

Current themes and recent advances in modelling species occurrences

Graeme S Cumming

Address: Percy FitzPatrick Institute, DST/NRF Centre of Excellence, University of Cape Town, Rondebosch, Cape Town 7701, South Africa

Email: graeme.cumming@uct.ac.za

F1000 Biology Reports 2009, 1:94 (doi:10.3410/B1-94)

The electronic version of this article is the complete one and can be found at: <http://F1000.com/Reports/Biology/content/1/94>

Abstract

Recent years have seen a huge expansion in the range of methods and approaches that are being used to predict species occurrences. This expansion has been accompanied by many improvements in statistical methods, including more accurate ways of comparing models, better null models, methods to cope with autocorrelation, and greater awareness of the importance of scale and prevalence. However, the field still suffers from problems with incorporating temporal variation, overfitted models and poor out-of-sample prediction, confusion between explanation and prediction, simplistic assumptions, and a focus on pattern over process. The greatest advances in recent years have come from integrative studies that have linked species occurrence models with other themes and topics in ecology, such as island biogeography, climate change, disease geography, and invasive species.

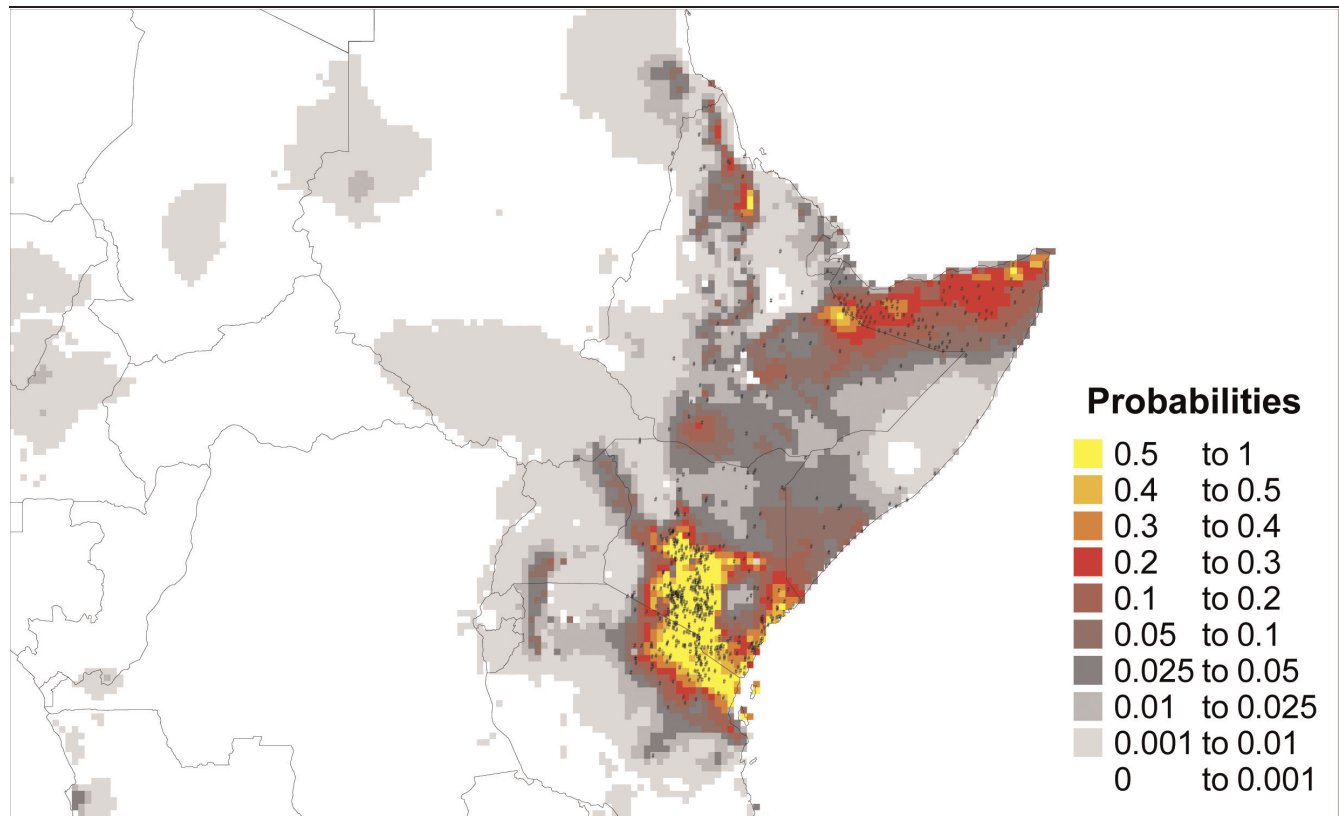
Introduction and context

Species occurrence models are used to develop spatially explicit interpolations from known species occurrences to unsampled areas. They are applied in ecology in a wide variety of ways that include (but are not limited to) the basic estimation of where a species can be expected to occur, explaining how species ranges may have changed in the past or predicting how they may do so in the future, understanding niches and the limits on species ranges, quantifying community-level patterns in biodiversity, and exploring alternative scenarios about the impacts of environmental change.

Species occurrence models (e.g., Figure 1) relate changes in a spatially explicit response variable (Y , the species occurrence, stated as either the number of individuals in a grid cell or species presence/absence) to changes in a spatially explicit set of predictor variables (X , which may be categorical or continuous and often include collinear variables such as temperature, rainfall, vegetation, and land cover). X variables are related to the Y variable via a link function, which defines the way in which the predictors relate to the response variable. Although link

functions are formally components of generalised linear models (as for identity, logit, or poisson links, for example), most non-linear models also require the selection of a link function (e.g., discriminant function analysis, fuzzy classifiers, or trainable algorithms such as neural networks).

The basic concerns of developing and applying species occurrence models were nicely laid out in the classic paper by Fielding and Bell [1]. A number of more recent papers [2-4] contribute in-depth summaries of important challenges, most of which are still relevant. The majority of current activity in the field can be classified into three interrelated themes: (a) development of new link functions and new statistical approaches; (b) exploration and resolution of issues relating to model fit and model comparisons for existing methods, including problems of scale, autocorrelation, and sampling; and (c) better integration with other themes in ecology, such as island biogeography, invasive species, disease ecology, and climate change impacts. I will expand on each of these three themes in a little more detail.

Figure 1. Example of a predictive species occurrence map

This map depicts the known distribution of a brown tick (Acari: Ixodidae; *Rhipicephalus pulchellus*) in East Africa and a species range map derived using rainfall and temperature data. The black dots are collection localities at which the tick was found, and shading indicates a probability of occurrence at a resolution of a quarter of a degree. Further methodological details can be found in the papers listed in [47].

The development of new statistical approaches to distribution modelling seems to have become something of a spin-off industry, and the range of approaches now on offer is bewildering and (arguably) unnecessary. Nonetheless, there have been a few genuine advances in this area in recent years, particularly in developing approaches to non-linear link functions (e.g., [5,6]). The tradeoff in many cases is between model interpretability and model accuracy.

Statistical questions remain an important research area in species occurrence modelling [7]. In addition to their ecological relevance, techniques for quantifying model fit are important for contrasting the strengths and weaknesses of alternative methods and for resolving questions about the influence of scale and sampling on model output. Under the influence of Fielding and Bell [1], there has been a gradual shift away from quoting kappa statistics or percentages of different errors and toward the use of ROC (Receiver Operating Characteristic) plots. Information criteria (particularly Akaike's Information

Criterion and Bayesian Information Criterion) are also widely used. There have been few great leaps forward in this area in recent years, but a number of solid papers that are gradually bringing clarity to the field have been published (e.g., [8,9]). There have been several clear demonstrations that simple statistical tricks, such as increasing the extent of the sampling area or decreasing the grain (resolution) of analysis while keeping the number of positive records constant, can increase a model's significance [10-12] (although the grain of available data for the analysis of some taxa may genuinely be critical [13]). Since the power of any frequentist statistical test is contingent on sampling frequency and sample size, recent criticisms of the AUC (Area Under the Curve) (e.g., [14]) do not, in my opinion, address the fundamental problem, which is the need for a multi-scale rather than a single-scale approach to spatial analysis [15].

There has been relatively little use of model averaging and Occam's window (a procedure in which a subset of well-fitting models is used to obtain an average solution)

[16-19] as ways of obtaining more reliable predictions, although some recent studies have explored the development of models that attempt to take both spatial autocorrelation and imperfect survey data into account (e.g., [20]) and consensus or ensemble methods are starting to be more widely used [21].

Species distribution models are increasingly being integrated with other themes in ecology, such as the influence of dispersal on species occurrences [22], the relevance of life history characteristics and fitness [23], the potential impacts of invasive species [24], and both forecasts and hindcasts about the impacts of climate change on species ranges (e.g., [25-27]) and community-level patterns [25,26]. A particularly fast-growing application is the development of models that are based on predictor variables (e.g., climate and land cover) that can be projected into the future under different scenarios to assist in the formulation of proactive strategies for problems such as changes in patterns of vector-borne and infectious diseases (e.g., [27-29]). The increasing availability of high-quality remotely sensed data sets and detailed atlas and survey records is also contributing to the development of more accurate occurrence predictions, though not inevitably so [27,28].

Major recent advances

In recent years, there has been a huge amount of research on predicting species occurrences. It is impossible to do full justice to this buzz of activity in such a short review; nonetheless, I will mention a few selected statistical and ecological highlights.

In the statistical arena, there has been considerable recent progress in dealing with autocorrelation [29-32] and in ways of thinking more effectively about non-linearities in species-habitat relationships, particularly in regard to the quantification of dispersal limitation [33,34] and environmental thresholds [35]. Useful insights into the problem of model transferability are also accumulating [36].

As methods for predicting species occurrences have improved and become more widely accepted, researchers have been able to turn their attention toward a range of interesting applications. Perhaps the most important advances in recent years have come from applications of occurrence models in fields like evolutionary biology [37], climate change, invasive species [38], the study of patterns of species richness [39,40], and disease geography [41]. Many of these studies, in turn, have offered further methodological and theoretical insights. The scale dependencies identified by Menke *et al.* [38], for example, constitute one of the most interesting of recent

results and should go well beyond their relevance for statistics.

Future directions

The field appears to be progressing in a number of interrelated ways. Some important methodological issues are still unresolved [42]: the development of ways to correct for the influences of prevalence and scale on model fit, rigorous resolution of the problems created by autocorrelation, and better integration of species distribution models with other approaches to the analysis of spatial pattern in ecology, such as metapopulation and metacommunity models [43].

The development of more effective ways of incorporating temporal variation in species occurrences into distribution models remains an important challenge, particularly in regard to climate change. Unbalanced sampling regimes create a constant danger that current models interpret temporal variation as spatial variation, or *vice versa*, and in this way may provide substantially inaccurate predictions. For example, I am not aware of any studies of species occurrences that have dealt with both spatial and temporal autocorrelation in the underlying data sets.

There have been some interesting recent developments relating to the conceptual foundations of species occurrence models [44,45], and some important theoretical challenges remain in thinking through the different assumptions that underlie occurrence models. One approach that has been little explored (but see, e.g., [46]) is to contrast statistical occurrence models with mechanistic or process-based predictions. As I have argued elsewhere [47], there is a strong need to develop and use cross-scale comparisons (and data from different levels of organization) to understand species occurrences. Perhaps the most fundamental problem in the field is that too many occurrence models are correlative desktop exercises that are light on ecology; statistically accurate but mechanism-free models do not necessarily mean accurate prediction [48,49] and frequently result in poor transferability [50].

Abbreviations

AUC, area under the curve; ROC, receiver operating characteristic.

Competing interests

The author declares that he has no competing interests.

Acknowledgements

I am grateful to four tough but anonymous reviewers for their useful comments.

References

1. Fielding AH, Bell JF: **A review of methods for the assessment of prediction errors in conservation presence/absence models.** *Environ Conserv* 1997, **24**:38-49.
 2. Austin MP: **Spatial prediction of species distribution: an interface between ecological theory and statistical modelling.** *Ecol Modell* 2002, **157**:101-18.
 3. Thuiller W, Albert C, Araújo MB, Berry PM, Cabeza M, Guisan A, Hickler T, Midgely GF, Paterson J, Schurr FM, Sykes MT, Zimmermann NE: **Predicting global change impacts on plant species' distributions: future challenges.** *Perspect Plant Ecol Evol Syst* 2008, **9**:137-52.
 4. Araujo MB, Guisan A: **Five (or so) challenges for species distribution modelling.** *J Biogeogr* 2006, **33**:1677-88.
 5. Williams JN, Seo CW, Thorne J, Nelson JK, Erwin S, O'Brien JM, Schwartz MW: **Using species distribution models to predict new occurrences for rare plants.** *Divers Distrib* 2009, **15**:565-76.
 6. Elith J, Graham CH, Anderson RP, Dudik M, Ferrier S, Guisan A, Hijmans RJ, Huettmann F, Leathwick JR, Lehmann A, Li J, Lohmann LG, Loiselle BA, Manion G, Moritz C, Nakamura M, Nakazawa Y, Overton JM, Peterson AT, Phillips SJ, Richardson KS, Scachetti-Pereira R, Schapire RE, Soberón J, Williams S, Wisz MS, Zimmermann NE: **Novel methods improve prediction of species' distributions from occurrence data.** *Ecography* 2006, **29**:129-51.
 7. Elith J, Graham CH: **Do they? How do they? WHY do they differ? On finding reasons for differing performances of species distribution models.** *Ecography* 2009, **32**:66-77.
 8. Platts PJ, McClean CJ, Lovett JC, Marchant R: **Predicting tree distributions in an East African biodiversity hotspot: model selection, data bias and envelope uncertainty.** *Ecol Modell* 2008, **218**:121-34.
 9. Jetz W, Sekercioglu CH, Watson JEM: **Ecological correlates and conservation implications of overestimating species geographic ranges.** *Conserv Biol* 2008, **22**:110-9.
- F1000 Factor 3.0 Recommended
 Evaluated by George Malanson 18 Mar 2008
10. Cumming GS: **Using between-model comparisons to fine-tune linear models of species ranges.** *J Biogeogr* 2000, **27**:441-55.
 11. Guisan A, Graham CH, Elith J, Huettmann F, Distri NS: **Sensitivity of predictive species distribution models to change in grain size.** *Divers Distrib* 2007, **13**:332-40.
 12. McPherson JM, Jetz W, Rogers DJ: **Using coarse-grained occurrence data to predict species distributions at finer spatial resolutions-possibilities and limitations.** *Ecol Modell* 2006, **192**:499-522.
 13. Trivedi MR, Berry PM, Morecroft MD, Dawson TP: **Spatial scale affects bioclimate model projections of climate change impacts on mountain plants.** *Glob Change Biol* 2008, **14**:1089-103.
 14. Peterson AT, Papes M, Soberon J: **Rethinking receiver operating characteristic analysis applications in ecological niche modeling.** *Ecol Modell* 2008, **213**:63-72.
 15. Wu J: **Effects of changing scale on landscape pattern analysis: scaling relations.** *Landscape Ecol* 2004, **19**:125-38.
 16. Raftery AE, Madigan D, Hoeting JA: **Bayesian model averaging for linear regression models.** *J Am Stat Assoc* 1997, **92**:179-91.
 17. Burnham KP, Anderson DR: *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach.* 2nd edition. New York, NY: Springer-Verlag; 2002.
 18. Hoeting JA, Raftery AE, Madigan D: **Bayesian variable and transformation selection in linear regression.** *J Comput Graph Stat* 2002, **11**:485-507.
 19. Hoeting JA, Madigan D, Raftery AE, Volinsky CT: **Bayesian model averaging: a tutorial.** *Stat Sci* 1999, **14**:382-401.
 20. Royle JA, Kery M, Gautier R, Schmid H: **Hierarchical spatial models of abundance and occurrence from imperfect survey data.** *Ecol Monogr* 2007, **77**:465-81.
 21. Roura-Pascual N, Brotons L, Peterson AT, Thuiller W: **Consensual predictions of potential distributional areas for invasive species: a case study of Argentine ants in the Iberian Peninsula.** *Biol Invasions* 2009, **11**:1017-31.
 22. Engler R, Randin CF, Vittoz P, Czaka T, Beniston M, Zimmermann NE, Guisan A: **Predicting future distributions of mountain plants under climate change: does dispersal capacity matter?** *Ecography* 2009, **32**:34-45.
 23. Betts MG, Rodenhouse NL, Sillett TS, Doran PJ, Holmes RT: **Dynamic occupancy models reveal within-breeding season movement up a habitat quality gradient by a migratory songbird.** *Ecography* 2008, **31**:592-600.
 24. Peterson AT: **Predicting the geography of species' invasions via ecological niche modeling.** *Q Rev Biol* 2003, **78**:419-33.
 25. Algar AC, Kharouba HM, Young ER, Kerr JT: **Predicting the future of species diversity: macroecological theory, climate change, and direct tests of alternative forecasting methods.** *Ecography* 2009, **32**:22-33.
 26. Elmendorf SC, Moore KA: **Use of Community-Composition Data to Predict the Fecundity and Abundance of Species.** *Conserv Biol* 2008, **22**:1523-32.
 27. Lozier JD, Aniello P, Hickerson MJ: **Predicting the distribution of Sasquatch in western North America: anything goes with ecological niche modelling.** *J Biogeogr* 2009, **36**:1623-7.
 28. Dormann CF, Purschke O, Marquez JRG, Lautenbach S, Schroder B: **Components of uncertainty in species distribution analysis: a case study of the great grey shrike.** *Ecology* 2008, **89**:3371-86.
 29. Dormann CF, McPherson JM, Araujo MB, Bivand R, Bolliger J, Carl G, Davies RG, Hirzel A, Jetz W, Kissling WD, Kühn I, Ohlemüller R, Peres-Neto PR, Reineking B, Schröder B, Schurr FM, Wilson R: **Methods to account for spatial autocorrelation in the analysis of species distributional data: a review.** *Ecography* 2007, **30**:609-28.
 30. Dormann CF: **Effects of incorporating spatial autocorrelation into the analysis of species distribution data.** *Glob Ecol Biogeogr* 2007, **16**:129-38.
 31. Hoeting JA: **The importance of accounting for spatial and temporal correlation in analyses of ecological data.** *Ecol Appl* 2009, **19**:574-7.
 32. Betts MG, Ganio LM, Huso MMP, Som NA, Huettmann F, Bowman J, Wintle BA: **Comment on "Methods to account for spatial autocorrelation in the analysis of species distributional data: a review".** *Ecography* 2009, **32**:374-8.
 33. Engler R, Guisan A: **MIGCLIM: Predicting plant distribution and dispersal in a changing climate.** *Divers Distrib* 2009, **15**:590-601.
 34. Munguia M, Peterson AT, Sanchez-Cordero V: **Dispersal limitation and geographical distributions of mammal species.** *J Biogeogr* 2008, **35**:1879-87.
 35. Betts MG, Forbes GJ, Diamond AW: **Thresholds in songbird occurrence in relation to landscape structure.** *Conserv Biol* 2007, **21**:1046-58.
 36. Sundblad G, Harma M, Lappalainen A, Urho L, Bergstrom U: **Transferability of predictive fish distribution models in two coastal systems.** *Estuar Coast Shelf Sci* 2009, **83**:90-6.
 37. Kozak KH, Graham CH, Wiens JJ: **Integrating GIS-based environmental data into evolutionary biology.** *Trends Ecol Evol* 2008, **23**:141-8.
 38. Menke SB, Holway DA, Fisher RN, Jetz W: **Characterizing and predicting species distributions across environments and scales: Argentine ant occurrences in the eye of the beholder.** *Glob Ecol Biogeogr* 2009, **18**:50-63.
 39. Buckley LB, Jetz W: **Environmental and historical constraints on global patterns of amphibian richness.** *Proc Biol Sci* 2007, **274**:1167-73.
 40. Broennimann O, Thuiller W, Hughes G, Midgley GF, Alkemade JMR, Guisan A: **Do geographic distribution, niche property and life form explain plants' vulnerability to global change?** *Glob Change Biol* 2006, **12**:1079-93.

41. Peterson AT, Williams RAJ: **Risk Mapping of Highly Pathogenic Avian Influenza Distribution and Spread.** *Ecol Soc* 2008, **13**:15.
42. Jimenez-Valverde A, Lobo JM, Hortal J: **Not as good as they seem: the importance of concepts in species distribution modelling.** *Divers Distrib* 2008, **14**:885-90.
43. Zanini F, Pellet J, Schmidt BR: **The transferability of distribution models across regions: an amphibian case study.** *Divers Distrib* 2009, **15**:469-80.
44. Soberon J: **Grinnellian and Eltonian niches and geographic distributions of species.** *Ecol Lett* 2007, **10**:1115-23.
45. Chase JM, Leibold MA: *Ecological Niches: Linking Classical and Contemporary Approaches.* Chicago, IL: University of Chicago Press; 2003.
46. Kearney M, Porter W: **Mechanistic niche modelling: combining physiological and spatial data to predict species ranges.** *Ecol Lett* 2009, **12**:334-50.
47. Cumming GS: **Global biodiversity scenarios and landscape ecology.** *Landsc Ecol* 2007, **22**:671-85.
48. McPherson JM, Jetz W: **Effects of species' ecology on the accuracy of distribution models.** *Ecography* 2007, **30**:135-51.
49. Vallecillo S, Brotons L, Thuiller W: **Dangers of predicting bird species distributions in response to land-cover changes.** *Ecol Appl* 2009, **19**:538-49.
50. Duncan RP, Cassey P, Blackburn TM: **Do climate envelope models transfer? A manipulative test using dung beetle introductions.** *Proc Biol Sci* 2009, **276**:1449-57.