Modeling the distribution of invasive species in small islands under future climates

Vasco Cabral[1], Ana Almeida Matos[1,2], Paulo A. V. Borges[3], Luís Borda-de-Água[4,5], and Eduardo Brito de Azevedo[6]

[1] Instituto Superior Técnico, Universidade de Lisboa
[2] Instituto de Telecomunicações
[3] cE3, Universidade dos Açores
[4] CIBIO, Campus de Vairão, Universidade do Porto
[5] CIBIO, Instituto Superior de Agronomia, Universidade de Lisboa
[6] Group of Climate, Meteorology and Global Change, Universidade dos Açores

The biodiversity of the Azores Archipelago faces a significant threat due to the introduction and proliferation of invasive species [4]. In response to this challenge, data science has emerged as a powerful tool in the field of conservation, in particular through the application of species distribution modeling (SDM) [9]. SDMs create predictions of the suitability of habitats for a species by analyzing the relationship between an occurrence of the species and a set of environmental variables that describe their location [9]. SDMs can inform monitoring and containment strategies for invasive species. However, developing a data preparation and modeling pipeline for predicting invasive species distribution in the Azores presents several challenges, including variable selection, multicollinearity, and limited sample sizes [14]. Uncertainty is a major concern in this modeling task, especially due to the objective of projecting the generated models onto future climate scenarios [2]. To address this, a common strategy is ensemble forecasting [14], which combines predictions from various models using consensus methods.

The project described in the article aims to develop a robust data preparation and modeling pipeline to predict the habitat suitability of invasive species in the Azores Archipelago under current and future climate scenarios. Initial findings suggest that ensemble modeling is a useful tool for reducing uncertainty, and also the prospect for a more automated predictor selection. We performed test cases for two species: *Gunnera tinctoria* (Molina) Mirb. and *Paspalum vaginatum* (Sw.) P.Fourn. In this summary, we will present our results for *Paspalum vaginatum*.

The considered study area is the Azores Archipelago, a group of nine volcanic islands situated in the North Atlantic Ocean. The species occurrence data is sourced from the Azores Biodiversity Platform (http://azoresbioportal.uac.pt/) [3] and focuses on the top 100 invasive species in the region [17]. The dataset predominantly comprises species from the Plantae Kingdom, followed by Animalia. *Paspalum vaginatum*, a perennial grass, has registered occurrences in Faial.

In this study, climate predictors were obtained from the CIELO model, a widely used model for downscaling global circulation models and climate change scenarios in small oceanic islands [7,8,16]. Two scenarios were considered: RCP 4.5, a baseline with greenhouse emissions peaking around 2040 and gradually

decreasing, and RCP 8.5, a worst-case scenario with emissions continuing to rise until 2100. The climate models included 19 climate predictors. In both statistical learning and machine learning methods, it is essential to avoid training models with highly collinear predictors. Multicollinearity can lead to poor model performance, particularly in multiple regression models, where important variables may be considered insignificant due to the correlation with other predictors [9]. Moreover, including unnecessary predictors in machine learning can lead to the curse of dimensionality. To address multicollinearity among climate predictors, the study utilized the Variance Inflation Factor (VIF). Predictors with a VIF exceeding the threshold of 10 were removed. This process is repeated until no predictor has a VIF exceeding the threshold, ensuring the final set of predictors used is free from significant multicollinearity [12].

When training species distribution models that need presence-background data, it is necessary to sample background locations to capture the environmental range of the study area where the species has not been sampled [13]. Various techniques exist for background sampling, including random and disk-based selection. In this study, we used a model called SRE. The SRE model, an Environmental Envelope Method, defines the potential species range by establishing bounds based on environmental predictors at occurrence locations. Background locations were sampled outside the predicted suitable occurrence area determined by the SRE model [5].

Ensemble forecasting is used to predict the current and future distributions of the considered invasive species, incorporating both statistical and machine learning models. The statistical modeling includes Generalized Linear Models (GLM) and Generalized Additive Models (GAM), while the machine learning methods consist of artificial neural networks (ANN), gradient boost machines (GBM), and Maxent. To evaluate the performance of the models, k-fold cross-validation is used considering 5 folds, meaning that a model is trained on 4 folds and evaluated on the remaining one. The process is repeated 5 times, then the resulting performance measures are aggregated using the average of the values. To generate habitat suitability maps, each cell's suitability value (0-1000) is calculated using a consensus method over the values of the models outputs given the current or future climate predictor values for that cell. The consensus methods used include the mean, median, and committee averaging. The implementation of both the modeling technique and the data preparation pipeline was performed using R [15], specifically utilizing the biomod2 package [18].

The performance of the models is evaluated using a combination of threshold-dependent and threshold-independent measures. For threshold-independent evaluation, the area under the receiver operating characteristic curve (AUC) is used to assess the model's ability to distinguish between random presence and random background sites. AUC values above 0.9 are excellent [6]. The Boyce Index is also used, its values ranging from -1 to 1, where 0 indicates performance equal to a random model [11]. For threshold-dependent evaluation, the true skill statistic (TSS) is employed, with values ranging from -1 to 1. TSS values above 0.8

indicate good to excellent performance [1]. Models exceeding a TSS value of 0.7 are considered suitable for inclusion in the ensemble. To convert the continuous habitat suitability index into binary presences and absences, a threshold value is required. The chosen approach for this study utilizes a sensitivity and specificity-combined approach, selecting the point on the ROC curve closest to the top-left corner, the perfect model.

Considering the previously described data preparation and modeling approach, we achieved promising results for *Paspalum vaginatum*. A total of 26 occurrences of the species were recorded in Faial. Two datasets were generated through two runs of background sampling, each containing 104 background locations. Then we performed environmental predictor selection using iterative VIF elimination, removing 13 highly collinear predictors, followed by PCA to identify the five variables that made the most significant contributions to the first two dimensions (which explained 66.5% of dataset variance) while not being highly correlated. A total of 50 different models were obtained through cross-validation, each technique producing five models, over two runs, one for each of the different background datasets. These models exhibited excellent performance in AUC and in TSS metrics on validation, indicating their high predictive capability. However, GAM appears to suffer from overfitting, since it has considerably degraded performance when assessed against the validation data.

Table 1. *Paspalum vaginatum* single and ensemble model performance

Model	Metric	Calibration	Validation
ANN	AUC	0.973 ±0.0293	0.981±0.0210
ANN	TSS	0.940±0.0452	0.953±0.0457
GAM	AUC	0.956±0.0587	0.847±0.137
GAM	TSS	0.910±0.121	0.703±0.278
GBM	AUC	0.999±0.0223	0.986±0.0244
GBM	TSS	0.994±0.0162	0.960±0.0724
GLM	AUC	0.984±0.0376	0.968±0.0313
GLM	TSS	0.958±0.0727	0.938±0.0611
MAXENT	AUC	0.985±0.0066	0.976±0.0332
MAXENT	TSS	0.927±0.0311	0.9204±0.0774
RF	AUC	0.999±0.0007	0.983±0.0196
RF	TSS	0.989±0.0135	0.957±0.0444

Model/Consensus Method	Metric	Value
Ensemble/Committee Averaging	AUC	0.999
Ensemble/Committee Averaging	TSS	0.986
Ensemble/Mean	AUC	0.997
Ensemble/Mean	TSS	0.981
Ensemble/Median	AUC	0.996
Ensemble/Median	TSS	0.981

The ensemble models produced in this study showed outstanding performance, with nearly perfect TSS and AUC scores. The Boyce Index analysis con-

firmed that all ensemble models outperformed a random model, with the mean consensus method demonstrating the best performance. However, the median and committee averaging approach exhibited significantly poorer performance. The projections for *Paspalum vaginatum* in Faial under future climate scenarios show an expansion in the suitable area in 2040-2069 and 2070-2099. Compared to current conditions, the models predict an average increase of approximately 17% in the suitable area under RCP 4.5 for 2040-2069 and around 40% under RCP 8.5. In the timeframe 2070-2099, compared to 2040-2069, the models estimate an average increase of approximately 23% under RCP 4.5 and 35% under RCP 8.5.

Table 1. *Paspalum vaginatum* Ensemble model performance – Boyce Index

Model/Consensus Method	Calibration	Validation
Ensemble/Mean	0.997	0.784
Ensemble/Median	0.996	0.242
Ensemble/Committee Averaging	0.999	0.382

The performance of the generated models provides validation for the proposed pipeline, showing favorable results when evaluated through statistical measures. However, caution is necessary due to certain limitations. One limitation is the lack of independent data to validate the models, the other is that the testing was limited to two specific species, therefore not being representative of the effectiveness of the pipeline for the diverse set of invasive species in the Azores Archipelago. The currently utilized approach also has inherent limitations, such as considering only climate predictors as determinants of species' potential distribution, disregarding the importance of biotic factors [2]. Additionally, uncertainty arises from the diversity of global circulation models available, and small sample sizes [10]. An approach for solving the latter is to use occurrence data from the species' native range, but this comes with its own potential challenges.

In summary, the results demonstrate the ability to generate models that accurately capture habitat suitability in the current environment, as verified by statistical measures. The proposal has significant practical implications, reducing reliance on expert knowledge and increasing efficiency. However, two limitations require consideration: small sample sizes may lead to underestimation of suitable areas, and inherent uncertainty persists even with ensemble models. This proposal serves as a foundation for future research, towards establishing a potential standard for modeling invasive species under future climate scenarios in small oceanic islands. It uses an innovative background sampling technique, which utilizes simple presence-only models to inform the selection of background locations. It represents a contribution to the field while offering avenues for further exploration and refinement of SDM methodologies.

References

1. Allouche, O., Tsoar, A., Kadmon, R.: Assessing the accuracy of species distribution models: Prevalence, kappa and the true skill statistic (tss). Journal of Applied Ecology **43**(6), 1223–1232 (2006). https://doi.org/10.1111/j.1365-2664.2006.01214.x
2. Beale, C.M., Lennon, J.J.: Incorporating uncertainty in predictive species distribution modelling. Philosophical Transactions of the Royal Society B: Biological Sciences **367**(1586), 247–258 (2012). https://doi.org/10.1098/rstb.2011.0178
3. Borges, P.A., Gabriel, R., Arroz, A.M., Costa, A., Cunha, R.T., Silva, L., Mendonça, E., Martins, A.M., Reis, F., Cardoso, P.: The azorean biodiversity portal: An internet database for regional biodiversity outreach. Systematics and Biodiversity **8**(4), 423–434 (2010). https://doi.org/10.1080/14772000.2010.514306
4. Borges, P.A., Santos, A.M., Elias, R.B., Gabriel, R.: The azores archipelago: Biodiversity erosion and conservation biogeography. Encyclopedia of the World's Biomes p. 101–113 (2020). https://doi.org/10.1016/b978-0-12-409548-9.11949-9
5. Buckland, C.E., Smith, A.J., Thomas, D.S.: A comparison in species distribution model performance of succulents using key species and subsets of environmental predictors. Ecology and Evolution **12**(6) (2022). https://doi.org/10.1002/ece3.8981
6. Coetzee, B.W., Robertson, M.P., Erasmus, B.F., van Rensburg, B.J., Thuiller, W.: Ensemble models predict important bird areas in southern africa will become less effective for conserving endemic birds under climate change. Global Ecology and Biogeography **18**(6), 701–710 (2009). https://doi.org/10.1111/j.1466-8238.2009.00485.x
7. Dutra Silva, L., Brito de Azevedo, E., Vieira Reis, F., Bento Elias, R., Silva, L.: Limitations of species distribution models based on available climate change data: A case study in the azorean forest. Forests **10**(7), 575 (2019). https://doi.org/10.3390/f10070575
8. Ferreira, M.T., Cardoso, P., Borges, P.A., Gabriel, R., de Azevedo, E.B., Elias, R.B.: Implications of climate change to the design of protected areas: The case study of small islands (azores). PLOS ONE **14**(6) (2019). https://doi.org/10.1371/journal.pone.0218168
9. Franklin, J., Miller, J.A.: Mapping species distributions: Spatial inference and prediction. Cambridge University Press (2014)
10. Goberville, E., Beaugrand, G., Hautekèete, N.C., Piquot, Y., Luczak, C.: Uncertainties in the projection of species distributions related to general circulation models. Ecology and Evolution **5**(5), 1100–1116 (2015). https://doi.org/10.1002/ece3.1411
11. Hirzel, A.H., Le Lay, G., Helfer, V., Randin, C., Guisan, A.: Evaluating the ability of habitat suitability models to predict species presences. Ecological Modelling **199**(2), 142–152 (2006). https://doi.org/https://doi.org/10.1016/j.ecolmodel.2006.05.017
12. James, G., Witten, D., Hastie, T., Tibshirani, R.: 3.3.3.6. Collinearity, p. 101–102. Springer (2017)
13. Liu, C., Newell, G., White, M.: The effect of sample size on the accuracy of species distribution models: Considering both presences and pseudo-absences or background sites. Ecography **42**(3), 535–548 (2018). https://doi.org/10.1111/ecog.03188
14. Marmion, M., Parviainen, M., Luoto, M., Heikkinen, R.K., Thuiller, W.: Evaluation of consensus methods in predictive species distribution modelling. Diversity and Distributions **15**(1), 59–69 (2009). https://doi.org/10.1111/j.1472-4642.2008.00491.x

15. R Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria (2023), https://www.R-project.org/
16. Silva, L.D., Costa, H., de Azevedo, E.B., Medeiros, V., Alves, M., Elias, R.B., Silva, L.: Modelling native and invasive woody species: A comparison of enfa and maxent applied to the azorean forest. In: Pinto, A.A., Zilberman, D. (eds.) Modeling, Dynamics, Optimization and Bioeconomics II. pp. 415–444. Springer International Publishing, Cham (2017)
17. Silva, L., Ojeda Land, E., Rodríguez Luengo, J.: Invasive Terrestial Flora & Fauna of Macaronesia. TOP 100 in Azores, Madeira y Canarias. (01 2008)
18. Thuiller, W., Georges, D., Gueguen, M., Engler, R., Breiner, F., Lafourcade, B., Patin, R.: biomod2: Ensemble Platform for Species Distribution Modeling (2023), r package version 4.2-3-5