OPEN DISCUSSION ON THE USE OF SPECIES DISTRIBUTION MODELLING IN THE DEEP-SEA USING R

Instructor: Dr José Manuel González-Irusta (ATLAS partner, Azores)

Date: May 23rd –25th, 2018

Language: English

Location: IMAR-DOP, Rua Prof. Doutor Frederico Machado 9901-862, HORTA, Portugal

Register: IMAR-DOP staff interested in the course can register sending an e-mail to gonzalezirusta@gmail.com before the 20th of May. Places limited to the room capacity.

Distribution Models (DMs), also called species distribution models, habitat suitability models or ecological niche models are correlative statistical models which combines georeferenced information of species distribution with environmental variables in a GIS format (raster) to predict the geographical distribution of species or habitats. Distribution models is a fast-moving field, which is receiving growing attention every year (one of the top 5 research front in ecology according to the ISI's Essential Science Indicators in July of 2012).

The aim of the talks is to teach students how to develop Distribution Models (DMs) using R. It is expected that students will acquire a basic knowledge about the ecological and statistical basis of DMs as well as the capacity to perform ecological niche models using R. The attendants of the course will learn about the differences between the available approaches (presence-absence, only presence, abundance), the advantages and disadvantages of each approach as well as how to use some of the most popular algorithms in DMs (GAMs, MaxEnt and Random Forest) and the differences between them.

Although the course will be developed using a case study provided by the IMAR team, attendants are also encouraged to **bring their own data** to produce their own preliminary results.

The course will be mainly for IMAR students, however few ATLAS partners developing DMs for different portion of the Atlantic Ocean will be invited to attend the course and their assistance will be prioritised. Places are limited to 10-20 students depending on the room capacity.

Requirements

All attendants must bring their **own laptop with R, R-studio and Q-GIS** (all free software) or other GIS software installed and be able to use R (basic-intermediate level) and other equivalent GIS software (basic knowledge).

Please also **install the following R packages** on your laptop before the start of the course: *raster, sp, rgeos, maptools, proj4, ncdf4, mgcv, car, dismo, rJava, ENMeval, randomForest, VSURF, randomForestSRC, ggRandomForests, ggplot2, rgdal, boot, gstat, automap, ape, utils*

Day 1 Preparing the data

During the first day students will receive a first introduction to DMs, they will learn how to work with spatial objects in R, how to compute depth derivates in R (slope, BPI, aspect) or how to work with oceanographic models freely available online.

Introduction to the concept of SDMs. Data types. Only presences, Presence/Absences, Abundance, other. Environmental layers. Rasters, working with Spatial data in R. Projections in R. The mask. Rescaling environmental layers. The raster stack. SpatialPixelDataFrames. Extract values from points. Depth and depth derivates. Environmental layers from oceanographic models (netcdf files). Correlation matrix from environmental layers. Variable selection.

Day 2 Run the models

On the second day students will receive a theoretical introduction to the most common statistical algorithms used in DMs: MAXENT, Random Forest and GAMs. Furthermore, they will learn how to apply these algorithms to the data processed on day one and they will produce their own model for each algorithm including the geographical model (the distribution map).

Type of models. Pseudoabsences, backgrounds absences, presence absences Background absence models (MAXENT). Other models (Random Forest and GAMs). Limitations of background absence models. Variable selection into the models. Response curves. Variable importance. Link function for presence/absence data. Link function for other type of data.

Day 3 Evaluate the models

On the third and last day students will learn how to evaluate DMs using cross-validation and how to detect spatial autocorrelation. They will receive a theorical introduction to the main evaluation metrics for DMs including Area Under the Curve (AUC), sensitivity, specificity, confusion matrix and kappa. Particular care will be paid to illustrate advantages and disadvantages of these metrics and how

they should be used to evaluate models. Finally, students will receive some information about the main techniques of threshold selection.

Spatial autocorrelation. Specificity, sensitivity, confusion matrix, AUC, Kappa. Cross-validation. Spearman and Pearson coefficients. Explained deviance. Threshold selection. Maximum kappa. Maximize sens-spec. Evaluation limitations. Spatial autocorrelation effect.

AGENDA

AN INTRODUCTION TO DISTRIBUTION MODELS USING R

May 23-25, 2018 Horta, Azores, Portugal

Wednesday, 23 May

10:00 Welcome

10:10-11:00 DMs. Main concepts

11:00:11:15 Coffee break

11:15-12:45 First practical session.

12:45-13:00 Coffee break

13:00-14:30 Second practical session.

Thursday, 24 May

10:00-11:00 GAMs, MaxEnt and Random Forest

11:00:11:15 Coffee break

11:15-12:45 Third practical session

12:45-13:00 Coffee break

13:00-14:30 Fourth practical session

Thursday, 24 May

10:00-11:00 Evaluation of DMs

11:00:11:15 Coffee break

11:15-12:45 Fifth practical session

12:45-13:00 Coffee break

13:00-14:30 Sixth practical session

RECOMMENDED LITERATURE

1. Bedia, J., Busqué, J., and Gutiérrez, J. M. 2011. Predicting plant species distribution across an alpine rangeland in northern Spain. A comparison of probabilistic methods. Applied Vegetation Science, 14: 415–432.
2. Breiman, L. 2001. RANDOM FOREST. Machine Learning, 45: 5–32. http://link.springer.com/10.1023/A:1010933404324.
3. Breiman, L. 2017. Classification and regression trees. New York: Routledge.
4. Brotons, L., Thuiller, W., Araújo, M. B., and Hirzel, A. H. 2004. Presence-absence versus presence-only modelling methods for predicting bird habitat suitability. Ecography, 27: 437–448.
5. Dorazio, R. M. 2012. Predicting the Geographic Distribution of a Species from Presence-Only Data Subject to Detection Errors. Biometrics, 68: 1303–1312.
6. Dormann, C. F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carré, G., Marquéz, J. R. G., *et al.* 2013. Collinearity: A review of methods to deal with it and a simulation study evaluating their performance. Ecography, 36: 27–46.
7. Dormann, f., McPherson, C.M., Araújo, J.B., Bivand, M., Bolliger, R., Carl, J., G., et al. (2007). Methods to account for spatial autocorrelation in the analysis of species distributional data: a review. Ecography, 30: 609-628.
8. Ehrlinger, J. (2016). ggRandomForests: random forests for regression. arXiv preprint arXiv:1501.07196.
9. Elith, J., and Leathwick, J. R. 2009. Species Distribution Models: Ecological Explanation and Prediction Across Space and Time. Annual Review of Ecology, Evolution, and Systematics, 40: 677–697.
10. Elith, J., Phillips, S. J., Hastie, T., Dudík, M., Chee, Y. E., and Yates, C. J. 2011. A statistical explanation of MaxEnt for ecologists. Diversity and Distributions, 17: 43–57.
11. Freeman, E. A., & Moisen, G. G. (2008). A comparison of the performance of threshold criteria for binary classification in terms of predicted prevalence and kappa. Ecological Modelling, 217: 48-58.

12. Genuer, Robin, Poggi, J.M., Tuleau-Malot, C. 2010. Variable selection using random forests." Pattern Recognition Letters 31: 2225-2236.
13. Graham, C. H., Elith, J., Hijmans, R. J., Guisan, A., Townsend Peterson, A., Loiselle, B. A., Anderson, R. P., *et al.* 2008. The influence of spatial errors in species occurrence data used in distribution models. Journal of Applied Ecology, 45: 239–247.
14. Hernández, P. A., Graham, C. H., Master, L. L., and Albert, D. L. 2006. The effect of sample size and species characteristics on performance of different species distribution modeling methods. Ecography, 29: 773–785.
15. Hutchinson, G.E. (1957) Concluding remarks. Cold Spring Harbor Symposia on Quantitative Biology, 22, 145-159.
16. Liu, C., White, M., & Newell, G. (2013). Selecting thresholds for the prediction of species occurrence with presence-only data. Journal of biogeography, 40: 778-789.
17. Lobo, J. M., Jiménez-Valverde, A., & Hortal, J. (2010). The uncertain nature of absences and their importance in species distribution modelling. Ecography, 33(1), 103-114.
18. Lozier, J. D., Aniello, P., & Hickerson, M. J. (2009). Predicting the distribution of Sasquatch in western North America: anything goes with ecological niche modelling. Journal of Biogeography, 36(9), 1623-1627.
19. Jiménez-Valverde, A. (2012). Insights into the area under the receiver operating characteristic curve (AUC) as a discrimination measure in species distribution modelling. Global Ecology and Biogeography, 21(4), 498-507.
20. Monk, J. 2014. How long should we ignore imperfect detection of species in the marine environment when modelling their distribution? Fish and Fisheries, 15: 352–358.
21. Naimi, B., Skidmore, A. K., Groen, T. A., and Hamm, N. A. S. 2011. Spatial autocorrelation in predictors reduces the impact of positional uncertainty in occurrence data on species distribution modelling. Journal of Biogeography, 38: 1497–1509.
22. Osborne, P. E., and Leitão, P. J. 2009. Effects of species and habitat positional errors on the performance and interpretation of species distribution models. Diversity and Distributions, 15: 671–681.
23. Phillips, S. J., Anderson, R. P., and Schapire, R. E. 2006. Maximum entropy modeling of species geographic distributions. Ecological Modelling, 190: 231–259.
24. Renner, I. W., and Warton, D. I. 2013. Equivalence of MAXENT and Poisson Point Process Models for Species Distribution Modeling in Ecology. Biometrics.
25. Steven, P., Dudik, M., Jane, E., Graham, C., Lehmann, A., Leathwick, J., and Ferrier, S. 2009. Sample selection bias and presence-only distribution models : implications for background and pseudo-absence data Reference

Sample selection bias and presence-only distribution models : implications for background and pseudo-absence data. Ecological Applications, 19: 181–197.
26. Stockwell, D. R. B., and Peterson, A. T. 2002. Effects of sample size on accuracy of species distribution models. Ecological Modelling, 148: 1–13.
27. Wisz, M. S., Hijmans, R. J., Li, J., Peterson, A. T., Graham, C. H., Guisan, A., Elith, J., *et al.* 2008. Effects of sample size on the performance of species distribution models. Diversity and Distributions, 14: 763–773.
28. Yackulic, C. B., Chandler, R., Zipkin, E. F., Royle, J. A., Nichols, J. D., Campbell Grant, E. H., and Veran, S. 2013. Presence-only modelling using MAXENT: When can we trust the inferences? Methods in Ecology and Evolution, 4: 236-243.
29. Zuur, A. F., Ieno, E. N., and Elphick, C. S. 2010. A protocol for data exploration to avoid common statistical problems: Data exploration. Methods in Ecology and Evolution, 1: 3–14.