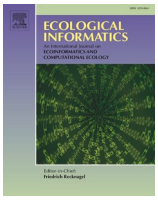
Ecological Informatics 75 (2023) 102080

Contents lists available at ScienceDirect

Ecological Informatics

journal homepage: www.elsevier.com/locate/ecolinf



Too many candidates: Embedded covariate selection procedure for species distribution modelling with the `covsel` R package

Antoine Adde [a,*], Pierre-Louis Rey [a], Fabian Fopp [b,c], Blaise Petitpierre [a,d], Anna K. Schweiger [e], Olivier Broennimann [a,f], Anthony Lehmann [g], Niklaus E. Zimmermann [b], Florian Altermatt [h,i], Loïc Pellissier [b,c], Antoine Guisan [a,f]

[a] Institute of Earth Surface Dynamics, Faculty of Geosciences and Environment, University of Lausanne, Lausanne, Switzerland
[b] Land Change Science Research Unit, Swiss Federal Institute for Forest, Snow and Landscape Research, WSL, Birmensdorf, Switzerland
[c] Ecosystems Landscape Evolution, Institute for Terrestrial Ecosystems, Department of Environmental System Sciences, ETH Zurich, Zurich, Switzerland
[d] InfoFlora, c/o Conservatoire et Jardin botaniques de Genève, Chambésy-Genève, Switzerland
[e] Department of Geography, Remote Sensing Laboratories, University of Zurich, Zurich, Switzerland
[f] Department of Ecology and Evolution, University of Lausanne, Lausanne, Switzerland
[g] EnviroSPACE, Institute for Environmental Sciences, University of Geneva, Geneva, Switzerland
[h] Department of Evolutionary Biology and Environmental Studies, University of Zurich, Zurich, Switzerland
[i] Department of Aquatic Ecology, Eawag: Swiss Federal Institute of Aquatic Science and Technology, Dübendorf, Switzerland

ARTICLE INFO

Keywords:
Automated covariate selection
Generalized additive model with null-space penalization
Generalized linear model with elastic-net regularization
Guided regularized random forest
Multicollinearity
Predictors
Species distribution models
R package

ABSTRACT

1. Selecting the best subset of covariates out of a panel of many candidates is a key and highly influential stage of the species distribution modelling process. Yet, there is currently no commonly accepted and widely adopted standard approach by which to perform this selection.
2. We introduce a two-step "embedded" covariate selection procedure aimed at optimizing the predictive ability and parsimony of species distribution models fitted in a context of high-dimensional candidate covariate space. The procedure combines a collinearity-filtering algorithm (Step A) with three model-specific embedded regularization techniques (Step B), including generalized linear model with elastic net regularization, generalized additive model with null-space penalization, and guided regularized random forest.
3. We evaluated the embedded covariate selection procedure through an example application aimed at modelling the habitat suitability of 50 species in Switzerland from a suite of 123 candidate covariates. We demonstrated the ability of the embedded covariate selection procedure to provide significantly more accurate species distribution models as compared to models obtained with alternative procedures. Model performance was independent of the characteristics of the species data, such as the number of occurrence records or their spatial distribution across the study area.
4. We implemented and streamlined our embedded covariate selection procedure in the `covsel` R package, paving the way for a ready-to-use, automated, covariate selection tool that was missing in the field of species distribution modelling. All the information required for installing and running the `covsel` R package is openly available on the GitHub repository https://github.com/N-SDM/covsel.

* Corresponding author at: Institute of Earth Surface Dynamics, Faculty of Geosciences and Environment, University of Lausanne, CH-1015 Lausanne, Switzerland.
E-mail address: antoine.adde@unil.ch (A. Adde).

https://doi.org/10.1016/j.ecoinf.2023.102080
Received 15 November 2022; Received in revised form 27 February 2023; Accepted 21 March 2023
Available online 22 March 2023
1574-9541/© 2023 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

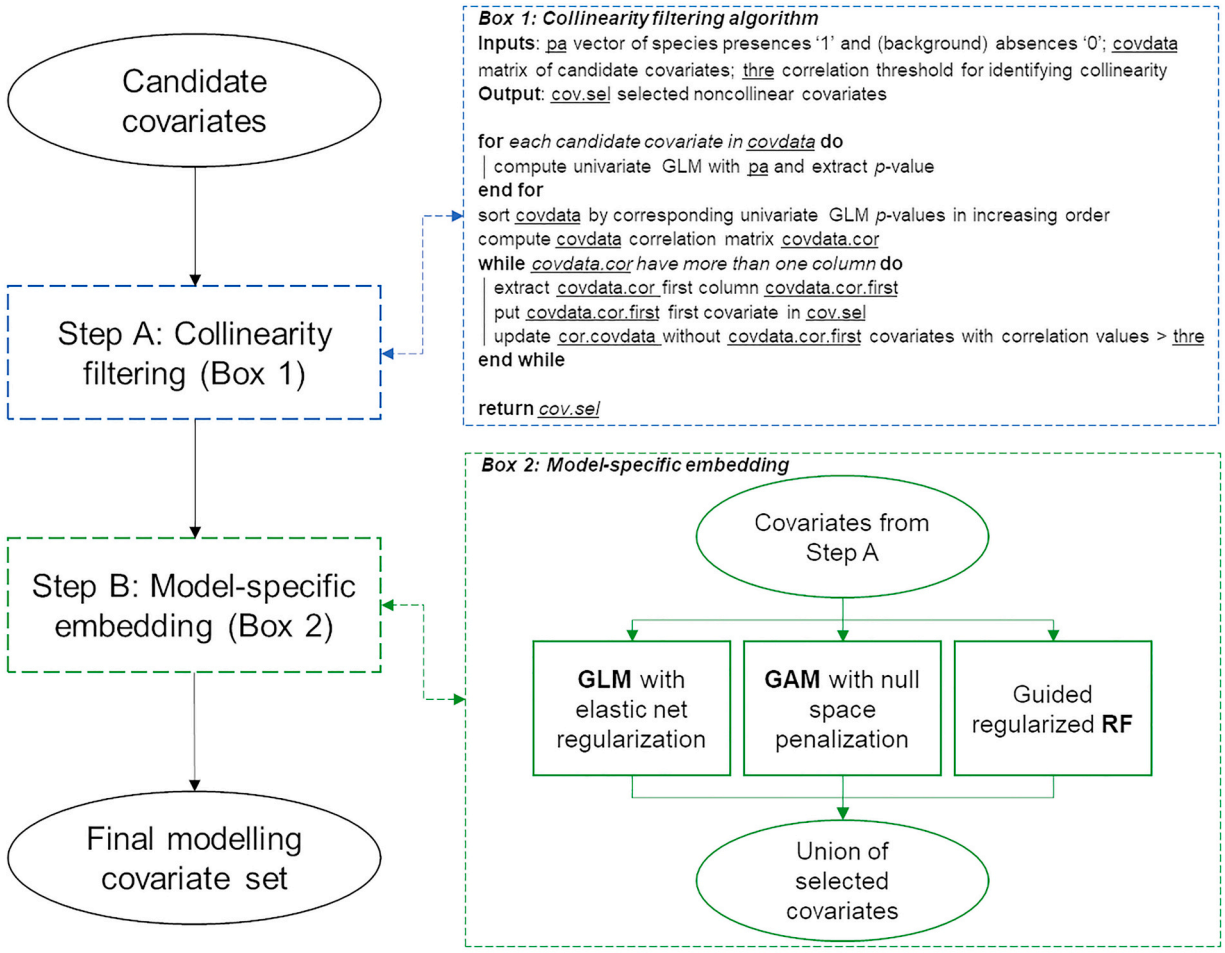


Fig. 1. Schema of the embedded covariate selection procedure proposed in this study. GLM: Generalized Linear Model. GAM: Generalized Additive Model. RF: Random Forest.

1. Introduction

Species distribution models (SDMs) relate species occurrence records with environmental covariates to predict and map species' habitat suitability (Franklin, 2010; Guisan et al., 2017; Peterson et al., 2011). Over the last decades, the implementation of new informatic tools and computational platforms have led to a significant increase in SDMs use (Araujo et al., 2019; Ferrier et al., 2016; Guisan et al., 2013). Despite recent efforts for standardizing SDM development steps and reporting (Araujo et al., 2019; Merow et al., 2019; Zurell et al., 2020), several methodological considerations still require further investigation. These include the covariate selection, a highly influential stage of the species distribution modelling process (Austin and Van Niel, 2011; Brun et al., 2020; Fourcade et al., 2018).

The goal of the covariate selection is to identify the "best" subset of covariates out of a panel of many candidates, both from ecological and statistical perspectives (Austin and Van Niel, 2011; Petitpierre et al., 2017; Yates et al., 2018). As highlighted in recent reviews (Fois et al., 2018; Fourcade et al., 2018; Melo-Merino et al., 2020), a common practice for covariate selection in SDM studies consists of building an expert-based set of ~20 to 30 covariates, which is then reduced in number after collinearity analyses. Collinearity analyses are usually based on variance inflation factors (VIFs) (Brauner and Shacham, 1998), or correlation tests (Dormann et al., 2013). Principal components analysis (PCA) has also been regularly applied for reducing the dimensionality of covariate spaces (De Marco and Nóbrega, 2018; Grenouillet et al., 2011; Raes et al., 2009).

Building on expert knowledge to select the covariate categories that are relevant from a species ecology perspective (e.g., bioclimatic, edaphic, hydrologic) is helpful for increasing the biological significance of the models (Mod et al., 2016; Petitpierre et al., 2017; Scherrer and Guisan, 2019). Expert input can also be useful to refine covariate sets or models obtained by using statistical methods for automated selection, but which seem unrealistic (Brandt et al., 2017). However, the expert approach alone can quickly become intractable as the number of candidate covariates and species to be modelled increases. Also, the expert approach may be inherently biased as selected covariates are generally the ones that are already well-known, reinforcing the use of "common covariates" in further studies and circularities in arguments. Automated covariate selection procedures are required for handling big data contexts, which are becoming the norm with the increasing amount of species occurrence data from citizen science initiatives (Amano et al., 2016; Dickinson et al., 2010; Pocock et al., 2017) and the availability of environmental layers for modelling them (Kuenzer et al., 2014; Soille et al., 2018; Sudmanns et al., 2020). Yet, there is currently no reference approach, or tool, to perform automated, data-driven, parsimonious, and fast covariate selection for species distribution modelling.

Here we introduce a two-step "embedded" covariate selection procedure aimed at optimizing the predictive ability and parsimony of SDMs fitted in a context of high-dimensional candidate covariate space. The procedure builds upon so-called "embedding methods", i.e., modelling algorithms equipped with their own built-in covariate selection procedures (Guyon and Elisseeff, 2003; Lal et al., 2006; Saeys et al., 2007). A key characteristic of these methods is that the covariate

Table 1
The three functions available in `covsel` (ver. 1.0.) with information on input data and arguments. See https://github.com/N-SDM/covsel and the function help files for additional details and examples.

Function	Description	Common inputs	Specific arguments
`covsel`. filteralgo()	Collinearity filtering (Step A)	*pa* numeric vector of species presences (1) and absences (0); *covdata* data frame with continuous covariate data; *weights* (optional) numeric vector with the weights for each value in *pa; force* (optional) character vector with the name(s) of the covariate(s) to be forced in the final set	*corcut* value of the correlation coefficient threshold used for identifying collinearity
`covsel`. embed()	Model-specific embedding (Step B)		*algorithms* character vector with the name(s) of the algorithms(s) used for the embedding procedure; *ncov* value for the target number of covariates to include in the final set; *maxncov* value for the maximum possible number of covariates to include in the final set; *nthreads* value for the number of cores to be used during parallel operations
`covsel`. filter()	Wrapper function applying the collinearity filtering algorithm at each target level(s) (e.g. i: variable level; ii: category level; iii: all remainders)		*corcut* value of the correlation coefficient threshold used for identifying collinearity; *categories* character vector with category-level covariate names; *variables* character vector with variable-level names

selection is done at the same time as model fitting, allowing to account early-on for both the specificities of the target modelling algorithm(s) and the multivariate context. Two major examples of embedded methods are LASSO (Tibshirani, 1996) and RIDGE (Hoerl and Kennard, 1970) regressions (see Guisan et al. (2002) in the context of SDMs). After providing a detailed description of our embedded covariate selection procedure, we assessed its performances relative to alternative simpler "filter" and "random" procedures through an example application aimed at modelling the habitat suitability of 50 species in Switzerland from a suite of 123 candidate covariates.

2. "Embedded" covariate selection procedure

Our embedded covariate selection procedure is developed around three main algorithms: Generalized Linear Model (GLM) (McCullagh and Nelder, 1989), Generalized Additive Model (GAM) (Hastie, 2017), and Random Forest (RF) (Breiman, 2001). These algorithms are among the most used in SDM studies (Hao et al., 2019) and are covering a gradient of flexibility and fitting methods that makes the ensemble of their results generalizable to many modelling frameworks. The procedure consists of two main steps: Step A "Collinearity filtering", and Step B "Model-specific embedding" (Fig. 1).

2.1. Step A: Collinearity filtering

In Step A, we reduce the dimensionality of the candidate covariate set by eliminating the less informative covariates among collinear pairs. This is done by iteratively reducing a correlation matrix in which the covariates are ordered based on univariate GLM *p*-values obtained using the *pa* vector of species presences/absences as response variable (see Fig. 1: Box 1 for details on the filtering algorithm). Collinear covariate pairs are identified using a user-specifiable Pearson correlation coefficient |r| threshold *corcut*, with a *corcut* default value of 0.70. From this step and onwards, it is possible to force specific covariates to be included in the final modelling step. For maximizing the diversity of selected covariate categories, the filtering step can be sequentially applied at three levels: (i) at the variable level (e.g., selecting the best covariate among those calculated for the same variable using multiple moving windows of 100-m, 500-m, or 1-km radii), (ii) the category level (e.g.: within thematic covariate categories), and (iii) using all remainders.

2.2. Step B: Model-specific embedding

In Step B, covariates selected after Step A are used to fit models with embedded selection procedures. We use GLM with elastic-net regularization (GLM-EN) (Zou and Hastie, 2005), GAM with null-space penalization (GAM-NP) (Marra and Wood, 2011), and guided regularized RF (RF-GR) (Deng and Runger, 2013). These algorithms are doing covariate selection at the same time as model fitting, allowing to account early-on for the specificities of the algorithms and the multivariate context. Furthermore, they have a more reasonable computational cost and limit overfitting compared to dredging or wrapping techniques, such as backward or forward selection strategies. Moreover, the three target algorithms are covering a gradient of fitting techniques (tree- and regression-based) and flexibility levels (GLM: parametric, GAM: semi-parametric, and RF: machine learning). A key benefit is that their covariate selection results are generalizable enough to be used as input for other popular SDM algorithms, such as Maxent (Phillips et al., 2006) or Gradient Boosting (Elith et al., 2008), even if they are not directly included in the initial procedure.

Details on the R packages and hyperparameter values used for fitting these three algorithms are provided in Supplementary material 1: Text S1. The three algorithms can be used all together (default), in combinations of two, or individually. For each algorithm, the *n* covariates retained after regularization are ranked from 1 ("best") to *n* ("worst"). The algorithm-specific ranking is done based on the maximum absolute values of the regularized regression coefficients for GLM, the chi-square statistic for GAM, and the Mean Decrease Gini index for RF. The final ranking of covariates is obtained by ordering the sum of the ranks for each covariate, starting with the covariates that were commonly selected by all algorithms, and then adding the remaining ones. The top *ncov* covariates are selected as the final modelling set, with *ncov* and $ncov_{max}$ being user-specifiable numbers. The default value for *ncov* is set to ceiling(log2(number of occurrences)), which, for example, results in 7 and 14 covariates for species with 100 and 14′000 occurrences, respectively. For species with less occurrences, rules such as *one predictor for ten occurrences* might be more parsimonious (Harrell et al., 1984). The default value of $ncov_{max}$ is set to 12 to limit the complexity of the models (Brun et al., 2020).

2.3. The `covsel` R package

The goal of the `covsel` R package is to implement and streamline the two steps of the embedded covariate selection procedure. It requires a standard installation of R (version≥4.0.0) and is openly available on the GitHub repository https://github.com/N-SDM/covsel, with all the information required for installing and running it. The README file of the repository (also provided as Supplementary material 2) navigates the user through the three functions currently available in `covsel` (ver. 1.0) (Table 1). It also includes an example application aimed at selecting the top 12 covariates, out of a panel of 75 candidates, for modelling the habitat suitability of the alpine marmot (*Marmota marmota*) in

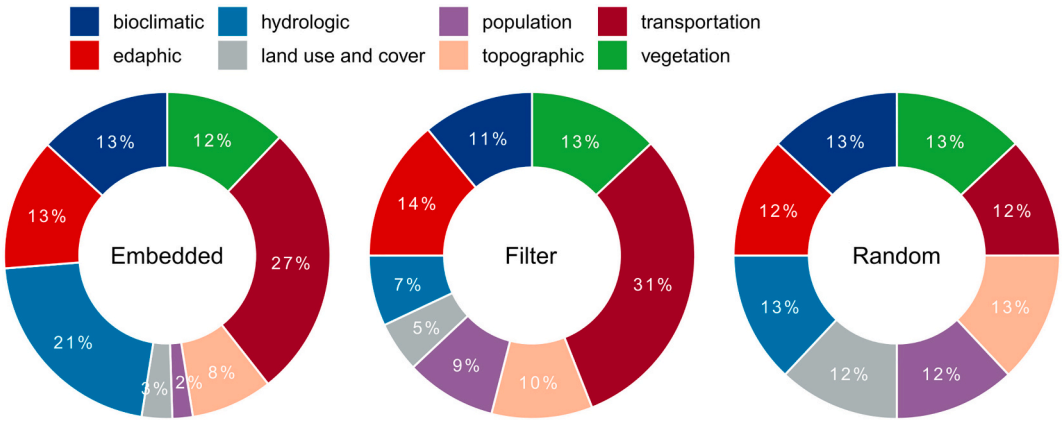


Fig. 2. Relative representation of the covariate categories selected for the 50 species under each of the three main covariate selection procedures ("embedded", "filter", and "random"). Each section represents the percentage of the ratio between the count of selected covariates from a given category and the overall number of covariates available in this category. See Supplementary material 2: Table S2 for details on covariate data.

Switzerland.

3. Performance demonstration

We assessed the performances of our embedded covariate selection procedure relative to alternative "filter" and "random" procedures through an example application aimed at modelling the habitat suitability of 50 species in Switzerland from a suite of 123 candidate covariates.

3.1. Species and covariate data

3.1.1. Species occurrence data

Occurrence records for 50 vertebrate and plant species in Switzerland were provided by the Swiss Species Information Center InfoSpecies (http://www.infospecies.ch) on August 23, 2021 (see Supplementary material 1: Table S1 for details on species data). Species were selected for maximizing the heterogeneity in terms of organismal groups, number of occurrence records, and spatial distribution in Switzerland. The number of occurrences available per species ranges from 130 to 13,462. For each species, 10,000 background absences were randomly generated across the study area to contrast the occurrence observations.

3.1.2. Covariate data

We used a suite of 123 candidate covariates from 8 environmental categories: 19 bioclimatic, 8 edaphic, 9 hydrologic, 67 land use and cover, 1 population, 12 topographic, 2 transportation, and 5 vegetation (see Supplementary material 2: Table S2 for details on covariate data). All covariates were extracted from a common 100-m resolution grid covering all of Switzerland and standardized to zero mean and unit variance.

3.2. Evaluated covariate selection procedures

For comparison with the embedded covariate selection procedure, we evaluated two alternative procedures: filter and random. For each procedure, we used `covsel` default values with *corcut* = 0.7, *ncov* = ceiling(log2(number of occurrences)), and $ncov_{max} = 12$.

The embedded procedure applied the full two-step approach described above. For the filter procedure, only Step A "collinearity filtering" was applied. For the random procedure, which was used as a null reference, *ncov* covariates were randomly sampled from the set of candidates. This random procedure was repeated 10 times for each species. For the embedded and filter procedures, the filtering algorithm (Step A) was first applied at the category level (i.e.: between covariates from each of the eight environmental categories), and then using all remainders. The three main procedures (embedded, filter, and random) were applied individually to each species.

In addition, we ran complementary procedures aimed at evaluating the "top-ranking" approach for selecting the set of covariates to be used in the final SDMs (see section 2.2.2 for details on the ranking approach). For the embedded procedure, this was done by comparing the results to those obtained using *ncov* randomly selected covariates among the set of non-regularized ones (after Step B). To obtain comparable outputs for the filter procedure, we replicated the random selection analysis on the subset of covariates selected after Step A. These two complementary random analyses were repeated 10 times each.

3.3. Model fitting and assessment

Selected covariate sets from the main (one embedded, one filter, and ten random) and complementary (ten random after Step B and ten random after Step A) procedures were used for fitting GLM, GAM and RF models. Details on the hyperparameters used for model fitting are provided in Supplementary material 1: Text S2. Model accuracy was evaluated using the Area Under the Curve' (AUC′) (or Somers' D, such as $AUC' = AUC * 2 - 1$) (Somers, 1962), the maximized True Skill Statistic (maxTSS) (Guisan et al., 2017), the Continuous Boyce Index (CBI) (Hirzel et al., 2006), and their average "Score" value obtained through a split-sample strategy repeated 100 times with 30% of the data kept aside for validation.

For each of the main procedure, we graphically summarized the relative representation of the eight environmental categories among selected covariates by displaying the percentages of the ratio between the count of selected covariates from a given category and the overall number of covariates available in this category. Differences in model accuracy between procedures were summarized using boxplots and Wilcoxon tests were used to assess their statistical significance. In addition, we computed the percentage of species for which a given procedure led to the top model. We assessed the sensitivity of our results to species data characteristics, by investigating model accuracy according to the number of occurrence records and their spatial coverage across Switzerland (Supplementary material 1: Table S1).

We compared the average computation time needed for (i) running the overall covariate selection procedure and model fitting steps under each of the three main procedures, and (ii) model fitting only. Analyses were run using a 10-core central processing unit strategy with AMD® EPYC 7402 on the University of Lausanne HPC cluster.

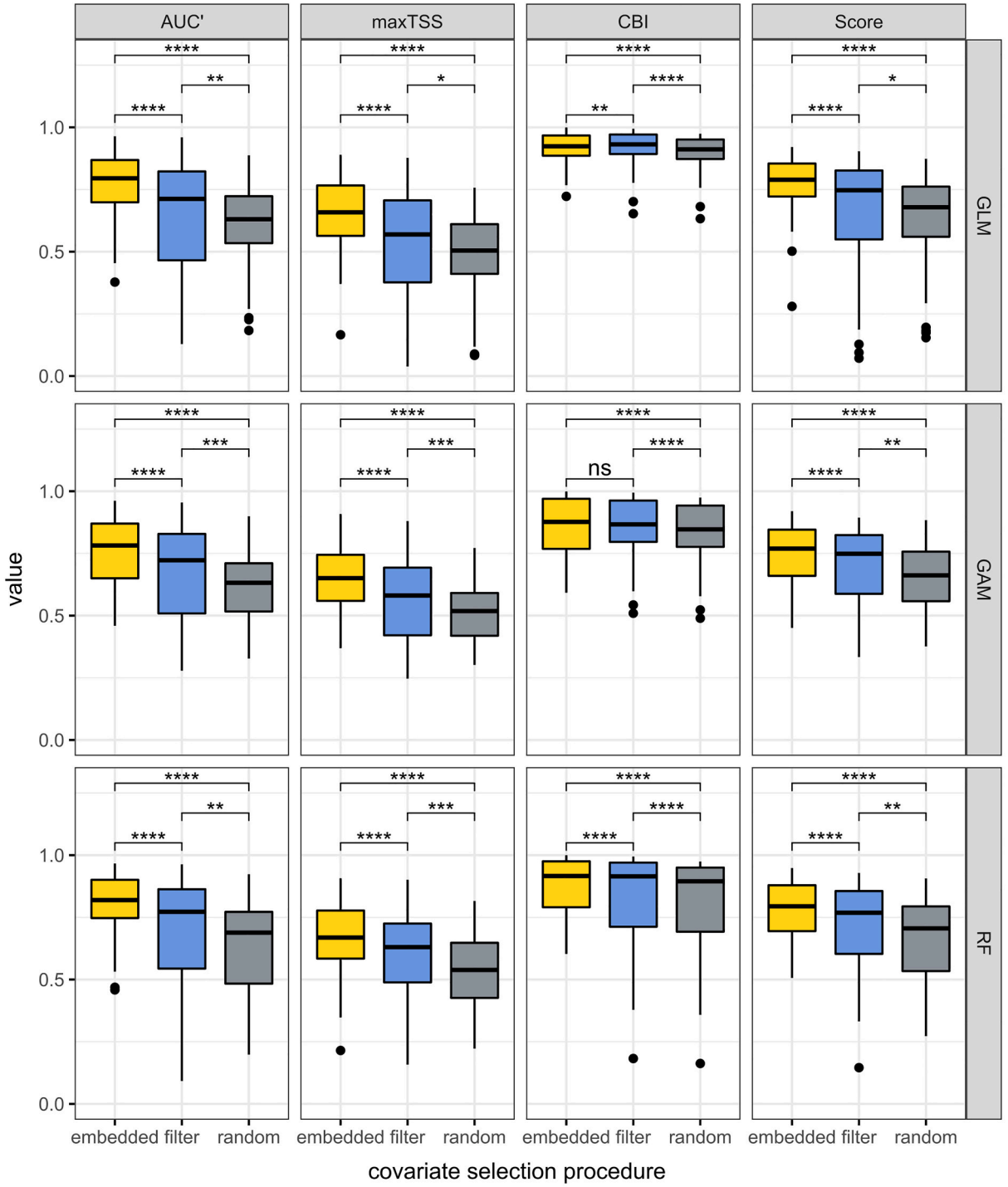


Fig. 3. Somers' D (AUC'), maximum True Skill Statistic (maxTSS), continuous Boyce index (CBI), and Score (average of AUC', maxTSS, and CBI) values of the models obtained for the 50 species under each of the three main covariate selection procedures ("embedded", "filter", and "random") and by modelling algorithm (GLM: Generalized Linear Model, GAM: Generalized Additive Model, and RF: Random Forest). For each boxplot, the central box represents the 1st quartile, the median, and the 3rd quartile. The two whiskers extend to the furthest non-outlier points (i.e., that are within 3/2 times the interquartile range of the 1st and 3rd quartiles). Wilcoxon tests were used to assess statistical significance in differences between methods with ****: $p < .0001$; ***: $p < .001$; **: $p < .01$; *: $p < .05$; ns: non-significant.

3.4. Results

3.4.1. Selected covariates

Covariate selection procedures and models were successfully run for all 50 species. Selected species-specific covariate sets included 8 to 12 covariates. Overall, covariate categories relative representation obtained for the embedded and filter procedures showed quite similar patterns (Fig. 2). For these two procedures, the "transportation" category was the most often selected relative to the overall number of covariates available in it, whereas "land use and cover" and "population" were among the least selected. The "hydrologic" category was the one with the largest difference in its relative representation between the embedded and filter procedures, with 21% and 7%, respectively. The random procedure selected covariate categories in a uniform way

Table 2
Percentage of species ($n = 50$) for which a given covariate selection procedure (Em: "embedded", Fi: "filter", and Rd: "random") led to the top Score value (average value of Somers' D, maximum True Skill Statistic, and continuous Boyce index) for each modelling algorithm (GLM: Generalized Linear Model, GAM: Generalized Additive Model, and RF: Random Forest). Results are shown for all species, species with the highest number of records (3rd tercile: "More records"), species with the lowest number of records (1st tercile: "Less records"), species with the widest spatial coverage in Switzerland (3rd tercile: "High coverage"), and species with the lowest spatial coverage (1st tercile: "Low coverage"). See Supplementary material 1: Table S1 for details on species data.

	All species			More records (3rd tercile)			Less records (1st tercile)			High coverage (3rd tercile)			Low coverage (1st tercile)		
Procedure	*Em*	*Fi*	*Rd*	*Em*	*Fi*	*Rd*	*Em*	*Fi*	*Rd*	*Em*	*Fi*	*Rd*	*Em*	*Fi*	*Rd*
GLM	76	22	2	65	35	0	82	18	0	75	25	0	88	12	0
GAM	82	18	0	83	17	0	76	24	0	81	19	0	94	6	0
RF	92	8	0	100	0	0	94	6	0	88	12	0	88	12	0

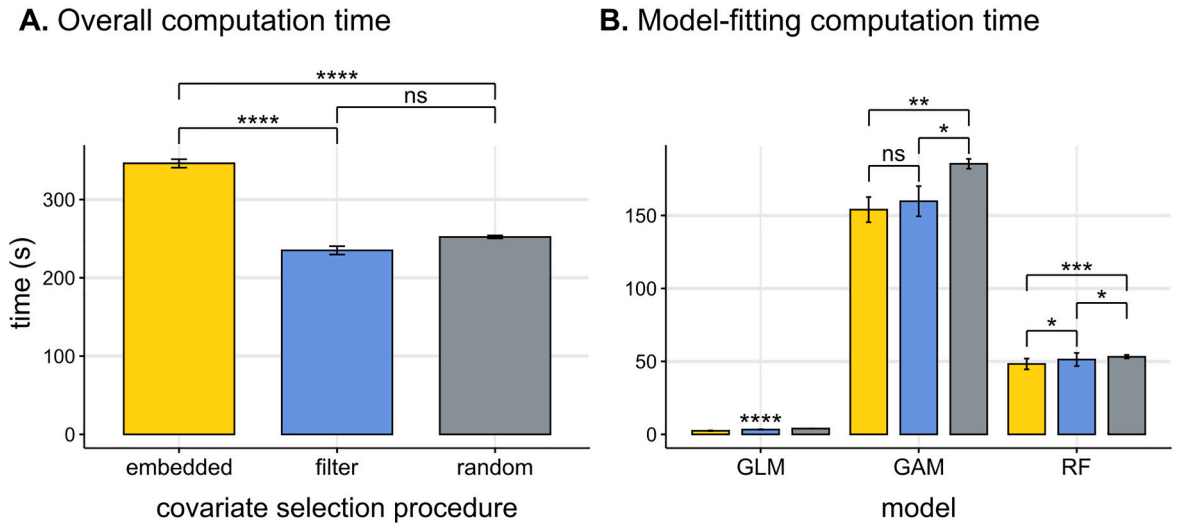


Fig. 4. Average computation time (seconds) for the 50 species needed for (A) running the overall covariate selection procedure and model fitting steps under each of the three main covariate selection procedures ("embedded", "filter", and "random"), and (B) model fitting only under each of the three main covariate selection procedures and by modelling algorithm (GLM: Generalized Linear Model, GAM: Generalized Additive Model, and RF: Random Forest). Wilcoxon tests were used to assess statistical significance in differences between methods with ****: $p < .0001$; ***: $p < .001$; **: $p < .01$; *: $p < .05$; ns: non-significant.

(Fig. 2).

3.4.2. Model accuracy

Models fitted with covariate sets obtained from the embedded procedure achieved the highest evaluation metrics (Fig. 3), with a mean ± standard deviation (SD) Score value for the 50 species and three algorithms of 0.77 ± 0.12. Models obtained with the random procedure had the lowest Score of 0.65 ± 0.18. The filter procedure had intermediate Score of 0.68 ± 0.21.

Results from the complementary analyses aimed at evaluating the top-ranking approach for selecting the covariates to be used in the final SDMs revealed that this method yielded an average increase in Score of 0.10 ± 0.08 and 0.12 ± 0.11 for the embedded and filter procedures (both significant at 0.001 level) by comparison to a random selection approach, respectively (see Supplementary material 1: Figs. S1 and S2 for detailed results on individual algorithms and metrics).

The embedded procedure led to highest Score values for >80% of the models (Table 2). The "embedded" > "filter" > "random" hierarchy was maintained across all modelling algorithms (Fig. 3 and Table 2). We showed that these results were consistent independently of the species data characteristics, with same findings obtained for species groups stratified by the number of occurrence records and spatial coverage across Switzerland (Supplementary material 1: Figs. S3 and S4). On average, the tendency of the embedded procedure to increase model accuracy was even more pronounced for RF, followed by GAM and GLM (Table 2).

3.4.3. Computation time

The overall average computation time for running both the covariate selection procedure and model fitting steps was ~1.5 time higher with the embedded procedure compared to the two alternatives (Fig. 4.A), which is the same order of magnitude as the gain in model accuracy. However, models from the embedded procedure were faster to fit (Fig. 4.B), presumably because the more relevant the covariate set, the faster the algorithms converged.

4. Conclusion

By combining a collinearity-filtering algorithm with model-specific embedded regularization techniques, we demonstrated the abilities of the two-step "embedded" covariate selection procedure to deliver accurate and parsimonious SDMs. Implemented and streamlined in the `covsel` R package, it offers an open and evolutive ready-to-use tool for automated covariate selection that was missing in the SDM field, with the potential to become the new standard by which to perform this step.

Capable of dealing with covariate sets ranging from several tens to thousands of candidates, the `covsel` R package can be easily run on any local computer or high-performance computing cluster. Despite being available for several decades (Hoerl and Kennard, 1970; Saeys et al., 2007; Tibshirani, 1996), embedding techniques have been little used in SDM studies. One of their main benefits compared to the more commonly applied filtering-only methods is their ability to interact directly with the target algorithms and to account for the multivariate context. Moreover, the combination of the three target algorithms (GLM, GAM and RF), that are covering a gradient of fitting techniques and

flexibility levels, makes our procedure particularly well-suited for ensemble SDM frameworks.

Measuring predictive accuracy with metrics such as AUC', maxTSS, or CBI was helpful to quantitatively compare and rank model performances. However, these metrics may capture only a partial picture of the quality of the model. Depending on the focus of the SDM study (e.g., conservation, climate change, biological invasions, ecological niche modelling, etc.), other model outputs, such as response curves and mapped predictions, should also be checked (Araujo et al., 2019; Zurell et al., 2012; Zurell et al., 2020). In addition, if the covariate selection is a key feature of the SDM process, all the other important steps with a potential influence on the predictions, including the complexity and tuning of model parameters, should also be carefully evaluated (Brun et al., 2020; Merow et al., 2014; Moreno-Amat et al., 2015).

All the complementary information required for the installation and use of `covsel`, along with sample data for an example application, are available on the `covsel` GitHub repository https://github.com/N-SDM/covsel. Anyone interested in contributing to its improvement is invited to suggest optimizations to existing pieces of code. To cite `covsel` or acknowledge its use, cite this article as follows, substituting the version of `covsel` that you used for "version 1.0": Adde et al. 2023. Too many candidates: embedded covariate selection procedure for species distribution modelling with the `covsel` R package (ver. 1.0).

Author contributions

Antoine Adde: Conceptualization (lead); Methodology (lead); Software (lead); Data curation (equal); Validation (lead); Writing – original draft (lead); Writing – review and editing (lead). **Pierre-Louis Rey**: Methodology (supporting); Data curation (lead); Validation (supporting); Writing – original draft (supporting); Writing – review and editing (equal). **Fabian Fopp**: Conceptualization (supporting); Methodology (supporting); Validation (supporting); Writing – review and editing (equal). **Blaise Petitpierre**: Conceptualization (supporting); Methodology (supporting); Writing – review and editing (equal). **Anna K. Schweiger:** Conceptualization (supporting); Writing – review and editing (equal). **Olivier Broennimann**: Conceptualization (supporting); Methodology (supporting); Writing – review and editing (equal). **Anthony Lehmann**: Conceptualization (supporting); Writing – review and editing (equal). **Niklaus E. Zimmermman**: Conceptualization (supporting); Writing – review and editing (equal). **Florian Altermatt**: Conceptualization (supporting); Writing – review and editing (equal). **Loïc Pellissier**: Conceptualization (supporting); Methodology (supporting); Validation (supporting); Writing – review and editing (equal). **Antoine Guisan**: Supervision (lead); Funding acquisition (lead); Conceptualization (supporting); Methodology (supporting); Validation (supporting); Writing – original draft (supporting); Writing – review and editing (equal).

Declaration of Competing Interest

The authors declare that they have no conflict of interest.

Data availability

All the data and information required for the installation and use of `covsel` are available on the `covsel` GitHub repository https://github.com/N-SDM/covsel.

Acknowledgments

We gratefully acknowledge financial support through the Action Plan of the Swiss Biodiversity Strategy by the Federal Office for the Environment (FOEN) for financing the ValPar.CH and SwissCatchment projects. The Swiss Species Information Center InfoSpecies (http://www.infospecies.ch) supplied Swiss-level species occurrence data and expertise on species' ecology, and we acknowledge their support regarding the database. This research was enabled in part by the support provided by the Scientific Computing and Research Unit of Lausanne University (https://www.unil.ch/ci/dcsr).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ecoinf.2023.102080.

References

Amano, T., Lamming, J.D.L., Sutherland, W.J., 2016. Spatial gaps in global biodiversity information and the role of citizen science. Bioscience 66, 393–400.
Araujo, M.B., Anderson, R.P., Barbosa, A.M., Beale, C.M., Dormann, C.F., Early, R., Garcia, R.A., Guisan, A., Maiorano, L., Naimi, B., O'Hara, R.B., Zimmermann, N.E., Rahbek, C., 2019. Standards for distribution models in biodiversity assessments. Sci. Adv. 5.
Austin, M.P., Van Niel, K.P., 2011. Improving species distribution models for climate change studies: variable selection and scale. J. Biogeogr. 38, 1–8.
Brandt, L.A., Benscoter, A.M., Harvey, R., Speroterra, C., Bucklin, D., Romañach, S.S., Watling, J.I., Mazzotti, F.J., 2017. Comparison of climate envelope models developed using expert-selected variables versus statistical selection. Ecol. Model. 345, 10–20.
Brauner, N., Shacham, M., 1998. Role of range and precision of the independent variable in regression of data. AICHE J. 44, 603–611.
Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32.
Brun, P., Thuiller, W., Chauvier, Y., Pellissier, L., Wuest, R.O., Wang, Z.H., Zimmermann, N.E., 2020. Model complexity affects species distribution projections under climate change. J. Biogeogr. 47, 130–142.
De Marco, P., Nóbrega, C.C., 2018. Evaluating collinearity effects on species distribution models: an approach based on virtual species simulation. PLoS One 13, e0202403.
Deng, H., Runger, G., 2013. Gene selection with guided regularized random forest. Pattern Recogn. 46, 3483–3489.
Dickinson, J.L., Zuckerberg, B., Bonter, D.N., 2010. Citizen science as an ecological research tool: challenges and benefits. Annu. Rev. Ecol. Evol. Syst. 41 (41), 149–172.
Dormann, C.F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carre, G., Marquez, J.R.G., Gruber, B., Lafourcade, B., Leitao, P.J., Munkemuller, T., McClean, C., Osborne, P.E., Reineking, B., Schroder, B., Skidmore, A.K., Zurell, D., Lautenbach, S., 2013. Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. Ecography 36, 27–46.
Elith, J., Leathwick, J.R., Hastie, T., 2008. A working guide to boosted regression trees. J. Anim. Ecol. 77, 802–813.
Ferrier, S., Ninan, K.N., Leadley, P., Alkemade, R., Acosta, L.A., Akçakaya, H.R., Brotons, L., Cheung, W.W.L., Christensen, V., Harhash, K.A., Kabubo-Mariara, J., Lundquist, C., Obersteiner, M., Pereira, H.M., Peterson, G., Pichs-Madruga, R., Ravindranath, N., Rondinini, C., Wintle, B.A., 2016. IPBES : The methodological assessment report on scenarios and models of biodiversity and ecosystem services. In: Secretariat of the Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services (IPBES), Bonn, DE.
Fois, M., Cuena-Lombrana, A., Fenu, G., Bacchetta, G., 2018. Using species distribution models at local scale to guide the search of poorly known species: review, methodological issues and future directions. Ecol. Model. 385, 124–132.
Fourcade, Y., Besnard, A.G., Secondi, J., 2018. Paintings predict the distribution of species, or the challenge of selecting environmental predictors and evaluation statistics. Glob. Ecol. Biogeogr. 27, 245–256.
Franklin, J., 2010. Mapping Species Distributions: Spatial Inference and Prediction. Cambridge University Press.
Grenouillet, G., Buisson, L., Casajus, N., Lek, S., 2011. Ensemble modelling of species distribution: the effects of geographical and environmental ranges. Ecography 34, 9–17.
Guisan, A., Edwards, T.C., Hastie, T., 2002. Generalized linear and generalized additive models in studies of species distributions: setting the scene. Ecol. Model. 157, 89–100.
Guisan, A., Tingley, R., Baumgartner, J.B., Naujokaitis-Lewis, I., Sutcliffe, P.R., Tulloch, A.I.T., Regan, T.J., Brotons, L., McDonald-Madden, E., Mantyka-Pringle, C., Martin, T.G., Rhodes, J.R., Maggini, R., Setterfield, S.A., Elith, J., Schwartz, M.W., Wintle, B.A., Broennimann, O., Austin, M., Ferrier, S., Kearney, M.R., Possingham, H. P., Buckley, Y.M., 2013. Predicting species distributions for conservation decisions. Ecol. Lett. 16, 1424–1435.
Guisan, A., Thuiller, W., Zimmermann, N.E., 2017. Habitat Suitability and Distribution Models, with Applications in R. Cambridge University Press, Cambridge.
Guyon, I., Elisseeff, A., 2003. An introduction to variable and feature selection. J. Mach. Learn. Res. 3, 1157–1182.
Hao, T.X., Elith, J., Guillera-Arroita, G., Lahoz-Monfort, J.J., 2019. A review of evidence about use and performance of species distribution modelling ensembles like BIOMOD. Divers. Distrib. 25, 839–852.
Harrell, F.E., Lee, K.L., Califf, R.M., Pryor, D.B., Rosati, R.A., 1984. Regression modelling strategies for improved prognostic prediction. Stat. Med. 3, 143–152.
Hastie, T.J., 2017. Generalized additive models. In: *Statistical Models in S*pp. Routledge, pp. 249–307.

Hirzel, A.H., Le Lay, G., Helfer, V., Randin, C., Guisan, A., 2006. Evaluating the ability of habitat suitability models to predict species presences. Ecol. Model. 199, 142–152.
Hoerl, A.E., Kennard, R.W., 1970. Ridge regression: biased estimation for nonorthogonal problems. Technometrics 12, 55–67.
Kuenzer, C., Ottinger, M., Wegmann, M., Guo, H., Wang, C., Zhang, J., Dech, S., Wikelski, M., 2014. Earth observation satellite sensors for biodiversity monitoring: potentials and bottlenecks. Int. J. Remote Sens. 35, 6599–6647.
Lal, T.N., Chapelle, O., Weston, J., Elisseeff, A., 2006. Embedded methods. In: Feature Extraction. Springer, pp. 137–165.
Marra, G., Wood, S.N., 2011. Practical variable selection for generalized additive models. Computat. Stat. Data Analys. 55, 2372–2387.
McCullagh, P., Nelder, J.A., 1989. Generalized Linear Models, 2nd edition. Chapman and Hall, London.
Melo-Merino, S.M., Reyes-Bonilla, H., Lira-Noriega, A., 2020. Ecological niche models and species distribution models in marine environments: a literature review and spatial analysis of evidence. Ecol. Model. 415, 108837.
Merow, C., Smith, M.J., Edwards, T.C., Guisan, A., McMahon, S.M., Normand, S., Thuiller, W., Wuest, R.O., Zimmermann, N.E., Elith, J., 2014. What do we gain from simplicity versus complexity in species distribution models? Ecography 37, 1267–1281.
Merow, C., Maitner, B.S., Owens, H.L., Kass, J.M., Enquist, B.J., Jetz, W., Guralnick, R., 2019. Species' range model metadata standards: RMMS. Glob. Ecol. Biogeogr. 28, 1912–1924.
Mod, H.K., Scherrer, D., Luoto, M., Guisan, A., 2016. What we use is not what we know: environmental predictors in plant distribution models. J. Veg. Sci. 27, 1308–1322.
Moreno-Amat, E., Mateo, R.G., Nieto-Lugilde, D., Morueta-Holme, N., Svenning, J.-C., Garcia-Amorena, I., 2015. Impact of model complexity on cross-temporal transferability in Maxent species distribution models: an assessment using paleobotanical data. Ecol. Model. 312, 308–317.
Peterson, A.T., Soberón, J., Pearson, R.G., Anderson, R., Martínez-Meyer, E., Nakamura, M., Araújo, M.P., 2011. Ecological Niches and Geographic Distributions. Princeton University Press, Princeton.
Petitpierre, B., Broennimann, O., Kueffer, C., Daehler, C., Guisan, A., 2017. Selecting predictors to maximize the transferability of species distribution models: lessons from cross-continental plant invasions. Glob. Ecol. Biogeogr. 26, 275–287.
Phillips, S.J., Anderson, R.P., Schapire, R.E., 2006. Maximum entropy modeling of species geographic distributions. Ecol. Model. 190, 231–259.
Pocock, M.J., Tweddle, J.C., Savage, J., Robinson, L.D., Roy, H.E., 2017. The diversity and evolution of ecological and environmental citizen science. PLoS One 12.
Raes, N., Roos, M.C., Slik, J.W.F., van Loon, E.E., ter Steege, H., 2009. Botanical richness and endemicity patterns of Borneo derived from species distribution models. Ecography 32, 180–192.
Saeys, Y., Inza, I., Larranaga, P., 2007. A review of feature selection techniques in bioinformatics. Bioinformatics 23, 2507–2517.
Scherrer, D., Guisan, A., 2019. Ecological indicator values reveal missing predictors of species distributions. Sci. Rep. 9.
Soille, P., Burger, A., De Marchi, D., Kempeneers, P., Rodriguez, D., Syrris, V., Vasilev, V., 2018. A versatile data-intensive computing platform for information retrieval from big geospatial data. Futur. Gener. Comput. Syst. 81, 30–40.
Somers, R.H., 1962. A new asymmetric measure of association for ordinal variables. Am. Sociol. Rev. 799–811.
Sudmanns, M., Tiede, D., Lang, S., Bergstedt, H., Trost, G., Augustin, H., Baraldi, A., Blaschke, T., 2020. Big earth data: disruptive changes in earth observation data management and analysis? Int. J. Dig. Earth 13, 832–850.
Tibshirani, R., 1996. Regression shrinkage and selection via the LASSO. J. Royal Stat. Soc. Ser. B-Methodol. 58, 267–288.
Yates, K.L., Bouchet, P.J., Caley, M.J., Mengersen, K., Randin, C.F., Parnell, S., Fielding, A.H., Bamford, A.J., Ban, S., Barbosa, A.M., 2018. Outstanding challenges in the transferability of ecological models. Trends Ecol. Evol. 33, 790–802.
Zou, H., Hastie, T., 2005. Regularization and variable selection via the elastic net. J. Royal Stat. Soc. Series B Stat. Methodol. 67, 301–320.
Zurell, D., Elith, J., Schroder, B., 2012. Predicting to new environments: tools for visualizing model behaviour and impacts on mapped distributions. Divers. Distrib. 18, 628–634.
Zurell, D., Franklin, J., Konig, C., Bouchet, P.J., Dormann, C.F., Elith, J., Fandos, G., Feng, X., Guillera-Arroita, G., Guisan, A., Lahoz-Monfort, J.J., Leitao, P.J., Park, D.S., Peterson, A.T., Rapacciuolo, G., Schmatz, D.R., Schroder, B., Serra-Diaz, J.M., Thuiller, W., Yates, K.L., Zimmermann, N.E., Merow, C., 2020. A standard protocol for reporting species distribution models. Ecography 43, 1–17.