Applications of Scale-Varying Functional Data Analysis to Biological Species Distribution Modelling

Allan Roberts, robera64@mcmaster.ca

Department of Biology, McMaster University, Hamilton, Ontario, Canada

INTRODUCTION

Types of functional data that are relevant to ecological modelling include observations of time-varying or distance-varying phenomena; for example a habitat characteristic may be observed as a function of time (e.g. temperature observed as a function of day of the year), or as a function of spatial scale (e.g. a landscape attribute observed as a function of distance). Such situations are often handled by identifying a single spatial or temporal scale at which to represent an explanatory variable; however, an alternative is to use a functional data approach that allows habitat characteristics to be regarded as functions of scale, rather than as values observed at a single particular scale. For example, Sims et al. (2007) uses time-varying functional models; a distance-varying method is used by Cornulier et al. (2015). One way to model datasets that include functional observations is with generalized additive models (GAMs). A GAM is a type of statistical model that allows the concurrent estimation of response curves for multiple environmental predictor variables and is a tool that is commonly used for species distribution modelling. R syntax (R Core Team, 2024) for including functional terms in a GAM is described in Wood (2017); however, such scale-varying functional terms are currently used rather infrequently in ecological modelling.

SCALE-VARYING RESOURCE USE

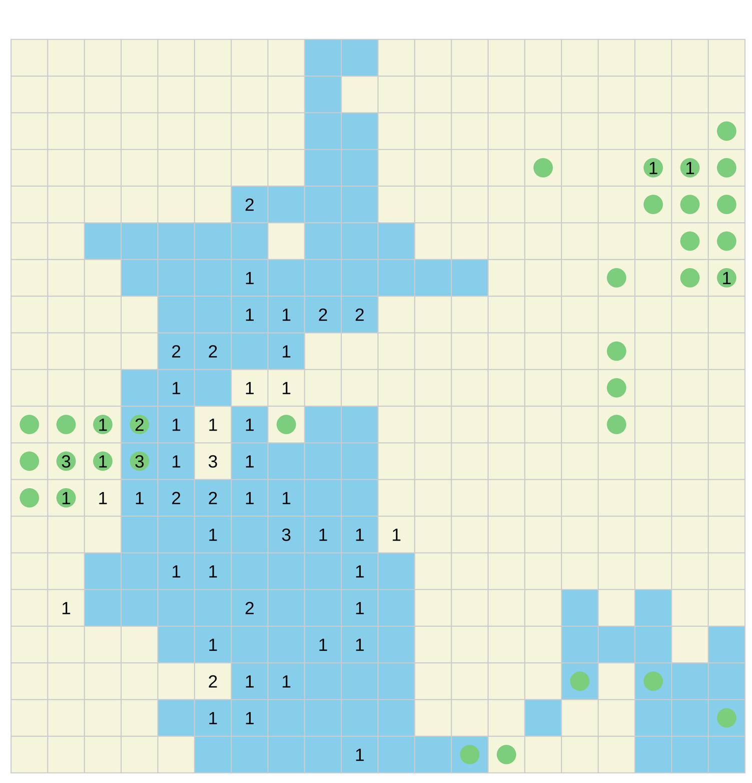

Fig. 1. This hypothetical landscape was generated with Gaussian random fields, using the R package `fields` (Nychka et al. 2021). This plot, and some others shown on this poster were made with the R package `ggplot2` (Wickham 2016). Numbers show species counts simulated from a Poisson distribution with a mean that is a function of the landscape.

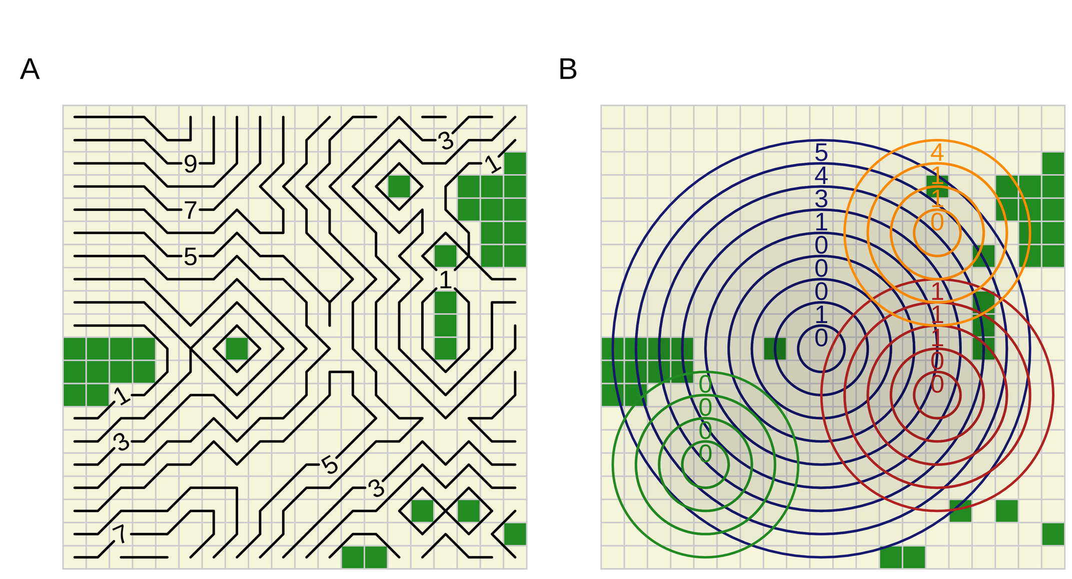

Fig. 2. Contour lines in Panel A show an explanatory variable that is the least distance to a particular resource type. Panel B illustrates the concept of a distance-varying *function* being observed at each location; such a distance-varying method is used by Cornulier et al. (2015) to model the distribution of Montagu's harrier *Circus pygargus*.

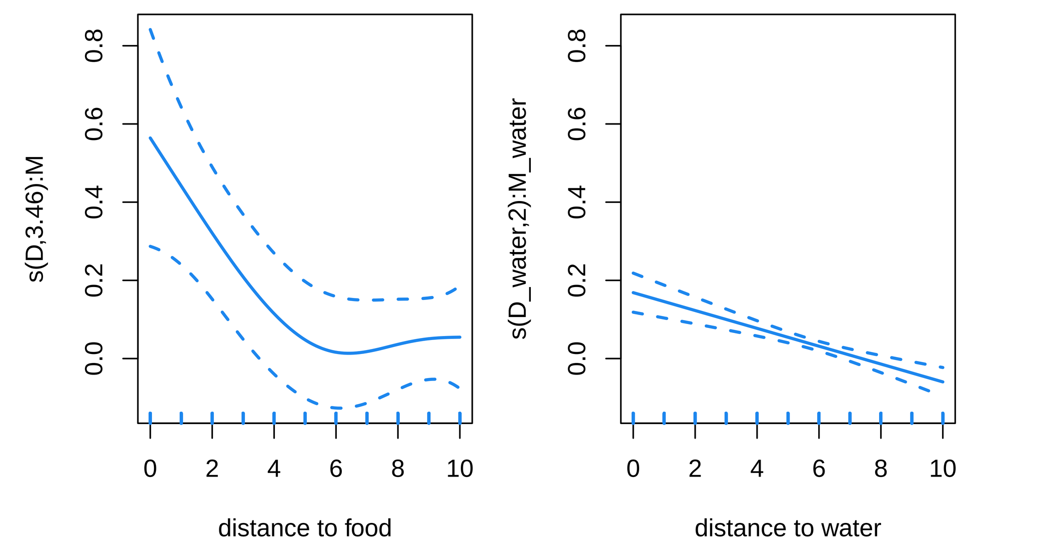

Fig. 3. Smoothers for the functional terms of a GAM fit to the simulated data shown in Fig. 1.

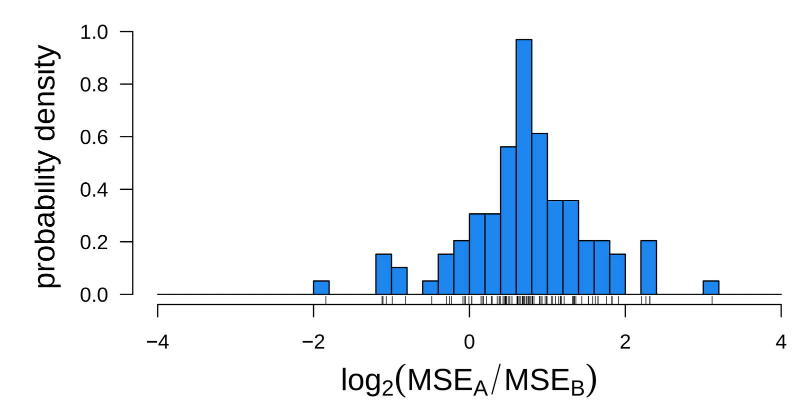

Fig. 4. Model performance compared for 98 landscapes with resources simulated on a 20x20 grid, surrounded by a buffer of width 10; 2 cases were omitted because of model fitting failure. MSE_A is the mean squared error for a nearest resource model; MSE_B is the mean squared error for a GAM with distance-varying functional terms, fit to the same data.

It is widely recognized that species distribution can be influenced by scale-dependent landscape attributes; however, this can result in attempts at finding a single best scale for modelling; treating a landscape attribute as a flexible function of distance is an alternative.

DISPERSAL KERNEL ESTIMATION

Gibson and Austin (1996) describe the spread of the *Citrus tristeza* virus in an orchard, with trees planted on an orthogonal lattice. For the sake of illustrating the use of a scale-varying functional smoother, we will consider a loosely analogous scenario. We will imagine a linear orchard, and discrete time steps, and we will consider not only local dispersal, but also distant dispersal that results in a probability a of immigration, to each location at each timestep. Furthermore, for the simulations run here, it is assumed that a tree recovers in the next time step, unless it is re-infected (perhaps by itself). In this example, observations are simulated mechanistically, but are fit using a regression-type model; a motivation for this is that we might think of a process as being mechanistic, but it might sometimes be more feasible to fit a regression-type model; on this point, see for example, Ovaskainen and Abrego (2020).

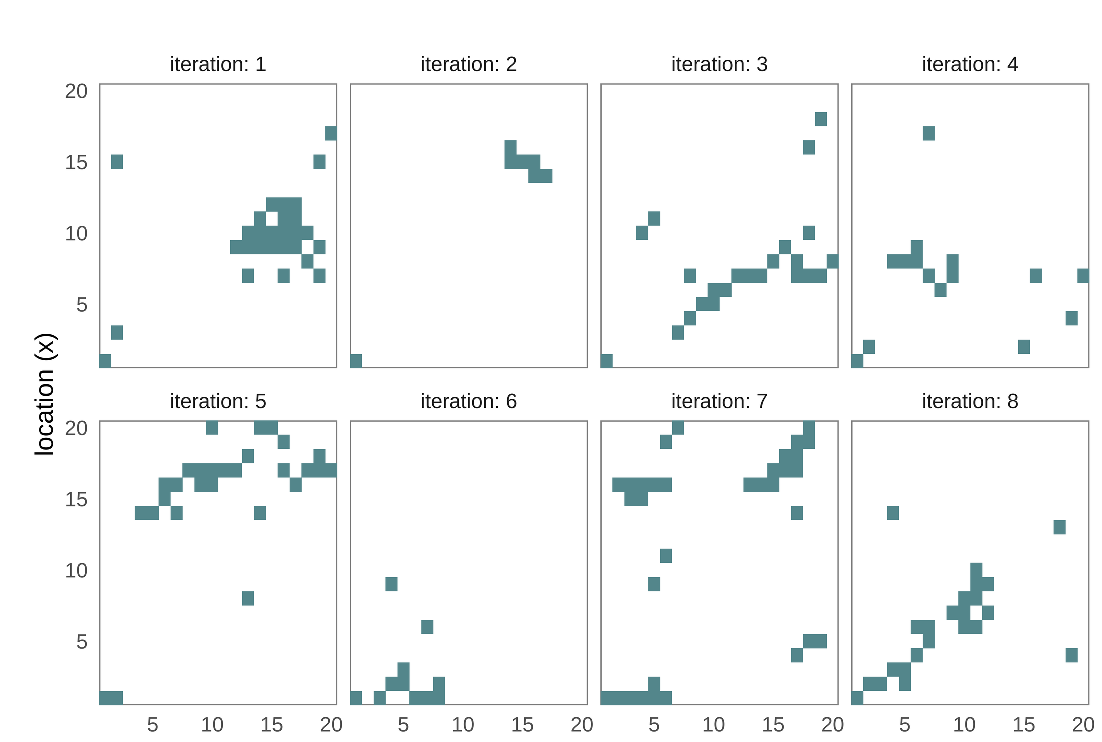

Fig. 5. A few examples of the simulated binomial outcomes ($n = 200$). Each iteration is independent of the others. The initial condition is a single presence at location $x = 1$, at time $= 1$. The function used for the probability of dispersal at distance d was $f(d) = 0.4(0.4^d)$, and the probability used to simulate immigration to a location from outside the local habitat is $a = 0.01$.

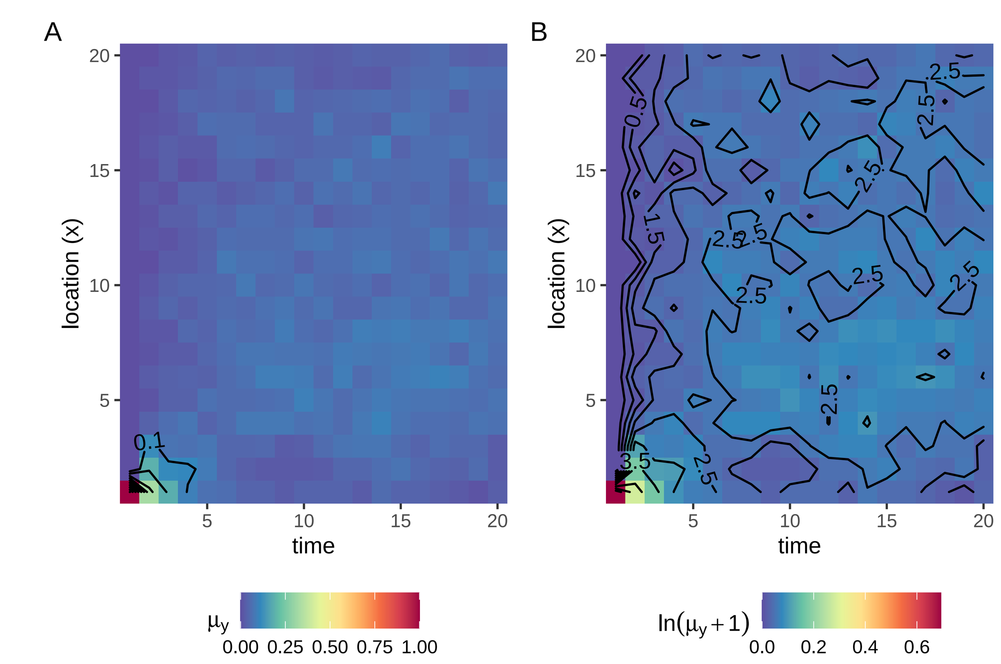

Fig. 6. Panel A shows mean outcomes, $\sum\left(\frac{y_{x,t}}{n}\right)$, across $n = 200$ simulated iterations; panel B shows the same means, transformed to better show local contrast.

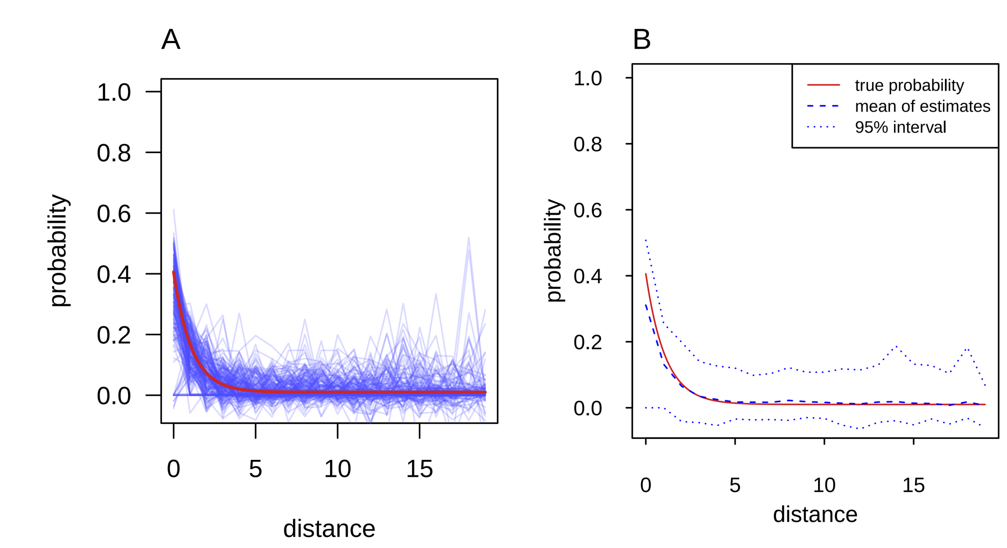

Fig. 7. In both panels A and B, the solid red curve shows the true dispersal kernel used to simulate the observations. In panel A, each blue curve shows the dispersal kernel estimated from a single iteration. In Panel B, the dashed blue curve shows the pointwise means of the estimated dispersal kernel functions; the dotted blue curves show pointwise quantiles.

DISCUSSION

Potential benefits of using scale-varying functional terms in species distribution models implemented with GAMs include improved interpretability and prediction accuracy; however, the increase in model complexity associated with the use of scale-varying functional terms may tend to increase computation time, and could perhaps sometimes result in decreased prediction accuracy in cases where a model becomes overly complex.

References

Cornulier, T., and A. Villers, 2015. Modelling Resource Selection Across Multiple Spatial Scales Using Varying Coefficient Regression. Conference: Spatial Statistics 2015 - Emerging patterns - 9-12 June 2015.

Gibson, G. J., and E. J. Austin, 1996. Fitting and Testing Spatio-Temporal Stochastic Models with Application in Plant Epidemiology. Plant Pathology 45 (2): 172–84.

Nychka, R. F., J. Paige, and S. Sain, 2021. fields: Tools for spatial data. R package version 15.2

Ovaskainen, O., and N. Abrego, 2020. Joint Species Distribution Modelling: with Applications in R. Cambridge University Press: Cambridge.

R Core Team, 2024. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Sims, M., D. A. Elston, A. Larkham, D. H. Nussey, and S. D. Albon, 2007. Identifying when weather influences life-history traits of grazing herbivores. Journal of Animal Ecology. 76, 761–770

Wickham, H., 2016. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag: New York.

Wood, S., 2017. Generalized Additive Models: An Introduction with R, 2nd Edition. CRC Press.

Acknowledgement

Prof. Ben Bolker, McMaster University, provided feedback on early drafts of some of the material presented in this poster.

McMaster University