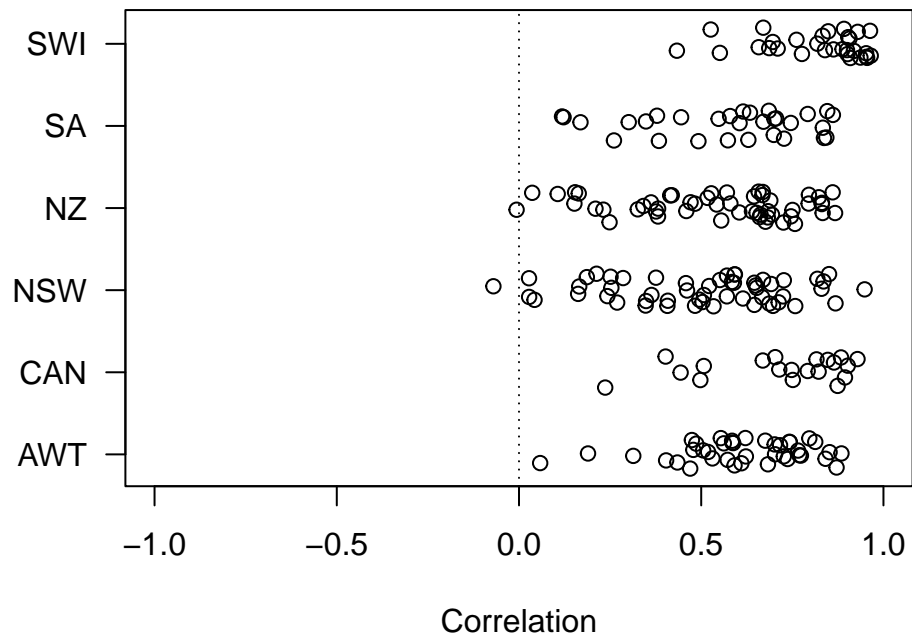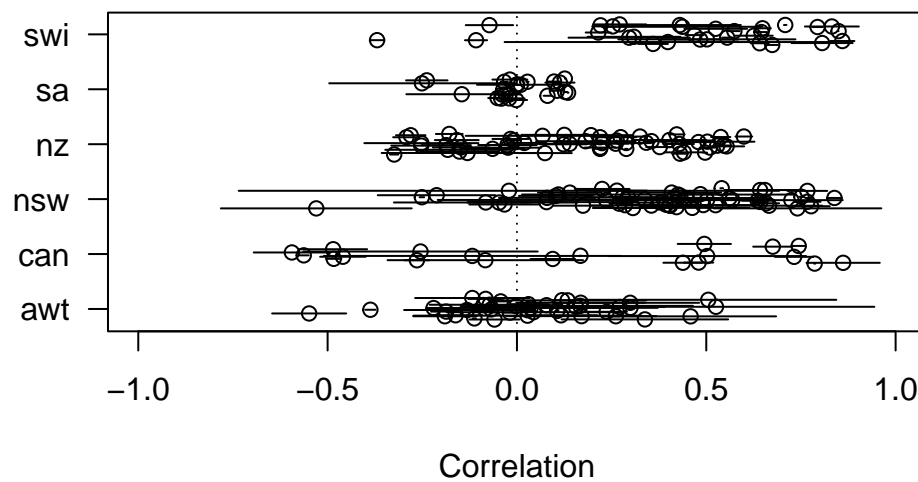
Figure 8

References

```
# Want to do this at the end
library(quarto)
quarto_render("ValidatingSDMsSupplInfo.qmd") # all formats
```

A. Lee-Yaw, Julie, Jenny L. McCune, Samuel Pironon, and Seema N. Sheth. 2022. "Species Distribution Models Rarely Predict the Biology of Real Populations." *Ecography* 2022 (6): e05877. https://doi.org/https://doi.org/10.1111/ecog.05877.

Aarts, Geert, John Fieberg, and Jason Matthiopoulos. 2012. "Comparative Interpretation of Count, Presence–Absence and Point Methods for Species Distribution Models." *Methods in Ecology and Evolution* 3 (1): 177–87. https://doi.org/https://doi.org/10.1111/j.2041-210X.2011.00141.x.

Abelson, R. P. 1995. *Statistics as Principled Argument.* L. Erlbaum Associates. https://books.google.no/books?id=TgmbosIA7N0C.

Box, G. E. P. 1979. "Robustness in the Strategy of Scientific Model Building." In. https://api.semanticscholar.org/CorpusID:121999337.

Carroll, Raymond J., Clifford H. Spiegelman, K. K. Gordon Lan, Kent T. Bailey, and Robert D. Abbott. 1984. "On Errors-in-Variables for Binary Regression Models." *Biometrika* 71 (1): 19–25. http://www.jstor.org/stable/2336392.

Dormann, Carsten F. 2007. "Effects of Incorporating Spatial Autocorrelation into the Analysis of Species Distribution Data." *Global Ecology and Biogeography* 16 (2): 129–38. https://doi.org/https://doi.org/10.1111/j.1466-8238.2006.00279.x.

Elith, Jane, Steven J. Phillips, Trevor Hastie, Miroslav Dudík, Yung En Chee, and Colin J. Yates. 2011. "A Statistical Explanation of MaxEnt for Ecologists." *Diversity and Distributions* 17 (1): 43–57. https://doi.org/https://doi.org/10.1111/j.1472-4642.2010.00725.x.

Elith, J., Graham, C.H., Valavi, R., Abegg, et al. 2020. "Presence-Only and Presence-Absence Data for Comparing Species Distribution Modeling Methods." *Biodiversity Informatics* 15: 69–80. https://doi.org/10.17161/bi.v15i2.13384.

Fithian, William, and Trevor Hastie. 2013. "Finite-sample equivalence in statistical models for presence-only data." *The Annals of Applied Statistics* 7 (4): 1917–39. https://doi.org/10.1214/13-AOAS667.

Phillips, Steven, and Miroslav Dudík. 2008. "Generative and Discriminative Learning with Unknown Labeling Bias." In *Advances in Neural Information Processing Systems*, edited by D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou. Vol. 21. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2008/file/9cf81d8026a9018052c429cc4e56739b-Paper.pdf.

Renner, Ian W., and David I. Warton. 2013. "Equivalence of MAXENT and Poisson Point Process Models for Species Distribution Modeling in Ecology." *Biometrics* 69 (1): 274–81. https://doi.org/https://doi.org/10.1111/j.1541-0420.2012.01824.x.

Smith, Jeffrey R., and Jonathan M. Levine. 2025. "Linking Relative Suitability to Probability of Occurrence in Presence-Only Species Distribution Models: Implications for Global Change Projections." *Methods in Ecology and Evolution* 16 (4): 854–65. https://doi.org/https://doi.org/10.1111/2041-210X.70003.

Valavi, Roozbeh, Gurutzeta Guillera-Arroita, José J. Lahoz-Monfort, and Jane Elith. 2022. "Predictive Performance of Presence-Only Species Distribution Models: A Benchmark Study with Reproducible Code." *Ecological Monographs* 92 (1): e01486. https://doi.org/https://doi.org/10.1002/ecm.1486.

Warton, David I., and Leah C. Shepherd. 2010. "Poisson point process models solve the

'pseudo-absence problem' for presence-only data in ecology." *The Annals of Applied Statistics* 4 (3): 1383–1402. https://doi.org/10.1214/10-AOAS331.

Yackulic, Charles B., Richard Chandler, Elise F. Zipkin, J. Andrew Royle, James D. Nichols, Evan H. Campbell Grant, and Sophie Veran. 2013. "Presence-Only Modelling Using MAXENT: When Can We Trust the Inferences?" *Methods in Ecology and Evolution* 4 (3): 236–43. https://doi.org/https://doi.org/10.1111/2041-210x.12004.