Generating Binary Species Range Maps – Supplementary Material

Filip Dorm[1] Christian Lange[1] Scott Loarie[2] Oisin Mac Aodha[1]

[1]University of Edinburgh [2]iNaturalist

A Additional Results

In this section, we present additional results and analysis. Unless stated otherwise, to generate these results we used the outputs from a single $\mathcal{L}_{\text{AN-full}}$ SDM and binarized the output using the LPT-R approach.

How much does the performance vary for different taxonomic groups? As seen in Fig. A1 for the IUCN dataset, LPT-R outperforms the other two approaches for the four different coarse taxonomic classes: amphibians, birds, mammals, and reptiles. This indicates that the results are stable across widely different taxonomic groups.

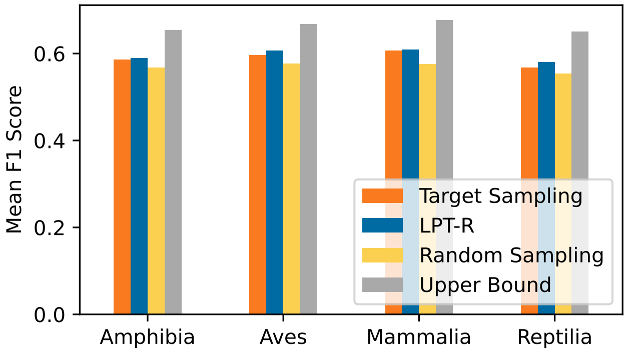

Fig. A1: Results across different taxonomic groups. Performance of the $\mathcal{L}_{\text{AN-full}}$ model on the IUCN task presented as the mean F1 score per taxonomic group.

How does the number of training samples influence the quality of binarized ranges? In Fig. A2 we display the relationship between the number of training samples per species against the F1 score, *i.e.* the measure of quality of the predicted binary range maps. Results are reported separately for the IUCN and S&T datasets. The overall trend is that the F1 score increases together with the number of training samples, *i.e.* species with more training presence observations have better predicted ranges.

What is the distribution of F1 scores across species? In Fig. A3 we display a histogram for the F1 scores for LPT-R on the IUCN dataset. We can see that most species obtain an F1 score of between 0.6 and 0.7 and that the

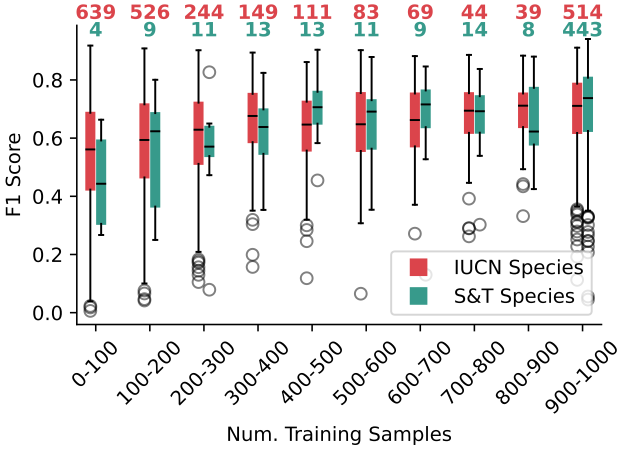

Fig. A2: Performance against number of training examples. Here we group species depending on how many training presence observations they have. The number of species for each bin is written on top of each box plot. The F1 score is calculated for the `LPT-R` method and the results are reported separately for the IUCN and S&T datasets. In general, performance improves with the number of training observations.

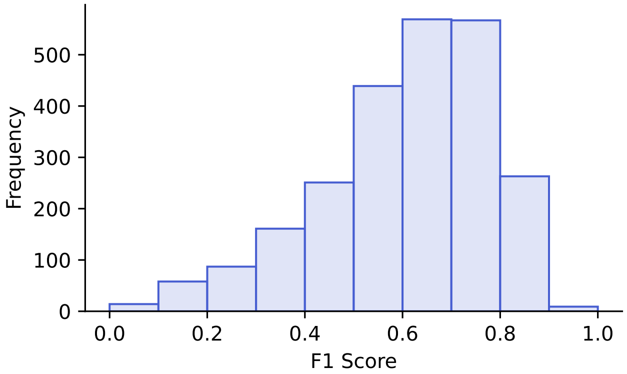

Fig. A3: Per-species binned performance. Histogram of scores on the IUCN dataset for $\mathcal{L}_{\text{AN-full}}$ binarized using `LPT-R`. The x-axis represents binned F1 score, and the y-axis is the number of species in each bin. In general, we observe that the distribution is skewed to the right.

distribution is skewed to the right. This indicates that thresholding results in plausible binary range maps for most species.

How many species obtain a boost in image classification performance as a result of using a binary range geo prior? In Fig. A4 we illustrate how using different geo priors influences the classification accuracy for computer vision models. We see that compared to the baseline of using continuous SDM predictions as a prior, the binarized range map results in fewer species with reduced performance as a result of using a prior (see left side of plot).

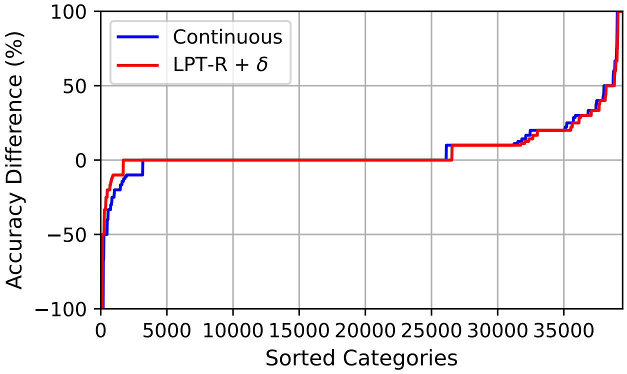

Fig. A4: Per-species image classification performance improvement. Here we sort species (*i.e.* categories) in the fine-grained image classification task according to how much the classification accuracy improves after using an SDM as a geo prior. Thus an accuracy difference of 0, indicates that the prior does not help for that particular species. The sorted accuracies after applying the original continuous predictions as a geo prior are shown in blue. Red shows the improvement resulting from using a binarized range map from our `LPT-R` with a small δ added.

How well do the thresholding techniques perform with models trained with additional environmental input features? In addition to the experiments in the main paper where we only evaluate SDMs that use coordinate features as input, here we evaluate SINR models that are trained using both environmental features and coordinates. SINR showed that these combined features yield the best continuous range estimation performance. Our results for binarizing these outputs with different thresholding techniques can be found in Tab. A1. The upper bound, *i.e.* the theoretically best possible range maps, for the model with environmental features is substantially higher at 73% (*vs.* 67.2% for the coordinate-only case). However, the resulting binarized range maps are not much more accurate than those of the coordinate-only variant. For the environmental model, the `Target Sampling` method outperforms the `LPT-R` method. This is consistent with earlier experiments where `Target Sampling` was superior in the S&T task, a set of species with more training data for which we expect the model to generate more accurate range maps. As the higher upper bound shows, there is potential for more accurate range maps, and this increased accuracy allows `Target Sampling` to outperform our `LPT-R`. The threshold classifier using an MLP also performs well, indicating a stronger correlation between model weights and optimal thresholds for certain model types. However, since the scores for this method are calculated using only a 25% subset of species from the evaluation set, conclusions drawn from these results should be approached with caution.

What impact does the choice of percentile have on `LPT-R`? In Tab. A2 we evaluate how varying the percentiles for the `LPT-R` method impacts performance across different models and loss functions. $\mathcal{L}_{\text{AN-SLDS}}$ seems to perform best with higher percentiles for `LPT-R` than the other two losses. This implies that there

Table A1: Binary range estimation performance of different thresholding techniques for models using additional environmental features. Similar to Tab. 1, here we report the average mean F1 score for five different random initializations of the $\mathcal{L}_{\text{AN-full}}$ SDM on the IUCN evaluation set. However, here the models are trained with coordinate *and* environmental input features, where the upper bound is 73.0%. (†) denotes threshold classifier scores which are computed on a 25% subset of the full evaluation set, as they are trained on the rest, and thus cannot be directly compared. (✓) indicates whether a thresholding technique uses true absences (TA), pseudo-absences (PA), or one single overall threshold (ST). Bold entries indicate best methods, and underline are second best.

Thresholding Method	ST	PA	TA	↑ Mean F1	↓ Upper Bound $\varDelta$
Threshold Classifier - RF †			✓	60.5	−12.5
Threshold Classifier - MLP †			✓	63.7	−9.3
Single Fixed Threshold - 0.5	✓			44.0	−29.0
Single Best Threshold - 0.1	✓		✓	$\underline{62.3}$	−10.7
Random Sampling - #Absences=#Presences		✓		54.4	−18.6
Random Sampling - #Absences=5#Presences		✓		58.5	−14.5
Random Sampling - #Absences=10#Presences		✓		57.8	−15.2
Random Sampling - 100 Absences		✓		55.5	−17.5
Random Sampling - 1000 Absences		✓		59.8	−13.2
Random Sampling - 10000 Absences		✓		45.1	−27.9
Target Sampling		✓		**62.8**	−10.2
Mean Predicted Threshold				42.5	−30.5
Lowest Presence Threshold (LPT)				41.6	−31.4
Lowest Presence Threshold - Robust (LPT-R)				60.2	−12.8

is more ‘noise’ associated with the presences used for identifying the thresholds, *i.e.* more of them need to be discarded when selecting an appropriate threshold. These results show that `LPT-R` can and should be tuned to the specific model it is applied to, ideally using a held-out validation set. Note, with the exception of the results in Tab. A2, this value was set to 5% by default for all other experiments and is not tuned.

B Additional Methods and Baselines

Here, we present some additional baselines to help contextualize the performance of the evaluated thresholding techniques. These baselines all use the held-out test set directly to find optimal thresholds. As a result, this is clearly not a viable approach but serves as a benchmark for the rest of the experiments.

Performance Upper Bound. First, we describe how we compute an upper bound on the possible mean F1 scores for the IUCN dataset. These values were used to represent the ‘Upper Bound $\varDelta$’ in Tab. 2, where we used the test data to select the optimal threshold for each species. These thresholds were obtained by generating predictions for each species for the test locations. Then for each species, the threshold is set to each unique value in the predictions until the one that maximized the F1 score was found. This means that each species had its own unique F1 score set. The upper bounds obtained are 67.2%, 60.1%, and 47.4% for $\mathcal{L}_{\text{AN-full}}$, $\mathcal{L}_{\text{AN-SSDL}}$, and $\mathcal{L}_{\text{AN-SLDS}}$, respectively. These results are the

Table A2: Impact of varying the percentile hyperparameter for `LPT-R`. Here, we report the mean F1 score for different variants of `LPT-R`, *i.e.* where we use different percentiles when setting the threshold. The scores are computed for the IUCN set as the average of five random initializations of SDMs, both in the setting where the models are only trained on coordinate inputs ('Crds.') and when trained with coordinates together with environmental inputs ('Env. + Crds.'). The scores are presented for different training losses. We can see that the models that also use environmental inputs perform best with a slightly larger percentile over our default of 5%.

Model	Thresholding Method	↑ Mean F1 Crds.	↑ Mean F1 Env. + Crds.
$\mathcal{L}_{\text{AN-full}}$	LPT	54.3	41.6
	LPT@2.5	60.6	57.0
	LPT@5.0	**60.8**	60.2
	LPT@7.5	60.0	**60.9**
	LPT@10	58.9	60.4
	LPT@12.5	57.6	59.3
	LPT@15	56.2	57.8
$\mathcal{L}_{\text{AN-SSDL}}$	LPT	49.9	35.2
	LPT@2.5	**53.7**	50.5
	LPT@5.0	53.5	**53.4**
	LPT@7.5	52.5	**53.4**
	LPT@10	51.3	52.5
	LPT@12.5	50.0	51.1
	LPT@15	48.7	49.5
$\mathcal{L}_{\text{AN-SLDS}}$	LPT	29.9	28.9
	LPT@2.5	36.9	41.6
	LPT@5.0	39.4	47.2
	LPT@7.5	40.7	50.7
	LPT@10	41.5	52.8
	LPT@12.5	42.0	54.2
	LPT@15	**42.2**	**55.1**

average of five runs with different random initializations of the input SDM. The upper bound for the ensemble of five $\mathcal{L}_{\text{AN-full}}$ SDMs is 68.3%. The maximum performance for each of these losses is represented by these scores, meaning that a good thresholding technique would find thresholds that match these as close as possible, and obviously can not be better. Similarly, the upper bounds were calculated for the S&T dataset. These scores were 76.1%, 69.7%, and 76.2% for $\mathcal{L}_{\text{AN-full}}$, $\mathcal{L}_{\text{AN-SSDL}}$, and $\mathcal{L}_{\text{AN-SLDS}}$, respectively, and 76.9% for the ensamble.

Subsampling Expert Data. We also explore the impact of using a fraction of the high-quality true data to compute the thresholds. One way of setting the thresholds for species is by using a small amount of true presence-absence data. For these experiments on the IUCN dataset, a subsample of the expert evaluation data was used (1%, 5%, and 10% randomly sampled) to maximize the F1 score and select the threshold. Since the subsample is random, species with a low number of presences might not have any presence-locations at all in the subsample. These results are presented in Tab. A3. Although these scores are high and close to the upper bound performance (obtained with 100% of the data), this method is again not viable for species without expert-derived range

Table A3: Using test data to determine thresholds. Here we report the mean F1 score for an $\mathcal{L}_{\text{AN-full}}$ SDM when different amounts of evaluation data are used to determine the threshold for each species. A fraction of 1% indicates that only 1% of the evaluation presence-absence data is utilized to identify the threshold for each species, and the remaining data is used for evaluation.

Model	Fraction Used	↑ Mean F1
$\mathcal{L}_{\text{AN-full}}$	1%	63.7
	5%	65.7
	10%	66.2
	100%	66.4

maps. As can be seen in this experiment, even a small amount of true data enables the identification of almost perfect thresholds for binarizing range maps. However, in practice, a sample of even 1% still requires an infeasible amount of survey locations to be checked. For context, the `LPT-R` approach obtains a mean F1 of 60.8%, without using pseudo or true absence data, which indicates that it is still quite competitive.

C SINR Training Losses

Here we outline the different loss functions used in SINR to train SDMs which are evaluated in Tab. 2 in the main paper. SINR introduces the following losses: "assume negative loss (same species, different location)" $\mathcal{L}_{\text{AN-SSDL}}$, "assume negative loss (same location, different species)" $\mathcal{L}_{\text{AN-SLDS}}$, and "full assume negative loss" $\mathcal{L}_{\text{AN-full}}$. The description of the $\mathcal{L}_{\text{AN-full}}$ loss can be found in Sec. 3.

The $\mathcal{L}_{\text{AN-SSDL}}$ loss pairs each species observation with a different randomly generated location as a negative (*i.e.* pseudo-absence). These randomly generated pseudo-absences are incorporated into the loss function as follows:

$$\mathcal{L}_{\text{AN-SSDL}}(\hat{\mathbf{y}}_i, \mathbf{z}_i) = -\frac{1}{n_{\text{pos}}} \sum_{j=1}^{S} \mathbb{1}_{[z_{ij}=1]} \left[\log(\hat{y}_{ij}) + \log(1 - \hat{y}'_j)\right], \tag{1}$$

where a randomly chosen location $\mathbf{r} \sim \text{Uniform}(\mathcal{X})$ is used together with $n_{\text{pos}} = \sum_{j=1}^{S} \mathbb{1}_{[z_{ij}=1]}$ to generate $\hat{\mathbf{y}}' = h_\phi(f_\theta(\mathbf{r}))$. In this way random absences are generated across the globe.

The $\mathcal{L}_{\text{AN-SLDS}}$ loss, on the other hand, associates every species observation with a pseudo-absence at the same location for a different species. This generates pseudo-absences that align with the distribution of the presence training data, so is referred to as target background sampling. This loss is computed as:

$$\mathcal{L}_{\text{AN-SLDS}}(\hat{\mathbf{y}}_i, \mathbf{z}_i) = -\frac{1}{n_{\text{pos}}} \sum_{j=1}^{S} \mathbb{1}_{[z_{ij}=1]} \left[\log(\hat{y}_{ij}) + \log(1 - \hat{y}_{ij'})\right], \tag{2}$$

where $j' \sim \text{Uniform}(\{j : z_{ij} \neq 1\})$.

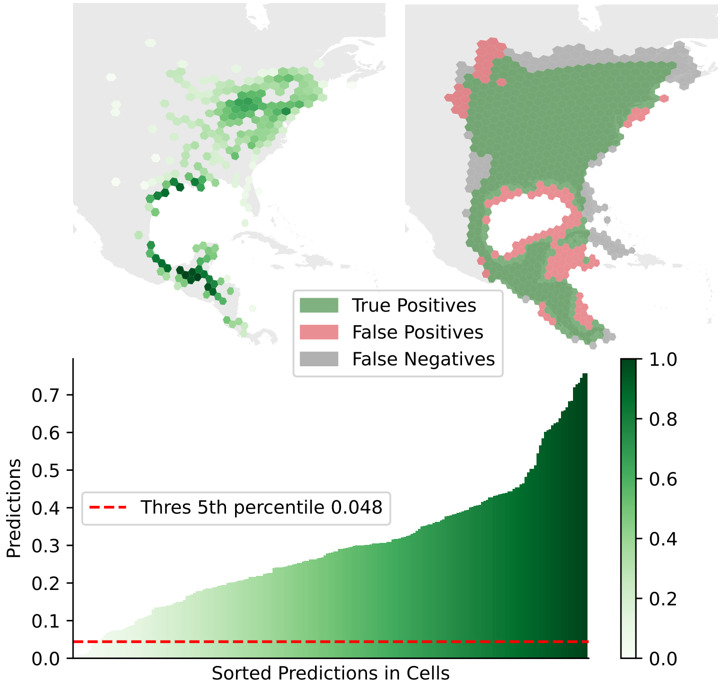

Fig. A5: Threshold selection for `LPT-R`. Here we illustrate the main steps of the `LPT-R` approach for the Wood Thrush. More specifically, the top left plot shows how the predictions are collected for all of the H3 cells of a specified resolution where the species has been observed. The 5th percentile of these predictions is then calculated and used as a threshold. The map is then binarized so that all cells with a prediction score higher than the threshold are marked as presences. In the top right plot, the output of this process is compared to the expert-derived range map. Green cells indicate true positives, red false positives, and dark gray false negatives.

D Additional Visualizations

Finally, we include Fig. A5 and Fig. A6 to visualize how different binarization methods work. Fig. A5 illustrates how thresholds are set through `LPT-R` using the *Wood Thrush* as an example. Similarly, this bird species is used to show how pseudo-absences are generated through `Target Sampling` and `Random Sampling` in Fig. A6.

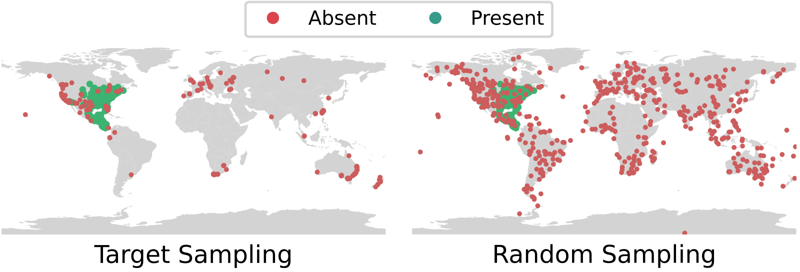

Fig. A6: Pseudo-absence generation. Here we visualize how the pseudo-absences (*i.e.* 'Absent') are generated for the two common sampling methods, target and random sampling for the Wood Thrush. For `Target Sampling`, the absences are clustered where most of the observations have been reported to iNaturalist, *i.e.* North America, Europe, and parts of Australasia. In contrast, the `Random Sampling` absences are uniformly distributed across the globe.