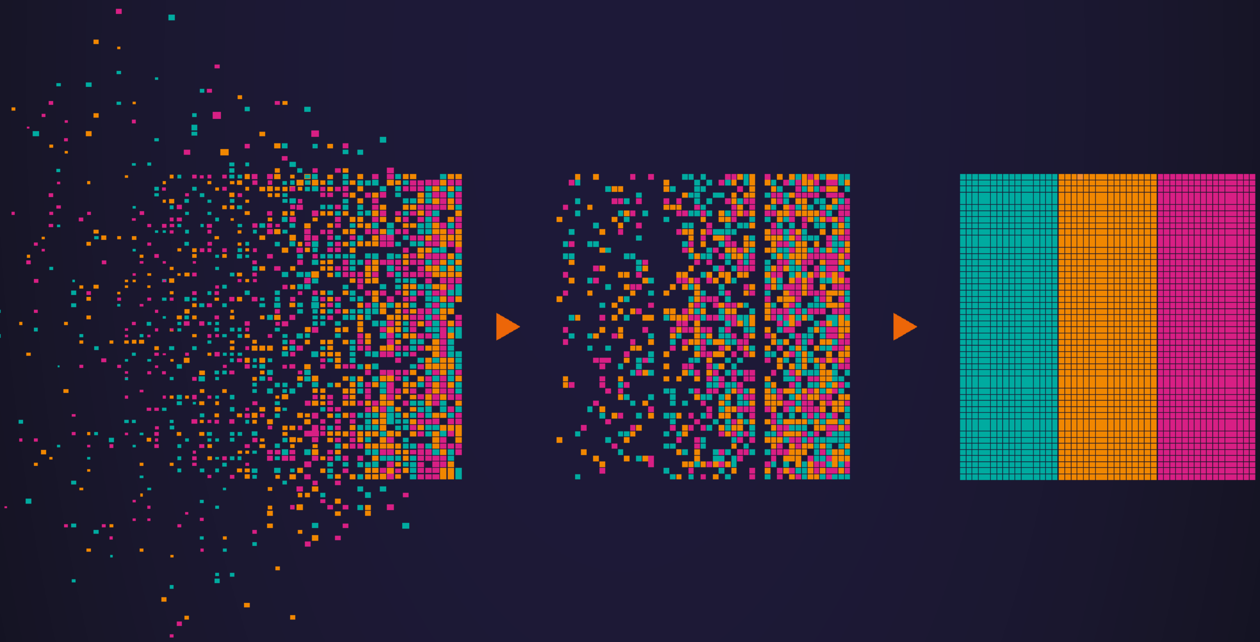
BEYOND ASSUMPTIONS:

Unraveling Data Limitations In Predictive Ecology



Lukáš Gábor

To Vítězslav Moudrý

My exceptional supervisor, mentor and more importantly friend.

This dissertation entitled *"Beyond Assumptions: Unraveling Data Limitations in Predictive Ecology"* is submitted to the Department of Spatial Sciences, Faculty of Environmental Sciences, Czech University of Life Sciences in Prague in July 2024 for the degree of *Doctor of Philosophy.* All sources of information, including text, illustrations, tables, and images, have been duly acknowledged and cited.

Essentially, all models are wrong, but some are useful.

-George Box

Acknowledgements

I would like to express my sincere gratitude to several people who have played a pivotal role in completing my dissertation.

First, I would like to recognize my supervisor, Vítek Moudrý, for your invaluable guidance, unwavering support, and exceptional mentorship throughout my Ph.D. journey. Your expertise, dedication, and encouragement have been instrumental in shaping the direction of my research and pushing me to reach my full potential.

I am also deeply grateful to the Fulbright Commission, Petr Keil, and Walter Jetz for the incredible opportunity to pursue a significant portion of my Ph.D. studies at Yale University. Yale's resources, intellectual environment, and collaborative opportunities have greatly enriched my research experience. Walter, I am truly indebted to you for your generosity and belief in my abilities.

Furthermore, I would like to thank my family for their love, encouragement, and relentless support. Mum and Dad, your patience, understanding, and certainty in my abilities have been an immense source of strength throughout this academic journey.

Special appreciation goes to Ruchika C. Rana for your unwavering support. You appeared unexpectedly in my life, but your presence has been a source of solace during the ups and downs of this, not only, academic endeavor. I am grateful for your insightful discussions, encouragement, and love you provided, which propelled me forward during challenging moments.

Lastly, I want to express my profound appreciation to all my friends, colleagues, and mentors who have supported me along the way. Your constructive feedback, stimulating discussions, and shared experiences have been invaluable in shaping my ideas and strengthening my research.

The scientific papers included in my dissertation were made possible through the generous support of the below-listed scientific grants. I am sincerely grateful for the financial support provided by these grants, as they enabled me to conduct the necessary research and contribute to advancing knowledge in my field.

- Technological grant agency of the Czech Republic (grant no. SS02030018 DivLand)

- OP RDE Improvement in Quality of the Internal Grant Scheme at CZU, reg. no. CZ.02.2.69/0.0/0.0/19_073/0016944 (grant no. 43/2021).

- Internal Grant Agency of Faculty of Environmental Sciences, Czech Univ. of Life Sciences Prague (grant no. 2020B0009)

- Internal grant agency of the faculty of environmental sciences, Czech Univ. of Life Sciences, Prague (grant no. 2021B0009)

Abstract

Context

Species distribution models (SDMs) have proven valuable in filling gaps in our knowledge of species occurrences. However, despite their broad applicability, SDMs exhibit critical shortcomings due to limitations in species occurrence data and in environmetal variables. Typical example of such limitation in species records is spatial uncertainty in their location, ranging from a few meters to tens of kilometers (e.g., positional uncertainty in the GBIF database can exceed 300 kilometers). Similarly, environmental variables may be limited by how much detailed information about the environment they provide. For example, land cover types typically indicate the amount of habitat within a spatial unit. However, it is possible that a simple binary presence/absence of suitable habitat may be the only information available or even more important aspect than area, determining species distribution.

Objectives

This dissertation had three main objectives. Firstly, I explored the potential of an alternative approach to incorporate environmental variables into models (i.e., binary versus continuous habitat information). Secondly, I investigated the influence of positional uncertainty in species records on the ecological interpretability of models. Thirdly, I evaluated whether the appropriateness of using binary or proportional (continuous) type of variables, and the influence of positional error, is affected by the scale of the analyses. Specifically, I addressed the following research questions: a) Can binary land cover predictors provide models of higher accuracy than traditionally used proportional variables? b) If so, what is the role of spatial grain in determining the usability of binary land cover predictors? c) To what extent does positional uncertainty in species occurrence data affect model parameter estimation and the ecological interpretability of species distribution models? d) What are the trade-offs between analysis grain and positional uncertainty in modeling species distributions?.

Results

Results indicated that models' performances were not affected by the type of the adopted habitat variable (proportional or binary but the usability of binary variables decreased with coarsening the resolution (i.e., binary representation of habitat is useful at finer grain sizes of approx. 1km^2. Results confirmed that model performance decrease with increasing positional error in species records, as demonstrated in prior studies. However, I have shown that coarsening the analysis grain to compensate for positional error did not improve model performance as was widely assumed. This, however, doesn't mean we should exclude species records with high positional uncertainty from our studies, because the negative consequences of positional uncertainty on model performance did not extend as strongly to the ecological interpretability of the models.

Conclusions

These findings are encouraging for practitioners using SDMs to reveal relationships between species occurrences and its environmental drivers as such relationship can be to some degree estimated using positionally uncertain data and simple environmental variables describing presence or absence of a habitat. On the other hand, my findings show that positional uncertainty in species data can cause inaccurate spatial predictions leading to inaccurate maps of species distributions, especially in heterogeneous environments and when using fine resolution environmental data. Therefore, such models are not suitable for tasks like setting up protected areas or prioritizing conservation efforts.

Abstract in Czech

Souvislosti

Modely druhové distribuce (SDMs) jsou důležitým nástrojem při doplňování mezer v našich znalostech o výskytech druhů. Navzdory tomu, že se tyto modely často používají v ekologických studiích, mají zásadní nedostatky kvůli nepřesnostem v datech o výskytech druhů a environmentálních prediktorech. Typickým příkladem takového omezení je polohová nejistota v záznamech druhů, která může být od několika metrů až po desítky kilometrů (např. polohová nejistota v databázi GBIF může přesáhnout 300 kilometrů). Environmentální prediktory pak mohou být omezeny tím, jak přesnou informaci o prostředí, ve kterém se druh nalézá, poskytují. Například proměnné krajinného pokryvu obvykle udávají rozlohu nebo podíl habitatu v rámci určité oblasti. Avšak co když, jednoduchá binární informace (přítomnost/absence) o vhodném habitatu může být jedinou dostupnou informací nebo dokonce důležitějším aspektem při určení druhové distribuce než informace o celkové rozloze habitatu?

Cíle

Tato disertační práce měla tři hlavní cíle. Za prvé jsem zkoumal možnosti použití nového typu environmentálních prediktorů (binárních dat), které obsahují pouze informaci o přítomnosti nebo absenci vhodného habitatu. Za druhé jsem zkoumal vliv polohové nejistoty v druhových záznamech na ekologickou interpretovatelnost modelů. Třetím cílem pak bylo posouzení role prostorového měřítka na modely, které používali binární prediktory a druhová data s různou polohovou nejistotou. Konkrétní výzkumné otázky byly: a) Mohou binární environmentální prediktory krajinného pokryvu zvýšit přesnost modelů? b) Pokud ano, jaká je role použitého prostorového měřítka? c) Do jaké míry ovlivňuje polohová nejistota v druhových datech ekologickou interpretovatelnost modelů? d) Jak spolu souvisí a polohová chyba druhových dat?

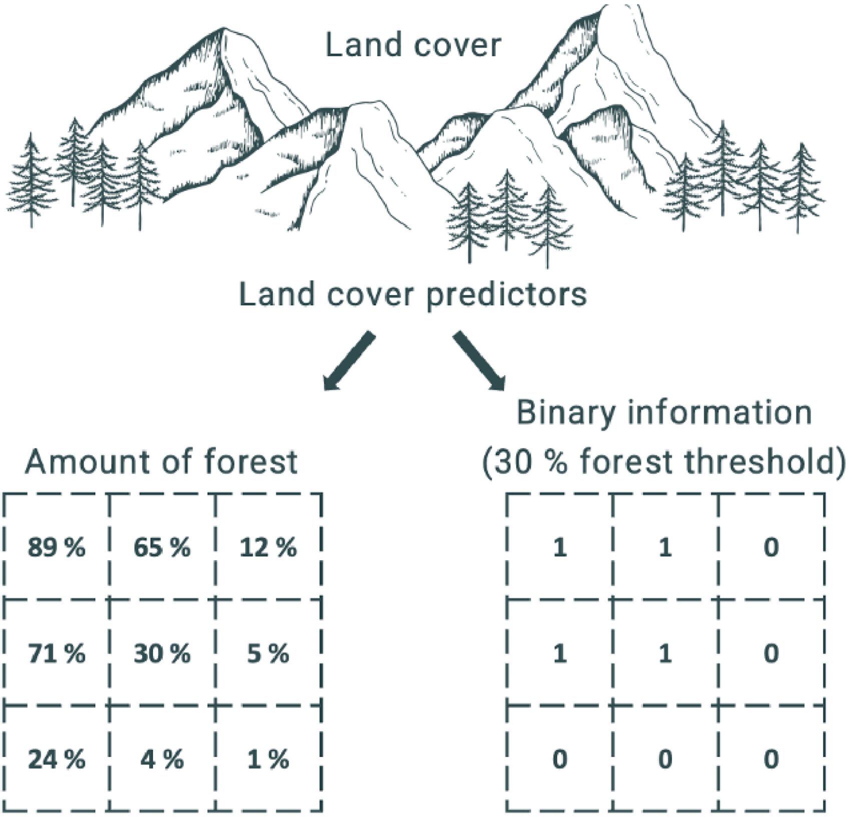
Výsledky

Výsledky ukázaly, že přesnost modelů nebyla významně ovlivněna typem použitých environmentální prediktorů (proporcionální nebo binární infomrace o vhodném habitatu). Je nicméně důležité říct, že ale použitelnost binárních prediktorů klesala s hrubším prostorovým rozlišením. To znamená, že binární reprezentace habitatu je užitečná především při použití prediktorů s vyšším prostorovým rozlišením (cca od 1km^2). Výsledky kromě toho potvrdily, že přesnost modelů klesá se zvyšující se polohovou nejistotou v záznamech druhů, tak jak bylo prokázáno v předchozích studiích. Důležitým závěrem práce je fakt, že zhoršení prostorového měřítka nekompenzuje negativní vliv polohové nejistoty, jak se všeobecně předpokládalo. To však neznamená, že bychom neměli při modelování záznamy druhů s vysokou polohovou nejistotou používat.

Závěry

Jak ukázaly výsledky, polohově neurčitá data a binární informaci o přítomnosti habitatu lze za určitých podmínek využít pro studium vztahů mezi organismy a prostředím. Na druhou stranu mapy druhové distribuce vycházející z modelů založených na datech s vysokou polohovu nejistotou, jsou nepřesné (zejména v heterogenním prostředí a při použití environmentálních prediktorů s vysokým prostorovým rozlišením) a nevhodné pro aplikace v ochraně přírody.

distribution projections. It is, therefore, imperative to acknowledge and quantify these uncertainties to ensure the robustness and reliability of SDMs.



Amount of forest

89 %	65 %	12 %
71 %	30 %	5 %
24 %	4 %	1 %

Binary information (30 % forest threshold)

1	1	0
1	1	0
0	0	0

Figure 2.2: Graphical representation of the hypothesis. Land cover predictors are typically incorporated by representing the amount of specific land cover types within individual sites. But what if, for some species, the total habitat area is less relevant than the simple fact that a particular habitat is present or absent?

2.2.2. Species records

In SDMs, species records are typically represented as presence-only or presence-absence data. Presence-only data include records that only indicate the presence of a species at a particular location without information on its absence. On the other hand, presence-absence data provide information on the presence and absence of a species at specific sites. While presence-absence data are considered more informative for modeling species distributions, they are often scarcer and require careful sampling design to ensure adequate representation of absence locations (Franklin 2010, Guillera-Arroita et al. 2015). Species records can be obtained from various sources, including field surveys, citizen science initiatives, museum collections, and literature reviews, or generated based on expert knowledge.

Field surveys involve systematic sampling or observations researchers conduct to record species' presence or absence at specific locations. These surveys often employ various sampling techniques, such as transect surveys, point counts, quadrat sampling, or trapping meth-

3.1.1. Abstract

The representation of a land cover type (i.e., habitat) within an area is often used as an explanatory variable in species distribution models. However, it is possible that a simple binary presence/absence of the suitable habitat might be the most important determinant of the presence/absence of some species and, thus, be a better predictor of species occurrence than the continuous parameter (area). We hypothesize that the binary predictor is more suitable for relatively rare habitats (e.g., wetlands) while for common habitats (e.g., forests) the amount of the focal habitat is a better predictor. We used the Third Atlas of Breeding Birds in the Czech Republic as the source of species distribution data and CORINE Land Cover inventory as the source of the landcover information. To test our hypothesis, we fitted generalized linear models of 32 water and 32 forest bird species. Our results show that for water bird species, models using binary predictors (presence/absence of the habitat) performed better than models with continuous predictors (i.e., the amount of the habitat); for forest species, however, we observed the opposite. Thus, future studies using habitats as predictors of species occurrences should consider the prevalence of the habitat in the landscape, and the biological role of the habitat type in the particular species' life history. In addition, performing a preliminary comparison of the performance of the binary and continuous versions of habitat predictors (e.g., using information criteria) prior to modelling, during variable selection, can be beneficial. These are simple steps that will improve explanatory and predictive performance of models of species distributions in biogeography, community ecology, macroecology, and ecological conservation.

Keywords: Binary data, Continuous data, Land cover, Niche models, Variable selection

habitat data in the field as simply recording the presence of a habitat is significantly less time-consuming than recording its total area.

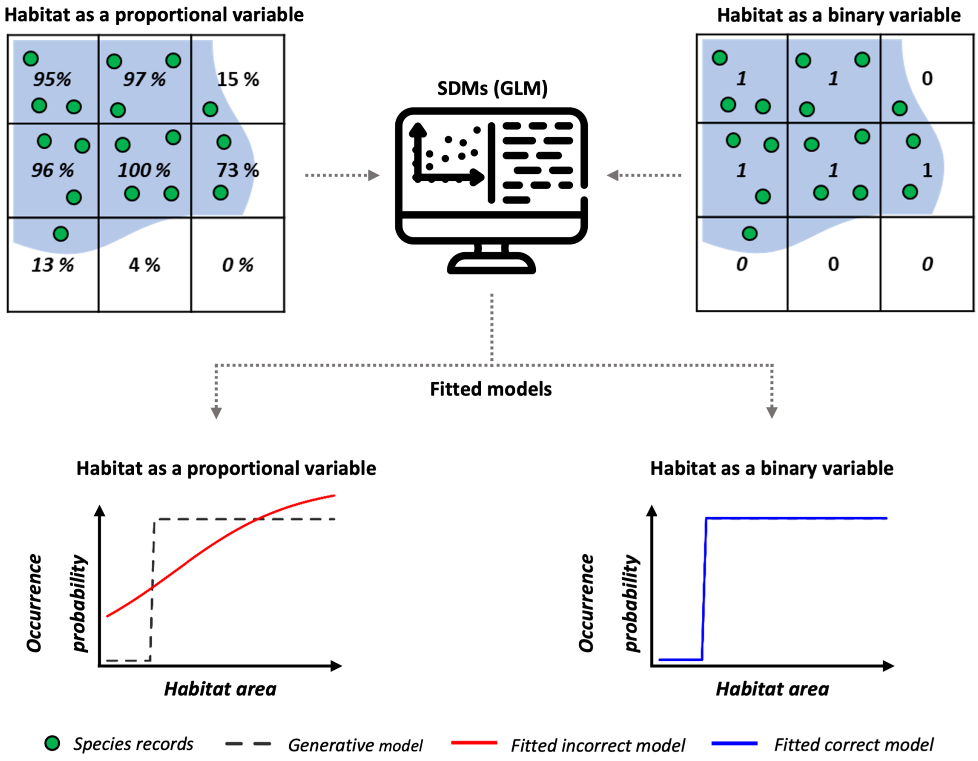


Figure 3.5: Graphical rationale for the hypothesis on binary variables. We modelled species occurrence probability as a step function (generative model, black dashed line) to generate 100 presences/absences of a species that needs at least 20% of land-cover within a spatial unit, drawn from a Bernoulli distribution with species occurrence probability as a parameter. Then, we fitted species distribution models (GLM) with either proportional (incorrect model, red line) and binary coverage (correct model, blue line), as a variable.

It is well-recognized that SDMs are grain-dependent (Elith and Leathwick 2009) and empirical evidence has shown that species exhibit stronger responses to their environment at certain grains than others (see reviews by Miguet et al. 2016 and Moudrý et al. 2023b). Therefore, the choice of the grain constitutes an important part of the modeling process as it can affect the ability to detect the response (Mertes and Jetz 2018, Luebert et al. 2022, Wunderlich, et al. 2022, Lu and Jetz 2023). In addition, the choice of the grain is often determined by data availability rather than study goals. This results in large variability of grains adopted in existing studies, from a few meters (e.g., Bazzichetto et al. 2018, Lecours et al. 2020, Casanelles-Abella et al. 2022, Stark and Fridley 2022) to many kilometers (e.g., Kleisner et al. 2017, Norberg et al. 2019, Zarzo-Arias et al. 2022). The grain size is also essential when selecting environmental variables (Pearson and Dawson 2003; Moudrý et al. 2023b), and it is likely to be critical when deciding whether to use land-cover as a binary or continuous variable when modelling species occurrence. However, this has never been thoroughly tested.

3.3. Species distribution models affected by positional uncertainty in species occurrences can still be ecologically interpretable

Lukáš Gábor, Walter Jetz, Alejandra Zarzo-Arias, Kevin Winner, Scott Yanco, Stefan Pinkert, Charles J. Marsh, Matthew S. Rogan, Jussi Mäkinen, Duccio Rocchini, Vojtěch Barták, Marco Malavasi, Petr Balej and Vítězslav Moudrý

Adapted from Ecography, 2023, e06358.

Publication metrics:
Quartile (2023): Ecology, Evolution, Behavior and Systematics (Q1)
Impact Factor (2023): 6.08
SCImago Journal Rank (2023): 2.37

The first author contributed to the study as follows: study conception and design (lead), data curation (lead), analysis and interpretation of results (lead), visualization (lead), writing – original draft (lead), writing – review and editing (equal), overall study supervision (equal), funding acquisition (lead).

The link to the published article and supplementary materials can be found here:
https://onlinelibrary.wiley.com/doi/full/10.1111/ecog.06358

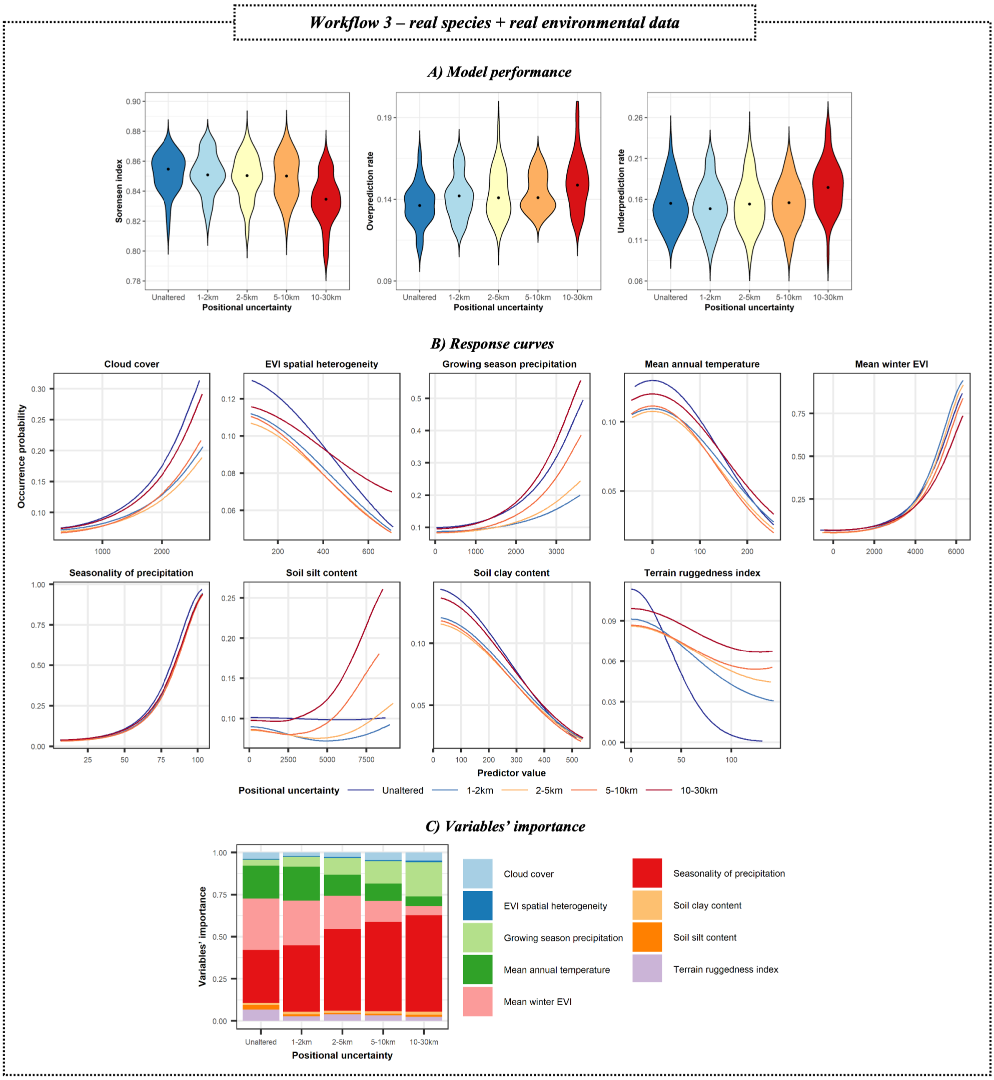


Figure 3.12: Resulting performance metrics (A), response curves (B) and variables' importance (C) of unaltered and altered models for all scenarios with real species and real environmental data. Values represent averages of the 50 repetitions.

The general increase in over and underprediction rates across all workflows implies that models fit to data with positional error tended to overpredict and, at the same time, underpredict species habitat suitability. Therefore, using positionally uncertain data might be highly risky for some ecological applications (e.g., nature conservation).

Variable importance

For Workflow 2 models correctly, across all modeled scenarios, detected the aspect and elevation, which were used to generate virtual species, as the most influential variables. (only these variables were used to generate virtual species; Figure 3.13 and Figure A4 in supplementary material). Increasing sample size increased the estimated importance of aspect and in the *gbm* package (ver. 2.1.5, Friedman et al. 2000), represented presence-absence methods, and MaxEnt, implemented in the *dismo* package (ver. 1.1-4, Phillips et al. 2006; ver. 3.4.3 of maxent.jar file, Phillips et al. 2020), a presence-background method. Using both presence-absence and presence-background methods allowed us to assess whether they are equally affected by positional errors and by coarsening of the analysis grain. The GLM was run with a logit link function and a binomial distribution. The quadratic terms of the environmental variables were included based on the known normal distribution curves of the response function. For BRT, we used Bernoulli distribution, shrinkage (learning rate) of 0.01, tree complexity of 1 (i.e., without interaction terms), bag fraction (the proportion of data used when selecting optimal tree number) of 0.5, and the maximum number of trees of 5,000. MaxEnt was used with default settings (i.e., auto features, logistic output format) and 10,000 backgrounds points. The only exception was for models with an analysis grain of 500 x 500 m, where the number of grids / cells was not sufficient to sample 10,000 background points, so we ended up with a smaller number of background points (see Tab. A1). The same three environmental variables (CHM, DTM and TWI) that were used in the process of generating virtual species were also used to fit the models in seven analysis grains (see the previous section).

Model evaluation

We used several discrimination metrics to evaluate the performance of the models. First, we used the Sørensen index (SI), which has been recommended for the evaluation of experiments testing SDM methodologies using virtual species (Li and Guo 2013, Leroy et al. 2018). We also aimed to determine whether predictions using erroneous/altered data tend to over- or underpredict species occurrences. Thus, we calculated the overprediction and underprediction rates. Overprediction refers to the proportion of observed absences in the predicted presence area, and underprediction measures the proportion of actual presences that were not predicted by the model (Barbosa et al. 2013, Leroy et al. 2018). However, these metrics use only three components (true positives, false positives and false negatives) of the confusion matrix and neglect the prediction of true negatives (Leroy et al. 2018). Because we manipulated the input data (i.e., introduced the positional error and changed the analysis grain), we were concerned that this might also affect the true negatives. Therefore, we added the area under the receiver operating characteristic curve (AUC; Fielding and Bell 1997; despite recent criticisms of this metric, see for example Lobo et al. 2008, Jiménez Valverde 2012) and the true skill statistics (TSS; Allouche et al. 2006), which are commonly used to assess the discriminatory power of models.

In addition, we took advantage of the virtual species approach and compared differences between the predicted distribution inferred from the models and the true probability of occurrence of virtual species in geographical space. However, it has been stressed that metrics used for niche comparison are seriously affected by the inclusion of large number of cells where the species are absent (i.e., with low occurrence probabilities) and it has been recommended to remove such cell from the evaluation (Rödder and Engler 2011). Therefore, for this evaluation, we extract occurrence probability only for occurrence data, which were used in the models. We used Spearman's rank correlation to quantify the differences. See Supplementary materials Figure A2 for visual comparison between virtual species true distribution and predicted probability of all modelled scenarios. Note that this comparison was performed using the same resolution for all models' predictions (i.e., 500m).

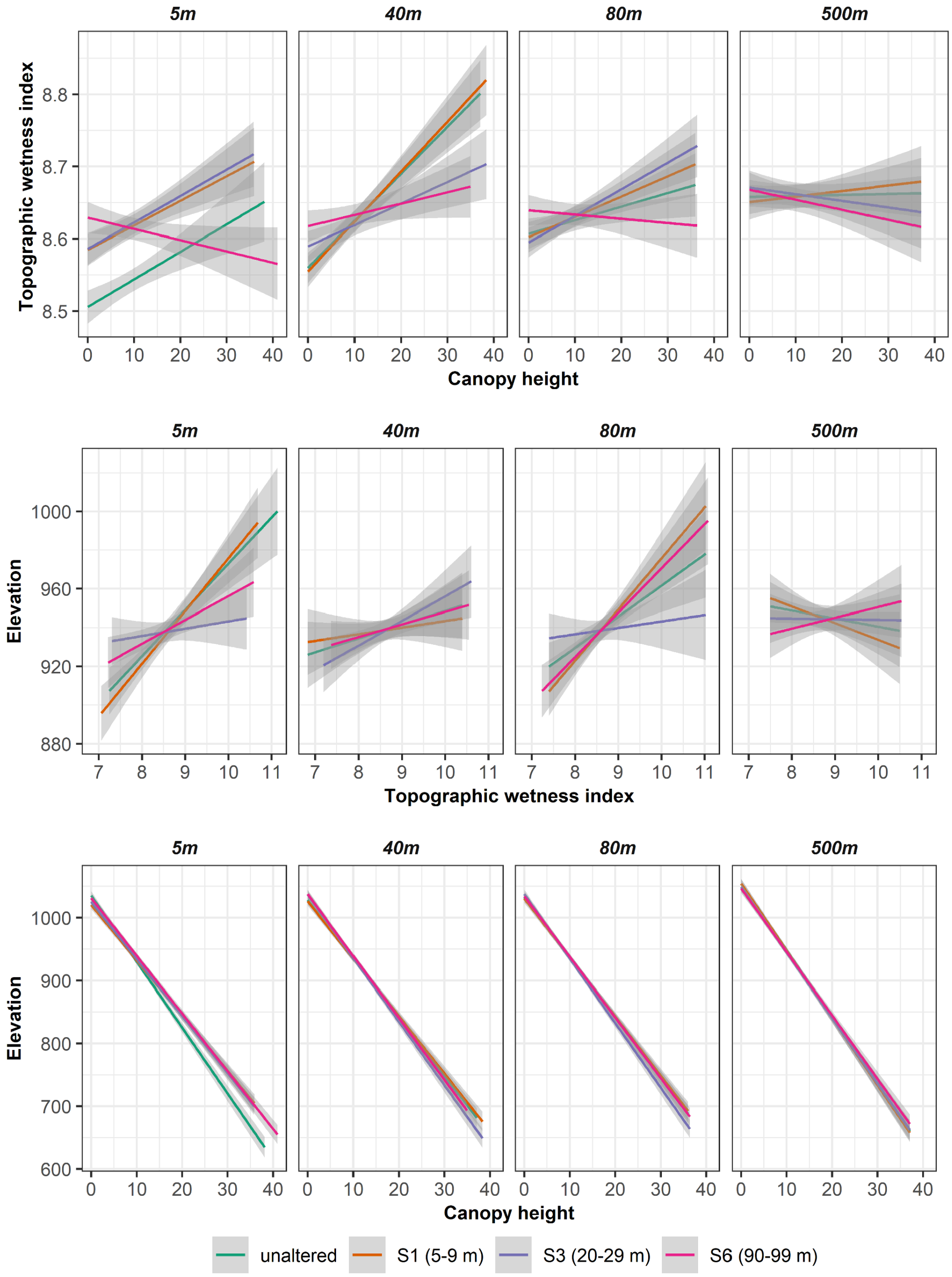


Figure 3.16: Comparison of changes in realized niche as a result of positional error in species occurrences and coarsening the analysis grain. Different colours show various levels of positional uncertainty while columns show different analysis grain. The line is obtained by linear regression and grey colour shows 95% confidence interval.

References

Agresti, A. (2003) Categorical data analysis. Vol. 482. John Wiley and Sons.

Ahmad Suhaimi, S. S., Blair, G. S., and Jarvis, S. G. (2021). Integrated species distribution models: A comparison of approaches under different data quality scenarios. Diversity and Distributions, 27(6), 1066-1075.

Alexander, J. M., Diez, J. M., and Levine, J. M. (2015). Novel competitors shape species' responses to climate change. Nature, 525(7570), 515-518.

Allouche, O., Tsoar, A., and Kadmon, R. (2006). Assessing the accuracy of species distribution models: prevalence, kappa and the true skill statistic (TSS). Journal of applied ecology, 43(6), 1223-1232.

Altizer, S., Ostfeld, R. S., Johnson, P. T., Kutz, S., and Harvell, C. D. (2013). Climate change and infectious diseases: from evidence to a predictive framework. Science, 341(6145), 514-519.

Amatulli, G., Domisch, S., Tuanmu, M. N., Parmentier, B., Ranipeta, A., Malczyk, J., and Jetz, W. (2018). A suite of global, cross-scale topographic variables for environmental and biodiversity modeling. Scientific data, 5(1), 1-15.

Ancillotto, L., Bosso, L., Smeraldo, S., Mori, E., Mazza, G., Herkt, M., ... and Russo, D. (2020). An African bat in Europe, Plecotus gaisleri: Biogeographic and ecological insights from molecular taxonomy and Species Distribution Models. Ecology and evolution, 10(12), 5785-5800.

Anderson, R. P., Lew, D., and Peterson, A. T. (2003). Evaluating predictive models of species' distributions: criteria for selecting optimal models. Ecological modelling, 162(3), 211-232.

Anderson, R. P., Araújo, M., Guisan, A., Lobo, J. M., Martínez-Meyer, E., Peterson, A. T., and Soberón, J. (2016). Final report of the task group on GBIF data fitness for use in distribution modelling. Global Biodiversity Information Facility.

Andren, H. (1994). Effects of habitat fragmentation on birds and mammals in landscapes with different proportions of suitable habitat: a review. Oikos, 355-366.

Araújo, M. B., and Guisan, A. (2006). Five (or so) challenges for species distribution modelling. Journal of biogeography, 33(10), 1677-1688.

Araújo, M. B., and New, M. (2007). Ensemble forecasting of species distributions. Trends in ecology and evolution, 22(1), 42-47.

Araújo, M. B., Anderson, R. P., Barbosa, A. M., Beale, C. M., Dormann, C. F., Early, R., ... and Rahbek, C. (2019). Standards for distribution models in biodiversity assessments. Science Advances, 5(1), eaat4858.

Arenas-Castro, S., Regos, A., Martins, I., Honrado, J., and Alonso, J. (2022). Effects of input data sources on species distribution model predictions across species with different distributional ranges. Journal of Biogeography, 49(7), 1299-1312.

Association, A.B. (2008). American Birding Association Checklist: Birds of the Continental United States and Canada.

Aubry, K. B., Raley, C. M., and McKelvey, K. S. (2017). The importance of data quality for generating reliable distribution models for rare, elusive, and cryptic species. PLoS One, 12(6), e0179152.

Cottenie, K. (2005). Integrating environmental and spatial processes in ecological community dynamics. Ecology letters, 8(11), 1175-1182.

Culina, A., Baglioni, M., Crowther, T. W., Visser, M. E., Woutersen-Windhouwer, S., and Manghi, P. (2018). Navigating the unfolding open data landscape in ecology and evolution. Nature ecology and evolution, 2(3), 420-426.

Curd, A., Chevalier, M., Vasquez, M., Boyé, A., Firth, L. B., Marzloff, M. P., ... and Dubois, S. F. (2023). Applying landscape metrics to species distribution model predictions to characterize internal range structure and associated changes. Global Change Biology, 29(3), 631-647.

D'Amen, M., and Azzurro, E. (2020). Lessepsian fish invasion in Mediterranean marine protected areas: a risk assessment under climate change scenarios. ICES Journal of Marine Science, 77(1), 388-397.

Davies, A. B., and Asner, G. P. (2014). Advances in animal ecology from 3D-LiDAR ecosystem mapping. Trends in ecology and evolution, 29(12), 681-691.

de Brooke, M. (2000). Why museums matter. Trends in Ecology and Evolution, 15(4), 136-137.

De Marco, P., and Nóbrega, C. C. (2018). Evaluating collinearity effects on species distribution models: An approach based on virtual species simulation. PloS one, 13(9), e0202403.

de Vries, J. P. R., Koma, Z., WallisDeVries, M. F., and Kissling, W. D. (2021). Identifying fine-scale habitat preferences of threatened butterflies using airborne laser scanning. Diversity and Distributions.

Didan, K. (2015). MOD13A2 MODIS/Terra Vegetation Indices 16-Day L3 Global 1km SIN Grid V006 [Data set]. NASA EOSDIS Land Processes DAAC.

Domisch, S., Jähnig, S. C., Simaika, J. P., Kuemmerlen, M., and Stoll, S. (2015). Application of species distribution models in stream ecosystems: the challenges of spatial and temporal scale, environmental predictors and species occurrence data. Fundamental and Applied Limnology, 45-61.

Dormann, F. C., M. McPherson, J., B. Araújo, M., Bivand, R., Bolliger, J., Carl, G., ... and Wilson, R. (2007). Methods to account for spatial autocorrelation in the analysis of species distributional data: a review. Ecography, 30(5), 609-628.

Dormann, C. F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carré, G., ... and Lautenbach, S. (2013). Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. Ecography, 36(1), 27-46.

Dormann, C. F., Bobrowski, M., Dehling, D. M., Harris, D. J., Hartig, F., Lischke, H., ... and Kraan, C. (2018). Biotic interactions in species distribution modelling: 10 questions to guide interpretation and avoid false conclusions. Global ecology and biogeography, 27(9), 1004-1016.

Drescher, M., Perera, A. H., Johnson, C. J., Buse, L. J., Drew, C. A., and Burgman, M. A. (2013). Toward rigorous use of expert knowledge in ecological research. Ecosphere 4: 1–26.

Drew, C. A., and Perera, A. H. (2010). Expert knowledge as a basis for landscape ecological predictive models. In Predictive species and habitat modeling in landscape ecology: Concepts and applications (pp. 229-248). New York, NY: Springer New York.

Mertes, K., and Jetz, W. (2018). Disentangling scale dependencies in species environmental niches and distributions. Ecography, 41(10), 1604-1615.

Mertes, K., Jarzyna, M. A., and Jetz, W. (2020). Hierarchical multi-grain models improve descriptions of species' environmental associations, distribution, and abundance. Ecological Applications, 30(6), e02117.

Meyer, C. B. (2007). Does scale matter in predicting species distributions? Case study with the marbled murrelet. Ecological Applications, 17(5), 1474-1483.

Meynard, C. N., and Kaplan, D. M. (2013). Using virtual species to study species distributions and model performance. Journal of Biogeography, 40(1), 1-8.

Meynard, C. N., Leroy, B., and Kaplan, D. M. (2019). Testing methods in species distribution modelling using virtual species: what have we learnt and what are we missing?. Ecography, 42(12), 2021-2036.

Mi, C., Huettmann, F., Guo, Y., Han, X., and Wen, L. (2017). Why choose Random Forest to predict rare species distribution with few samples in large undersampled areas? Three Asian crane species models provide supporting evidence. PeerJ, 5, e2849.

Michener, W. K. (2015). Ecological data sharing. Ecological informatics, 29, 33-44.

Miguet, P., Jackson, H. B., Jackson, N. D., Martin, A. E., and Fahrig, L. (2016). What determines the spatial extent of landscape effects on species?. Landscape ecology, 31, 1177-1194.

Milanesi, P., Herrando, S., Pla, M., Villero, D., and Keller, V. (2017). Towards continental bird distribution models: environmental variables for the second European breeding bird atlas and identification of priorities for further surveys. Vogelwelt, 137, 53-60.

Miller, J. (2010). Species distribution modeling. Geography Compass, 4(6), 490-5.

Miller, J. A. (2014). Virtual species distribution models: Using simulated data to evaluate aspects of model performance. Progress in Physical Geography, 38(1), 117-128.

Misiuk, B., Lecours, V., and Bell, T. (2018). A multiscale approach to mapping seabed sediments. PLoS One, 13(2), e0193647.

Mitchell, P. J., Monk, J., and Laurenson, L. (2017). Sensitivity of fine-scale species distribution models to locational uncertainty in occurrence data across multiple sample sizes. Methods in Ecology and Evolution, 8(1), 12-21.

Mod, H. K., Scherrer, D., Luoto, M., and Guisan, A. (2016). What we use is not what we know: environmental predictors in plant distribution models. Journal of Vegetation Science, 27(6), 1308-1322.

Mohammadi, A., Almasieh, K., Nayeri, D., Adibi, M. A., and Wan, H. Y. (2022). Comparison of habitat suitability and connectivity modelling for three carnivores of conservation concern in an Iranian montane landscape. Landscape Ecology, 37(2), 411-430.

Monk, J. (2014). How long should we ignore imperfect detection of species in the marine environment when modelling their distribution?. Fish and Fisheries, 15(2), 352-358.

Montgomery, R. A., Roloff, G. J., and Hoef, J. M. V. (2011). Implications of ignoring telemetry error on inference in wildlife resource use models. The Journal of Wildlife Management, 75(3), 702-708.

Title, P. O., and Bemmels, J. B. (2018). ENVIREM: an expanded set of bioclimatic and topographic variables increases flexibility and improves performance of ecological niche modeling. Ecography, 41(2), 291-307.

Trainor, A. M., Schmitz, O. J., Ivan, J. S., and Shenk, T. M. (2014). Enhancing species distribution modeling by characterizing predator–prey interactions. Ecological Applications, 24(1), 204-216.

Tuanmu, M. N., and Jetz, W. (2015). A global, remote sensing-based characterization of terrestrial habitat heterogeneity for biodiversity and ecosystem modelling. Global Ecology and Biogeography, 24(11), 1329-1339.

Tulowiecki, S. J., Larsen, C. P., and Wang, Y. C. (2015). Effects of positional error on modeling species distributions: a perspective using presettlement land survey records. Plant Ecology, 216, 67-85.

Tulowiecki, S. J. (2020). Modeling the historical distribution of American chestnut (Castanea dentata) for potential restoration in western New York State, US. Forest Ecology and Management, 462, 118003.

Urban, M. C., Bocedi, G., Hendry, A. P., Mihoub, J. B., Pe'er, G., Singer, A., ... and Travis, J. M. (2016). Improving the forecast for biodiversity under climate change. Science, 353(6304), aad8466.

Václavík, T., Kupfer, J. A., and Meentemeyer, R. K. (2012). Accounting for multi-scale spatial autocorrelation improves performance of invasive species distribution modelling (iSDM). Journal of Biogeography, 39(1), 42-55.

Valavi, R., Guillera-Arroita, G., Lahoz-Monfort, J. J., and Elith, J. (2022). Predictive performance of presence-only species distribution models: a benchmark study with reproducible code. Ecological Monographs, 92(1), e01486.

Valavi, R., Elith, J., Lahoz-Monfort, J. J., and Guillera-Arroita, G. (2023). Flexible species distribution modelling methods perform well on spatially separated testing data. Global Ecology and Biogeography, 32(3), 369-383.

Van Moorter, B., Kivimäki, I., Panzacchi, M., Saura, S., Brandão Niebuhr, B., Strand, O., and Saerens, M. (2023). Habitat functionality: Integrating environmental and geographic space in niche modeling for conservation planning. Ecology, e4105.

van Proosdij, A. S., Sosef, M. S., Wieringa, J. J., and Raes, N. (2016). Minimum required number of specimen records to develop accurate species distribution models. Ecography, 39(6), 542-552.

VanDerWal, J., Shoo, L. P., Graham, C., and Williams, S. E. (2009). Selecting pseudo-absence data for presence-only distribution modeling: how far should you stray from what you know?. Ecological modelling, 220(4), 589-594.

Varela, S., Anderson, R. P., García-Valdés, R., and Fernández-González, F. (2014). Environmental filters reduce the effects of sampling bias and improve predictions of ecological niche models. Ecography, 37(11), 1084-1091.

Velásquez-Tibatá, J., Graham, C. H., and Munch, S. B. (2016). Using measurement error models to account for georeferencing error in species distribution models. Ecography, 39(3), 305-316.

Author's Academic CV

Personal

Name: Lukáš Gábor
Date of birth: January 1^{st}, 1991, Czech Republic
E-mail: gabor.lucas@gmail.com
ORCID: https://orcid.org/0000-0001-6137-0994
Google Scholar: https://scholar.google.cz/citations?user=pLQXY5wAAAAJ&hl=cs

Affiliations

2021 – 2023
Yale University
Jetz Lab
Department of Ecology and Evolutionary Biology
Yale University

2017 – 2024
Department of Spatial Sciences,
Faculty of Environmental Sciences,
Czech University of Life Sciences Prague

Education

2020 – present
Department of Spatial Sciences,
Faculty of Environmental Sciences,
Czech University of Life Sciences Prague
PhD studies in Spatial Sciences
Thesis topic: Beyond Assumptions: Unraveling Data Limitations in Predicting Species Distribution

2016 – 2020
Department of Applied Geoinformatics and Spatial Planning,
Faculty of Environmental Sciences,
Czech University of Life Sciences Prague
PhD studies in Applied and Landscape Ecology
Dissertation topic: The Quality of Spatial Data and its Effect on Species Distribution Models

2014 – 2016
Faculty of Environmental Sciences,
Czech University of Life Sciences Prague
Master's degree in Applied Ecology
Thesis topic: Do Environmental Filters Improve Predictions of Species Distribution Models?

Recognizing ChatGPT's Contribution

I would like to acknowledge that the AI tool ChatGPT was utilized as an assistance in organizing my thoughts and refining the text for the introduction and discussion sections of this dissertation. Throughout the writing process, I was mindful of the limitations and potential biases of the AI tool. I approached it critically, carefully evaluating its suggestions and ensuring they aligned with academic standards. I was responsible for critically reviewing and refining the content, ensuring its accuracy, coherence, and adherence to scholarly conventions. Therefore, the intellectual integrity of the work lies with me as the author, with the AI tool serving as a support in the writing process.

Presented insight, ideas, and thoughts are my own and result from extensive research, critical thinking, and scholarly engagement in ecological predictive modeling. More importantly, the references used in this text were independently sourced through reputable academic platforms such as Google Scholar or Web of Science. The proper citation of these references throughout the dissertation demonstrates that they were not generated by the AI tool. While writing and completing my dissertation (summer 2024), there were no explicit regulations or guidelines at the Czech University of Life Sciences or at the Faculty of Environmental Sciences that prohibited using AI tools like ChatGPT for dissertation writing.

Meet the Mind Behind the Words

The author (see Figure 9.1), an data science and ESG expert, has always been fascinated by the profound influence of data on our lives and the world around us. This early fascination sparked a keen interest in exploring various types of data and their potential applications. Consequently, when it came time to choose a topic for Ph.D. studies, he made a deliberate decision to delve into the realm of spatial data quality and types within the field of predictive ecology.

It soon became clear that while numerous studies focused on modeling methodologies and theoretical frameworks, there was a need for deeper exploration into the critical aspects of spatial data quality and the wide range of spatial data types applicable in predicting species distribution.

The presented work represents the culmination of years of academic pursuit and a profound passion for unraveling the intricacies of ecological predictive modeling. The research within these pages strives to shed light on the influence of spatial data quality, challenge commonly held assumptions, and offer insights that can enhance the accuracy, reliability, and robustness of ecological predictive models. May the ideas presented within these pages ignite curiosity, foster collaboration, and deepen appreciation for the marvelous intricacies of predictive modeling in ecology.

Figure 9.1: Author, Fall 2023

„The dilemma is, as it is often said, correlation does not imply causation. The discovery of a predictive relationship between A and B does not mean one causes the other, not even indirectly. No way, nohow.“

— Eric Siegel

„You can have all of the fancy tools, but if data quality is not good, you're nowhere.“

— Veda Bawo

„Torture the data, and it will confess to anything.“

— Ronald Coase

„How do we start to regulate the mathematical models that run more and more of our lives? I would suggest that the process begin with the modelers themselves. Like doctors, data scientists should pledge a Hippocratic Oath, one that focuses on the possible misuses and misinterpretations of their models.“

—Cathy O'Neil

„With too little data, you won't be able to make any conclusions that you trust. With loads of data you will find relationships that aren't real. Big data isn't about bits, it's about talent.“

— Douglas Merrill

„Big data is like teenage sex; everyone talks about it, nobody really knows how to do it, everyone thinks everyone else is doing it, so everyone claims they are doing it.“

— Dan Ariely

CZU Czech University
of Life Sciences Prague

Faculty of Environmental Sciences
Department of Applied Geoinformatics and Spatial Planning
2024
Dissertation Thessis