

# Common Species Distribution Models in Biodiversity Analysis and Their Challenges and Prospects in Application

Le ZOU<sup>1</sup>, Renyan DUAN<sup>1</sup>, Chenzhong JIN<sup>1,2</sup>, Xiansheng TAN<sup>1,2\*</sup>

1. College of Agriculture and Biology Technology, Hunan University of Humanities Science and Technology, Loudi 417000, China; 2. Collaborative Innovation Center of Farmland Weed Control Technology and Application, Loudi 417000, China

**Abstract** Species distribution models have been widely used to explore suitable habitats of species, the impact of climate change on the distribution of suitable habitats of species, and the construction of ecological reserves. This paper introduced species distribution models commonly used in biodiversity analysis, as well as model performance evaluation indexes, challenges in the application of species distribution models, and finally prospected the development trend of research on species distribution models.

**Key words** Species distribution models; Evaluation index; Challenge; Development trends

**DOI:**10.19759/j.cnki.2164-4993.2023.04.025

Species distribution models (SDMs) are the association models between species distribution and environmental variables by associating the distribution sample information of target organisms with certain environmental control variables or relevant information technology. Association models can be applied to research on a certain region to evaluate the distribution of specific species. The models are built on the basis of ecological niche, which mainly refers to the location data of species in the ecosystem and the relationships between species<sup>[1]</sup>. Hutchinson mathematically described ecological niche as a kind of "super volume"<sup>[1]</sup>, which can maintain a constant population distribution in the space defined by various types of environmental variables. With regard to the concept of "super volume", ecologists have made many attempts to obtain multiple types of species distribution models based on different concepts and methods. Today, species distribution models have evolved into one of the indispensable tools of basic ecology and biogeography. They are mainly used to explore the relationship between species distribution and climate in the environment of intensified global climate change<sup>[2]</sup>, the level of climate change and the level of plant community composition and distribution changed by climate in local areas<sup>[3]</sup>, the monitoring and early warning of ecosystem management functional groups and core species<sup>[4]</sup>, the management and maintenance of ecosystems at different scales and their impact on diversity<sup>[5]</sup>, the prediction of the invasion area by introduced species, and the distribution prediction and protection planning of potential key organisms in restoration of ecosystem construction and system management<sup>[5]</sup>. In the general use of generalized population distribution models,

consideration should be given to evaluating the performance of models. Therefore, the strategy for evaluating model performance also needs to be carefully considered during the modeling process. This paper expounded the classification and practical functions of species distribution models, summarized the discovery and development process of species distribution models and the difficulties that may be faced by the current application of species classification models, and finally put forward prospects for the future of species distribution models.

## Common Species Distribution Models

The establishment of species distribution models began with the development and application of biological climate models<sup>[4]</sup>. Subsequently, a variety of models were gradually developed, including the ecological niche model based on computation (BIOCLIM)<sup>[5]</sup>, DOMAIN model based on the Gower distance algorithm<sup>[6]</sup>, the MD model, the ecological niche factor model (ENFA model)<sup>[7]</sup>, the MaxEnt model<sup>[8]</sup>, the GLM model<sup>[7]</sup>, the generalized additive model (GAM)<sup>[9]</sup>, the classification tree model (GTA), discriminant analysis (FDA) (or linear discriminant analysis LDA)<sup>[9]</sup>, the booster method based on the boosting algorithm (GBM) multivariate autoregressive spline model (MARS), random forest (RF), and rule-based artificial neural network (ANN), among which BIOCLIM, DOMAIN model, MaxEnt model, and artificial neural network (ANN) are the most widely used.

## Evaluation Indexes of Species Distribution Models

The commonly used evaluation indexes include Kappa coefficient, TSS value, overall accuracy, predictive specificity, sensitivity, and AUC, which are calculated in different ways, thus playing different roles. When comparing multiple types of models, we should choose different models as references based on the practical meanings they represent. The Kappa coefficient and AUC value

Received: April 17, 2023 Accepted: June 28, 2023

Supported by Natural Science Foundation of Hunan Province (2021JJ30375); Natural Science Foundation of Hunan Provincial Department of Education (20A275); Science and Technology Innovation Team Project of Hunan Province (201937924).

Le ZOU (1998 -), female, P. R. China, master, devoted to research about biodiversity.

\* Corresponding author.

are the two most commonly used evaluation indexes.

The Kappa coefficient is an indicator used for consistency testing and can also be used to measure the effectiveness of classification. For classification problems, the so-called inhibition refers to whether the predicted results of the model are consistent with the actual distribution results. The calculation of Kappa coefficient is based on the confusion matrix, and the value is between  $-1$  and  $1$ , usually greater than  $0$ .

The AUC value is the area under the subject's operating characteristic curve (ROC), that is, the area enclosed with the coordinate axis under the ROC. An AUC value closer to  $1$  stands for better performance of the model.

## Challenges in Application of Species Distribution Models

Species distribution models are not only suitable for basic ecological research, but also suitable for environmental application research. They are also used to study different characteristics of the same organism, such as functional characteristics, secondary groups, or genes. On this basis, other professional tools and data can also be integrated to analyze and study biogeography, main system development, population genetics, population development dynamics, *etc.* However, practical research still faces following challenges.

**Sample selection** Firstly, the selection of distribution samples may lead to bias in the prediction results. Due to the interference of natural or human factors on the spatial distribution of species during the sampling process, the deviation in sample collection has been caused. In other words, the ecological niche represented by a sample set is only a part of the actual ecological niche. The selection and processing without sample sets for the purpose of unbiased selection can lead to the condition that the model algorithm overuses samples from certain regions to fit the relationship between species and environmental variables in the calculation process, resulting in overfitting problems. Meanwhile, due to the small number of samples outside the part, it leads to underfitting.

**Selection of species distribution models** Scientists have developed many models to estimate relationships between species and related environmental variables. In the modeling process, scholars need to select different modeling algorithms according to different modeling purposes, species' ecological niche characteristics and modeling data bases. Generally speaking, the more complex the model algorithm, the higher the statistical accuracy, but in some cases, simple models also have irreplaceable advantages. For example, the generalized linear model is dominant, and the algorithm can provide the coefficients of explanatory variables, according to which scholars can intuitively determine the role and importance of environmental variables on species distribution. In contrast, classification and regression tree models are difficult to explain in ecology due to their complexity. Complex machine learning algorithms, such as random forest and artificial neural network models, are easy to use and relatively accurate, but the modeling

process has almost no ecological significance.

**Effects of environmental variables** The selection of environmental variables also affects the prediction results of models. At present, many studies use the Bioclim data set, the 19 bioclimate variables in which include extreme values such as temperature, precipitation and variability. In addition, aspect, slope, altitude, vegetation index, vegetation coverage and land use efficiency are also used as environmental variables to estimate the potential distribution of terrestrial ecosystem species. However, due to the large amount of information introduced by these variables in the model prediction process, the selection of environmental variables should be considered first when modeling the potential distribution of species. Therefore, in the subsequent model development process, the key issue in the development of species distribution models is to add a variable selection module to eliminate the impact of redundant information.

**Extrapolation** Extrapolation<sup>[10]</sup> refers to the extrapolation of a model in space, time, and scale. The spatial extrapolation limit method is also widely used in evaluating the risk of invasion and biological invasion. It is worth noting that in the invasion phase, organisms may produce appropriate ecological position drift in the target area. Representative studies on time extrapolation include assessing the possible distribution of a specific species within the same region based on climate characteristics. In this type of study, it should be noted that in addition to determining meteorological factors (temperature and precipitation), biological diffusion levels and various external factors (soil and vegetation) need to be included, especially in studying plant populations with significant biological action ability in certain typical areas. Species distribution models are often unable to be transformed at the spatial scale, which is because that in general, species ecological niche shows different properties in different spatial scales, and environmental thresholds are also different. Under such condition, research should be supported by other data, especially downscaling information of species distribution models.

**Modeling strategies** A reasonable model strategy includes not only selecting appropriate model statistical algorithms, but also comprehensive strategies calculated through models. In a certain research context, a simple model framework cannot complete modeling because the model requires more complex structures, such as certain special habitat species (such as fungi, mushrooms, and desert vegetation) that are more dependent on certain non-climatic environmental factors. This non climatic impact can lead to differences in the local distribution patterns of such organisms, as well as affecting their diffusion ability and even their ability to adapt to the climate environment. In order to solve this difficulty, some scholars have proposed a method to establish comprehensive species distribution models. First, the a species distribution trend model is established based on climate factors and environmental variables, and then, species distribution models are established based on limiting environmental factors. Next, the species distribution models are comprehensively analyzed by comprehensively

analyzing the two factors, and the conclusion serves as the preliminary modeling result. Studies have shown that this method is more conducive to analyzing the distribution status of animals in special habitats and the correlation with climate factors.

## Development Trend of Species Distribution Models

Nowadays, species distribution models are gradually applied in ecology, environmental science, biogeography and other multidisciplinary directions. Meanwhile, the progress of basic theories in various fields has also accelerated the development of species distribution models, and its specific development direction can be classified into following points.

### Integrating historical climate information on current distribution pattern of species

The precondition of species distribution models is the ecological needs of organisms when they reach a stable state, but the negative effects of historical reasons on the distribution areas and distribution structures of organisms are usually not fully considered when building models. Such equilibrium hypothesis is an ideal state, but it may also affect and lead to errors in the results of species distribution models. In 2014, Patsiou *et al.*<sup>[11]</sup> simulated the potential distribution area of the rare plant *Saxifraga* using the aggregated species distribution model, high-resolution paleoclimatic characteristics and geomorphic data sets. They proved that the geomorphic and paleoclimatic characteristics have a profound impact on the current and future distribution areas of this plant. In 2015, Svenning *et al.*<sup>[12]</sup> elaborated on the important impact of climate historical factors on current spatial distribution of biological richness, and gave full evidence of the important impact of climate change during the freezing period of the Quaternary on the current biological distribution pattern, as well as the important impact of historical meteorological factors on the development and functional diversity of the current biological system. Due to the increasing difficulty of collecting data on modern weather environments, it has become possible to accurately model the distribution patterns of organisms in ancient climate environments, which will improve scholars' understanding on the geographical pattern that constitute the current distribution patterns of organisms. Biodiversity is related to the functional diversity of ecosystems. It is necessary to gradually strengthen research on the impact of climate change on national biological distribution and local biodiversity.

### How to effectively simulate biological interaction within the framework of species distribution models

At present, most of species distribution models do not include biological interaction factors in the initial stage of construction, and only comprise limited biological interaction data. Therefore, only abiotic environmental information is usually used in model construction. However, the introduction of biological action information into models will help to increase the efficiency of constructing species distribution patterns and enable taking environmental factors into account. The simulation can be based on the

following strategies; the numbers of other species present in the grid<sup>[12]</sup>, target species proportion data<sup>[13]</sup>, competition coefficient, Eltonian noise hypothesis<sup>[9]</sup>, *etc.* Conceptually, the response of species to climate gradient changes also reflects Hutchinson's multidimensional super volume ecological niche theory. Since species distribution patterns are jointly affected by abiotic predictors and biological interaction variables at various scales, the impact of biological interaction varies based on the changes in regions and scales. In short, in the future, biological factors should be considered in building models, while also paying attention to the impact of scale effects.

### Using big data of earth science as model input

Based on technological innovation, big data analysis technology has become one of the main forces in multidisciplinary development<sup>[14]</sup>. Although this technology has been able to provide more detailed and accurate environmental and species distribution data, how to more efficiently integrate massive geographical data of various resolutions and stages is still the main challenge in the modeling process of species distribution models. At present, scholars have begun testing remote sensing information as model input to provide more detailed modeling information. For example, remote sensing information is combined with biological sampling point information analysis, and remote sensing data are used as environmental variable indexes. At present, through the analysis of a large quantity of remote sensing data, people have mastered the products of remote sensing technology with ecological functions, such as evapotranspiration, vegetation index, leaf area coefficient, normalized vegetation index and other vegetation coverage indexes, which can significantly increase the accuracy of species distribution models. Vega *et al.*<sup>[15]</sup> collected further research results on the synthesis of 28 types of satellite data, and applied algorithms and interpolation rules consistent with meteorological stations to obtain 19 bioclimatic variables derived from satellite data. The above information has significantly expanded its coverage compared with WorldClim's information, such as Antarctica, while significantly improving the accuracy of BIO variables in certain areas where meteorological stations have not been built.

## Conclusions

In recent decades, with the development of geography, statistics, computer science and other related disciplines and science and technology, species distribution models have made great progress. However, during the use of models, scholars often do not have a clear understanding of the research on ecological issues, lack proficiency in computing or building models, and often blindly imitate the work of predecessors, drawing too much from past data, which makes it difficult to have a subjective understanding of the modeling process and related algorithms. In this way, there may be significant personal equation, and it is often difficult to apply models. The research expounds the construction of species distribution models, explains the relevant theories supported by previous models, analyzes the methodological advantages and

application value of various model algorithms, and puts forward the difficulties and future prospects that may be encountered in the establishment and application of models. In addition, in the modeling practice of species distribution models, the rationality and accuracy of model results are closely related to the correct selection of prediction factors and modeling methods, the interaction between time and space scales and between environmental and geographical factors, and the impact of model extrapolation. Therefore, scholars need to understand the supporting framework of models as much as possible, clarify the way to build models, and correctly select model algorithms. Future research should attach great importance to important historical and geographical factors that lead to current species distribution patterns, and possible role of species ecological niche evolution. In the modeling process, the genealogical biogeography and landscape genetics theories are combined, and the natural processes such as biological interaction are reasonably introduced into the model framework. Meanwhile, a large amount of geographic data (such as remote sensing data) has been effectively utilized, greatly improving the accuracy of models. Therefore, with the development of various models and the improvement of their accuracy, the application of species distribution models in the study of various biological habitats and the prediction of changes in biological habitats caused by climate change will be further promoted, providing a reference for the establishment of biological reserves.

## References

- [1] JIANG X, NI J. Species-climate relationships of 10 desert plant species and their estimated potential distribution range in the arid lands of northwestern China[J]. Chinese Journal of Plant Ecology, 2005, 29(1): 98–107. (in Chinese).
- [2] ZHAI TQ, LI XH. Climate change induced potential range shift of the crested ibis based on ensemble models [J]. Acta Ecologica Sinica, 2012, 32(8): 2361–2370. (in Chinese).
- [3] ZHANG L, LIU SR, SUN PS, *et al.* Comparative evaluation of multiple models of the effects of climate change on the potential distribution of Pi-

- nus massoniana[J]. Chinese Journal of Plant Ecology, 2011, 35(11): 1091–1105. (in Chinese).
- [4] YEE THOMAS, MITCHELL ND. Generalized additive models in plant ecology[J]. Journal of Vegetation Science, 1991(2): 587–602.
- [5] HIRZEL AH, HAUSSER J, CHESELL D, *et al.* Ecological niche factor analysis: How to compute habitat-suitability maps without absence data [J]. Ecology, 2002(83): 2027–2036.
- [6] ZHU Y, KANG MY. Application of ordination and GLM/GAM in the research of the relationship between plant species and environment [J]. Chinese Journal of Ecology, 2005, 24(7): 807–811. (in Chinese).
- [7] HOLZMANN I, AGOSTINI I, DEMATTEO K, *et al.* Using species distribution modeling to assess factors that determine the distribution of two parapatric howlers (*Alouatta* spp.) in South America[J]. International Journal of Primatology, 2014, 36(1): 1–15.
- [8] WALKER PA, COCKS KD. HABITAT: A procedure for modeling a disjoint environmental envelop for a plant or animal species[J]. Global Ecology and Biogeography, 1991(1): 108–118.
- [9] HEIKKINEN RK, LUOTO M, VIRKKALA R, *et al.* Biotic interactions improve prediction of boreal bird distributions at macroscales[J]. Global Ecology and Biogeography, 2010, 16(6):754–763.
- [10] LIU C, NEWELL G, WHITE M. On the selection of thresholds for predicting species occurrence with presence-only data[J]. Ecology and Evolution, 2016, 6(1): 337–348.
- [11] PATSIU TS, CONTI E, ZIMMERMANN NE, *et al.* Topo-climatic microrefugia explain the persistence of a rare endemic plant in the Alps during the last 21 millennia[J]. Global Change Biology, 2014, 20(7): 2286–2300.
- [12] SVENNING JC, EISERHARDT WL, NORMAND S, *et al.* The influence of paleoclimate on present-day patterns in biodiversity and ecosystems[J]. Annual Review of Ecology Evolution and Systematics, 2015, 46(1): 551–572.
- [13] FORDHAM DA, SALTRÉ F, BROWN SC, *et al.* Why decadal to century timescale paleoclimate data is needed to explain presentday patterns of biological diversity and change[J]. Global Change Biology, 2018(24): 1371–1381.
- [14] MEIER ES, EDWARDS TC, KIENAST F, *et al.* Co-occurrence patterns of trees along macro-climatic gradients and their potential influence on the present and future distribution of *Fagus sylvatica* L.[J]. Journal of Biogeography, 2011, 38(2): 371–382.
- [15] VEGA GC, PERTIERRA LR, OLALLA-TÁRRAGA MÁ. MERRAclim, a high-resolution global dataset of remotely sensed bioclimatic variables for ecological modelling[J]. Scientific Data, 2017(4): 170078.

Editor: Yingzhi GUANG

Proofreader: Xinxu ZHU

(Continued from page 114)

- [24] WANG H, ZHOU Y. The method of ABTS assay for screening and evaluating antioxidant [J]. Guangzhou Chemical Industry, 2012, 40(22): 41–43.
- [25] ZHANG H, ZHANG G, CHEN J. Study on the sunproof potential for thirty kinds of chemical constituents from Chinese herb medicine[J]. China Pharmacist, 2010, 13(7): 935–938. (in Chinese).
- [26] ZHANG XL, LI SY, SUN JN, *et al.* Optimization of extraction process and antioxidant activity of total flavonoids from passion fruit peel by response surface methodology[J/OL]. Feed Research, 2022(16): 83–88. (in Chinese).
- [27] HU XH, LONG JF, YANG SZ, *et al.* Study on total flavonoids content and antioxidant activity of *Allium tuberosum* in three regions of south-

- east Guizhou[J]. Journal of Kaili University, 2022, 40(3): 67–74. (in Chinese).
- [28] LI JH, WENG GY, ZHU M, *et al.* Microwave-assisted extraction and antioxidant activities of total flavonoids from kiwi fruit residues [J]. Food Research and Development, 2022, 43(12): 79–85. (in Chinese).
- [29] ZAHRA ABOTORABIL, MOHSEN KHORASHADIZADEH, MINA ARAB, *et al.* Jujube and green tea extracts protect human fibroblast cells against UVB-mediated photo damage and MMP-2 and MMP-9 production[J]. Avicenna Journal of Dhyytmedicine, 2020, 10(3): 287–296.
- [30] LI LY, LAN C, HUANG T, *et al.* Natural products and extracts from plants as natural UV filters for sunscreens: A review [J]. Animal Model and Experiment Medicine, 2022(00): 1–13.

Editor: Yingzhi GUANG

Proofreader: Xinxu ZHU