Igor Bychkov
Victor Voronin   *Editors*

Information Technologies in the Research of Biodiversity

Proceedings of the International Conference "Information Technologies in the Research of Biodiversity", 11–14 September, 2018, Irkutsk, Russia

Springer

Springer Proceedings in Earth and Environmental Sciences

The series Springer Proceedings in Earth and Environmental Sciences publishes proceedings from scholarly meetings and workshops on all topics related to Environmental and Earth Sciences and related sciences. This series constitutes a comprehensive up-to-date source of reference on a field or subfield of relevance in Earth and Environmental Sciences. In addition to an overall evaluation of the interest, scientific quality, and timeliness of each proposal at the hands of the publisher, individual contributions are all refereed to the high quality standards of leading journals in the field. Thus, this series provides the research community with well-edited, authoritative reports on developments in the most exciting areas of environmental sciences, earth sciences and related fields.

More information about this series at http://www.springer.com/series/16067

Igor Bychkov · Victor Voronin
Editors

Information Technologies in the Research of Biodiversity

Proceedings of the International Conference "Information Technologies in the Research of Biodiversity", 11–14 September, 2018, Irkutsk, Russia

Springer

The I-test, used for taxa with overlapping ranges, provides a comparative analysis of ecological and climatic niches using measures I—the standardized Hellinger distance and the D—Schöner index [20].

At the same time, as Zink [22] notes, if different populations live in different climatic conditions, the I-test can have a high value, but this does not yet indicate their environmental-climatic divergence, since they can exist in different conditions due to environmental plasticity. In order to identify the ecological factors, and in this case, ecologo-climatic divergence at the genetic level, it is necessary to use a more complex B-test, which allows us to detect the presence of a genetically determined discrepancy for niches. Both tests are recommended to be carried out with all 19 variables in 100-fold repetition (100 replicates).

Figure 4 shows the result of the test for the identity of niches of two species of bluegrass—*P. palustris* and *P. nemoralis*. As a result of pairwise comparison of ecological and climatic niches constructed on the basis of all 19 biologically significant climatic variables for these species, the histograms with total information of 100 replicates were obtained. On the graphs (Fig. 4) the arrows indicate the position of the value, indicating overlapping of ecological-climatic niches between species. The diagram on the right illustrates the distribution of overlapping in replicates. On all graphs, the arrows stand at a considerable distance from the diagrams. This indicates that the null hypothesis about the identity of niche models should be rejected (Fig. 5).
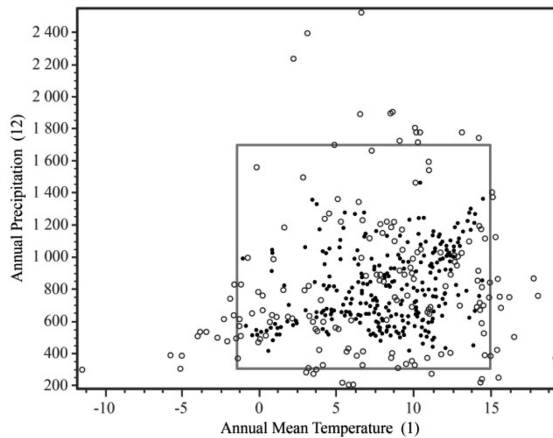

Fig. 4. Ecologo-climatic "envelope", obtained for *Poa palustris* 2 climatic variable.

The second important component of environmental and climate research is to obtain a model for the likely distributed of the species in accordance with its climatic needs. To obtain and evaluate the model, the sample is divided into 2 parts training and testing samplings. The training sampling serves to identify the ecological and climatic profile, and the test samplings is for verifying the quality of the obtained model. The resulting model is then projected onto an electronic map of the region under study.
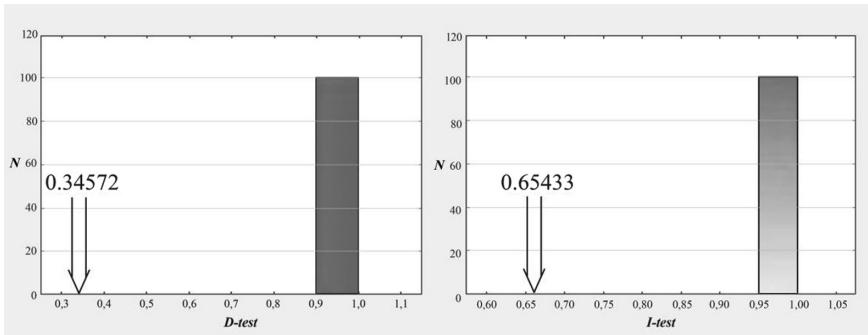
Fig. 5. I-тест–similarity of ecologo-climatic niches of Poa palustris and P. nemoralis in Asian Rusia. **A** – the standardized Hellinger distance. **B** – the Schöner index.

It shows the potential distribution of the species and the gradation of colors, from dark to light, determines the areas where this species can grow, and where it can spread in the future [3].

A darker tone on the map indicates areas with the most favorable combinations for each type of climatic characteristics. Since MaxEnt is currently recognized as one of the best algorithms for modeling the range of species, it is dynamically developing. Developers are taking into account the new aspects, and new opportunities that significantly improve the quality of the model. Such additional features are implemented in the SDMtoolbox application [23, 24] which is a freely distributed application to the ArcGIS licensed program [25]. Improvements relate both to training sampling and climatic characteristics.

In order to improve the quality of the training sampling, this application takes into account the features of the geographic coordinate system, where the area, enclosed between segments of parallels and meridians, varies in latitudinal direction due to convergence of meridians at the poles. The application also takes into account the uniformity of the selection of the material, since the collection is usually the most intensive near the settlements, along rivers, and generally in more inhabited areas, which distorts the picture of the climate preferences of the species. In addition, the application provides processing, thinning the training sample, depending on the uniformity of the geographical distribution of climatic parameters. The special options are supposed for operation of small samplings [26, 27].

All this allows you to more accurately determine the climate profile of the species and improves the quality of the model. Another problem is the correlation between bioclimatic data. As the developers of the program indicate, a high correlation between bioclimatic characters makes the evaluation of their contribution very difficult, and sometimes impossible. The SDMtoolbox application reduces the correlation between climatic characteristics by identifying and removing highly correlated variables.

The program provides an assessment of the predictive capability of the model, namely, to predict the finding of a species where it is, according to climatic needs, occurs, and its absence where it can not grow. This is of great importance in studying the ranges of invasive species and predicting their distribution, as well as in plant

introductions. The model is estimated using the Receiver Operating Curve (ROC) and the area under this curve—Area Under the Curve (AUC). ROC—the graph, which allows to estimate the quality of binary classification [12, 26, 28].

It reflects the ratio between the total number of characteristic carriers, correctly classified as carrying characteristics (ordinate axis), and the proportion of objects that do not carry characteristics, that are erroneously classified as bearing the characteristic (the abscissa axis). The quantitative interpretation of ROC is given by the AUC indicator—the area bounded by the ROC and the axis of the fraction of false positive classifications. AUC is an estimate of the ability of the model to indicate the presence of a species at the point of the raster where it is most likely to be located.

AUC is a measure of the area under ROC, ranging from 0.5 (random accuracy) to 1 (ideal discrimination). If the AUC is equal to or below 0.5, the model has no predictive value. A value of 0.8 for AUC means an 80% probability that, where the species is predicted to be, it will indeed be located [10].

The method also makes it possible to evaluate the contribution of each climatic variable in the obtained model of the species distribution, therefore we can evaluate the role of each biologically significant factor, included in the analysis. So, we receive a valuable environmental information that characterizes the species. The evaluation of the contribution of each variable in MaxEnt can be made in three independent ways: direct evaluation of the contribution as a percentage, evaluation after permutation and using the jackknife option. The most important are the last two. When permutation, the contribution of each variable is determined by randomly changing the value of this variable in the analysis, including all the data participating in the "learning" process: both the real presence points of the species, and the "background" ones. In this case, the value of each variable is expressed in a change in the evaluation of the obtained model. The more the estimation decreases, the more the model depends on this variable. In order to express this influence in percent, the values are normalized. The jackknife option consists of three steps. First, all variables are exclude from the analysis in turn, and the model is created with the remaining variables. Then the model is created with only one variable (each in turn). Finally, for comparison, a model is created with the participation of all variables.

Figure 6 presents a model of the potential distribution of *Poa palustris* in the territory of Asian Russia, taking into account the threshold of 10 percentile. The predictive capability of this model is estimated as good—the values of AUC-training and testing were 0.887, and 0.871, respectively. Both characteristics exceed 0.8, which corresponds to a good estimate [10]. The standard deviation was 0.016.

In Fig. 4 a richness of morphotypes of the *Stenopoa* section in the territory of Asian Russia is shown [29]. The widespread introduction of molecular methods into practice led to the development of phylogeographic studies and, as a result, the development of new GIS programs. At present, the applications for mapping of the spatial distribution of genetic diversity [10] and reconstruction of the most probable migration routes [23, 24]. have been developed.

In this brief overview not all the possibilities of using modern GIS technologies in botanical studies are listed, but undoubtedly that their application opens a new page in ecological and geographical studies of species (Fig. 7).
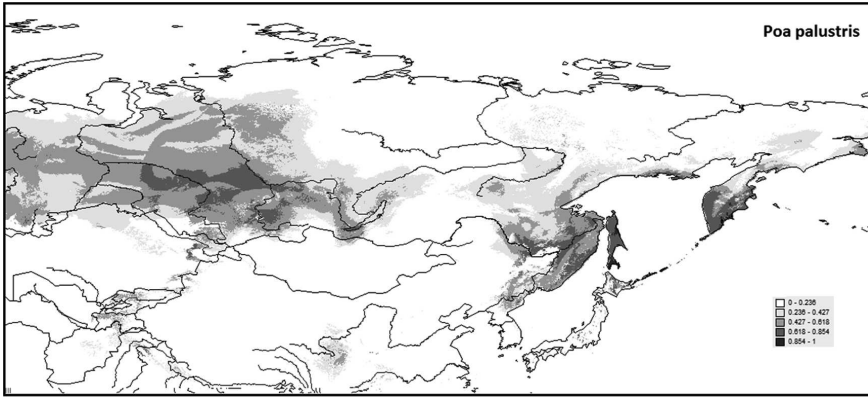
Fig. 6. The model of potential distribution of *Poa palustris* in Asian Russia, obtained in MaxEnt for the current climate, based the ecologo-climatic variables Bio1, Bio2, Bio5, Bio7, Bio8, Bio12, Bio15.
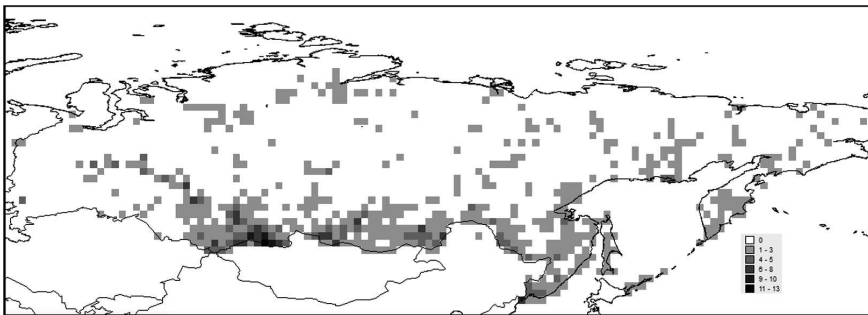

Fig. 7. The richness of flora of Asian Russia with morphotypes of bluegrasses of section Stenopoa.

References

1. Goodwin, B.J., McAllister, A.J., Fahrig, L.: Predicting invasiveness of plants species based on biological information. Concerv. Biol. **13**, 422–426 (1999)
2. Welk, E.: Constrains in range predictions of invasive plant species due to non-equilibrium distribution patterns: Purple loosestrife (*Lythrum salicaria*) in North America. Ecol. Model. **179**, 551–567 (2004)
3. Ward, D.F.: Modelling the potential geographic distribution of invasive ant in New Zeland. Biol. Invasions **9**, 723–735 (2007)
4. Hellmann, J.J., Byers, J.E., Bierwagen, B.G., Dukes, J.S.: Five potential consequences of climate change for invasive species. Conserv. Biol. **22**(3), 534–543 (2008)

5. Hutchinson, G.E.: Concluding remarks. Cold Spring Harb. Symp. Quant. Biol. **22**, 415–422 (1957)
6. MacAthur, R.H.: The theory of the niche. In: Lewontin, R.C. (ed.) Population Biology and Evolution, pp. 159–176. Syracuse University Press, Syracuse (1968)
7. Khlebosolov, E.I.: The theory of niche in ecology: the history and modern status. Russ. Ornithol. J. **11**(203), 1019–1037 (2002)
8. Hijmans, R.J., Guarino, L., Jarvis, A., O'Brien, R., Mathur, P., Bussink, C., Cruz, M., Barrantes, I., Rojas, E.: DIVA-GIS, version 5.2. http://www.diva-gis.org/DIVA-GIS5_manual.pdf. Accessed 18 Apr 2018
9. Franklin, J.: Mapping Species Distribution: Spatial Inference and Prediction. Cambridge University Press, Cambridge (2009)
10. Scheldeman, X., van Zonneveld, M.: Training Manual on Spatial Analysis of Plant Diversity and Distribution. Biodiversity International, Rome (2010)
11. Phillips, S.J.: AT&T. Research. A brief tutorial on MaxEnt. http://www.cs.princeton.edu/~schapire/maxent/tutorial/tutorial.doc. Accessed 21 Jan 2018
12. Elith, J., Graham, C.H., Anderson, R.P., Dudik, M., Ferrier, S., Guisan, A., Hijmans, R.J., Huettmann, F., Leathwick, J., Schapire, R., Soberso, J., Williams, S., Wisz, M., Zimmerman, N.: Novel methods improve prediction of species' distributions from occurrence data. Ecogeography **29**, 129–151 (2006)
13. Ramirez, J., Bueno-Cabrera, A.: Working with climate data and niche modeling. Creation of bioclimatic variables. http://gisweb.ciat.cgiar.org/GCMPage?docs?tutorial_bcvars_creation.pdf. Accessed 21 Nov 2017
14. Beaumont, L.J., Hughes, L., Poulsen, M.: Predicting species distributions: use of climate parameters in BIOCLIM and its impact on predicting of species' current and future distributions. Ecol. Model. **176**, 250–269 (2005)
15. Philips, S.J., Anderson, R.P., Schapire, R.E.: Maximum entropy modeling of species geographic distributions. Ecol. Model. **190**, 231–259 (2006)
16. Phillips, S.J., Dudic, M.: Modeling of species distribution with Maxent: new extentions and a comprehensive evaluation. Ecography **31**, 161–175 (2008)
17. Hijmans, R.J., Guarino, L., Mathur, P.: DIVA-GIS Version 7.5. http://diva-gis.org/docs/DIVA-GIS_manual_7.pdf. Accessed 25 Jan 2018
18. Hijmans, R.J., Cameron, S., Parra, J.: Climate date from Worldclim. http://www.worldclim.org. Accessed 25 Jan 2018
19. Olonova, M.V., Gudkova, P.D.: Bioclimatic Modeling. Tutorial for Students. Tomsk State University, Tomsk (2017)
20. Warren, D.L., Glor, R.E., Turelli, M.: Environmental niche equivalency versus conservatism: quantitative approaches to niche evolution. Evolution **62**, 2868–2883 (2008)
21. Warren, D.L., Glor, R.E., Turelli, M.: ENMTools User Manual v 1.3. http://www.danwarren.net/enmtools/builds/ENMTools_1.4.3.zip. Accessed 13 Apr 2018
22. Zink, R.M.: Genetics, morphology, and ecological niche modeling do not support the subspecies status of the endangered Southwestern Willow Flycatcher (Empidonax traillii extimus). Condor **117**(1), 76–86 (2015)
23. Brown, J.L.: SDMtoolbox User Guide. http://www.jasonleebrown.org/SDMtoolbox/current/User_Guide_SDMtoolbox.pdf. Accessed 15 Mar 2018
24. Brown, J.L.: SDMtoolbox: a python-based GIS toolkit for landscape genetic, biogeographic and species distribution model analyses. Methods Ecol. Evol. **5**(7), 694–700 (2014)
25. ArcGIS (ESRI) Desktop and Spatial Analyst Extension: Release 10.1. Environmental Systems Research Institute, 2012. Redlands, CA. http://www.esri.com. Accessed 16 Apr 2018

26. Anderson, R.P., Lew, D., Peterson, A.T.: Evaluating predictive models of species' distributions: criteria for selecting models. Ecol. Model. **162**, 211–232 (2003)
27. Shcheglovitova, M., Anderson, R.P.: Estimation optimal complexity for ecological niche models: a jackknife approach for species with small sample sizes. Ecol. Model. **269**, 9–17 (2013)
28. Araújo, M.B., Pearson, R.G., Thuiller, W., Erhard, M.: Validation of species-climate impact models under climate change. Glob. Change Biol. **1**, 1504–1513 (2005)
29. Olonova, M.V., Mezina, N.S., Gorina, N.V., Vysokikh, T.S.: Phenetic variability of bluegrasses (Poa L., *Poaceae*) section *Stenopoa* Dum in Asian Russia. Int. J. Environ. Stud. **74**(5), 715–723 (2017)

Diatom Analysis Using SOQL Language Interpreter

Y. V. Avramenko[1(✉)], R. K. Fedorov[1], and A. D. Firsova[2]

[1] Matrosov Institute for System Dynamics and Control Theory of the Siberian Branch of the Russian Academy of Sciences, Irkutsk 664033, Russia
`avramenko@icc.com, fedorov@icc.ru`
[2] Limnological Institute Siberian Branch of the Russian Academy of Sciences, Irkutsk 664033, Russia
`firsova@lin.irk.ru`

Abstract. Article proposes the method based on logic programming image recognition in order to solve the diatom selection problem. This method is suitable for diatoms selection, as it allows describing every type of diatoms and performing the search. This analysis is based on the study of micro algae organisms, particularly counting the number of diatoms remaining. Counting is usually performed manually by the microscope operator, so, in this paper, we suggest using the SOQL interpreter to find objects to help the operator to estimate the amount of remaining diatom algae.

Keywords: Deformable models · Pattern recognition · Diatoms

1 Introduction

One of the methods of monitoring the ecological state of Lake Baikal is the analysis of diatom algae. Diatom algae belong to the leading group of phytoplankton participates in the food chain and reacts to changes in environmental conditions. The process of estimating the amount of diatom algae in Lake Baikal is performed by the operator and takes a long time. The amount of diatom algae is taken from a sample obtained with a microscope. This process consists of a sequential visual analysis of the magnified parts of the original image. Existing methods and programs implementing them, designed to process data and structural characteristics of phytoplankton, are mainly oriented to a certain set of microalgae. They are equipped with various filters to improve the perception of the image. Here are descriptions of some methods and software.

1.1 Methods

In [1], the Center Supported Segmentation (CSS) method is proposed for segmenting objects on an image by calculating the center of mass of the regions. The method is more accurate than segmentation by a threshold or a watershed method. The limitation of the method is the absence of the possibility of classification.

In [2], a method for detecting a set of overlapping objects using a HOG-SVM detector with modification is presented. The HOG-SVM detector operates on the basis