Understanding spatial effects in species distribution models

Iosu Paradinas*[1,2], Janine Illian [3], and Sophie Smout [1]

[1]Scottish Ocean's Institute. University of St Andrews. East sands, St Andrews, UK.
[2]AZTI, Txatxarramendi Ugartea z/g, 48395, Sukarrieta, Bizkaia, Spain
[3]School of Mathematics and Statistics, University of Glasgow, Glasgow, G12 8QQ, UK

January 2022

Abstract

Most Species Distribution Models include spatial effects to improve prediction at unsampled locations and reduce Type I errors. Ecologists tend to try ecologically interpret the spatial patterns displayed by the spatial effect. However, spatial autocorrelation may be driven by many different unaccounted drivers, which complicates the ecological interpretation of fitted spatial effects. This study wants to provide a practical demonstration that spatial effects are able to smooth the effect of multiple unaccounted drivers. To do so we use a simulation study that fit model-based spatial models using both geostatistics and 2D smoothing splines. Results show that fitted spatial effects resemble the sum of the unaccounted covariate surface(s) in each model.

1 Introduction

Understanding and predicting species spatial patterns through Species Distribution Models (SDM) is pivotal for ecology, evolution and conservation (Zurell et al., 2020)). SDMs quantify the relationship between species occurrence or abundance with biotic and abiotic factors in order to gain ecological and evolutionary understanding (Elith and Leathwick, 2009)). This way SDMs allow us to predict distributions across landscapes and make future predictions based

---

*Iosu Paradinas: ip30@st-andrews.ac.uk

on identified drivers, as well as other latent variables such as spatial or spatio-temporal correlation effects. Generally, a SDM is composed by three types of predictors: non-spatial covariates; spatially structured covariates; and spatial or spatio-temporal autocorrelation effects that accommodate the spatial or spatiotemporal autoccorelation of the data that is unaccounted by our covariates.

Spatial autocorrelation refers to the dependence between pairs of observations in space. In SDMs, spatial effects allow us to predict better and reduce Type I errors in the presence of covariates (Lennon, 2000; Legendre et al., 2002)). In species distribution, spatial autocorrelation may arise as a combination of different factors such as: a geographical range dispersion process, e.g. colonisation; unaccounted environmental or biotic drivers; and other highly dynamic processes such as wind and current (Keitt et al., 2002; Dormann, 2007; De Knegt et al., 2010)). These drivers can influence species distribution at all scales, from micrometres to continental and ocean-wide scales (Legendre, 1993)). However, the size, spacing and extent of sampling units will constrain the scale of inferable drivers, and the scale of spatial autocorrelation (Dungan et al., 2002; De Knegt et al., 2010)). In other words, if we sample at a kilometer scale, we cannot infer processes at a smaller scale, and inversely, if our study area is one kilometer long, we cannot infer processes that affect at a larger scale.

The statistical interpretation of a spatial effect is related to the sign and link function of our linear predictor, but in general terms, positive values refer to areas where we expect more than that predicted by the rest of the linear predictor and vice versa. Ecologically, many SDM studies have linked spatial effects to biological features like home-range (Keitt et al., 2002)), hot-spot size (Ungaro et al., 2014)) and unaccounted environmental drivers (Borcard and Legendre, 1994)), providing reasonable arguments. For example, given a species that is driven by two environmental variables, one that drives the large-scale variation and another that drives the small-scale variation, the residual spatial pattern of a SDM that includes one of the two covariates will resemble the pattern of the unaccounted explanatory variable, either the large-scale or small-scale one. However, as we mentioned before, reality behind ecological processes is often high dimensional and variables that drive spatial correlation can occur at several different scales. In fact, SDMs are seldom able to identify more than a small portion of all the drivers that influence the distribution of the species under study. This results on spatial effects that are potentially driven by many different unaccounted drivers, diluting their interpretability in terms of an individual process. Although this interpretation issues have sporadically been addressed in the literature (Perry et al., 2002; Diniz-Filho et al., 2003; Dormann, 2007; Legendre et al., 2009; De Knegt et al., 2010; Pasanen et al., 2018; Flury et al., 2021)), many modellers fail to acknowledge this probably due to the lack of an explicit study that shows this.

The aim of this article was to provide a practical demonstration that spatial effects are able to smooth the effect of multiple unaccounted drivers, making the biological interpretation of spatial effects rather complicated. To do so, we used model-based spatial models applied over simulated species distribution surfaces. Simulated fields were based on three spatially structured environmental covariates acting at different spatial scales, and a geographical range dispersion process.

2 Simulation

We used an iterative simulation approach to produce spatially aggregated distributions (link to code in Annex A). At each iteration we added a fixed number of new specimens to the study area based on a probability surface constituted by three spatially structured covariates, each operating at different scales (i.e., small, medium and large scale), plus a spatial aggregation process driven by the abundance of the neighbouring areas, mimicking the colonization of a plant species for example. As a result, our simulated species distributions were driven by the sum of four different effects (Figure 1): the influence of three explanatory environmental variables operating at different spatial scales (S = small, M = medium and L = large) and a spatial dispersal effect that increase the spatial autocorrelation of the response variable.

We simulated fifty different scenarios, selected 100 random samples for each scenario and fitted all the possible combinations of model-based spatial models that ranged from a purely spatial model to a full model that accounted for the three covariates (see Table 1). We used two spatial modelling approaches, geostatistics through the Intergated Nested Laplace Approximation approach (INLA) (Lindgren et al., 2015)) and 2D smoothing splines through the MGCV package for R (Augustin et al., 2013; Wood, 2017)).

Our aim was to assess the resemblance between fitted spatial effects and unaccounted covariate surface combinations. Resemblence was assessed through the similarity in pattern score (SIP) (Jones et al., 2016)). SIP scores are bound between zero and one, and high scores denote high similarity in pattern and vice versa. For each simulated scenario, we calculated the SIP score between the spatial effect of every fitted model (rows in Table 2) and all the possible different combinations of covariate surfaces (columns in Table 2), and recorded the absolute difference between the best SIP score and the rest (i.e., SIP differences calculated per row in Table 2). This way, the spatial effect that best resembled a given combination of covariate surfaces scored a zero and that with the worst resemblance recorded the highest value (see Annex for a more detailed explanation of the procedure). As a result, we obtained fifty scores per model and combination of covariate surfaces. Finally, we summarised these scores by their mean and standard deviation. All the R script is available at `https://tinyurl.com/2p8n3e4r`.

3 Results

Results show that fitted spatial effects resemble the sum of the unaccounted covariate surfaces in each model (see highlighted diagonal scores in Table 2). Fitted 2D splines using generalized additive models (GAM) seemed to perform a little worse than model based-geostatistics, probably due to the default selection of knots, but the overall pattern is very similar. This result suggests that spatial effects are able to smooth complex residual spatial patterns originated by a set of covariates that operate at very different scales. For example, model M_M, which only accounts for the mid-scale covariate, estimates a spatial effect that resembles the aggregation of the small-scale and large-scale covariates (S and L respectively). Similarly, the spatial effect of model M_0, which is a purely spatial model (no covariates included), mirrors the combination of all three covariate surfaces (S, M and L). In the particular cases where we included two covariates (i.e., only one unaccounted covariate), spatial effects resembled the missing covariate. At this point, the question is: how many times do SDMs account for all but one driver? One can only speculate this answer but our guess would be: hardly ever.

4 Discussion

Many studies have analyzed the characteristics of spatial effects to describe the unaccounted ecological mechanisms that drive the distribution of species and try to associate spatial effect patterns to single unaccounted drivers. However, most species distributions are driven by a large number of factors and we are seldom able to identify most of these drivers in our statistical models. As a consequence, SDM spatial effects constitute a combination of many unaccounted factors (Keitt et al., 2002; Dormann, 2007; De Knegt et al., 2010)).

This study used a simulation study to illustrate the difficulty in interpreting spatial effects with regards to unaccounted environmental drivers. Readers must realize that did not attempt an exhaustive account of all possible cases, instead, we aimed at illustrating our point using a simple and intuitive approach. Fitted spatial effects resembled the sum of the unaccounted covariate surfaces, including spatial patterns originated by covariates that operated at very different scales. Therefore the biological interpretation of spatial effects may only be valid when the unexplained spatial heterogeneity of the data is characterised by a single dominant driver. However, the environmental and ecological processes that drive the distribution of species are complex and diverse, and one could only arbitrarily assume that there is only one covariate missing in our SDM predictor to make biological interpretations over fitted spatial effects.

In this regard, one could use a multiresolution decomposition approach to identify dominant features within the residual spatial correlation of the data (Pasanen et al., 2018; Flury et al., 2020)). This method essentially estimates the range of spatial correlation at different resolutions of the data, or in this case, residuals of the SDM to help us identify the scale-dependent features within the spatial effect of the residuals. Then, assuming that each scale is characterized by a single dominant driver (Perry et al., 2002)), one could relate them to underlying process generating mechanisms.

5 Conclusions

Spatial autocorrelation is a common feature in ecological data. As a consequence, spatial correlation models are important to correctly estimate covariate standard errors and therefore reduce Type I errors. Additionally, spatial correlation terms estimate the residual spatial structure of the data, improving the predictive capacity of our models at locations that are within range. In ecology, residual spatial patterns are potentially driven by complex multivariate and multi-scaled systems, which can be accommodated by a single spatial effect. Therefore, the biological interpretation of spatial effects is very difficult. A multiresolution decomposition of residual spatial patterns (Flury et al., 2020)) could help us identify the scale-dependent features within the spatial correlation structure of the residuals assuming that each scale is characterized by a single dominant driver.

6 Figure and Tables

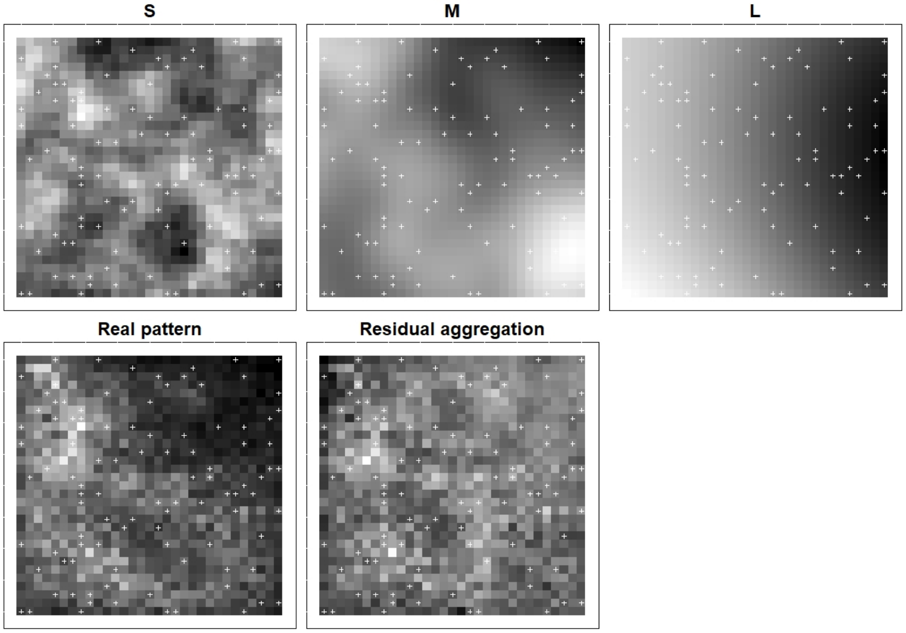

Figure 1: Visualization of the different autocorrelated drivers that influence the abundance pattern in a simulated scenario. S, M and L refer to the small, medium and large scaled covariate fields, respectively. Residual aggregation refers to the geographical range dispersion. White crosses refer to the simulated 100 samples.

Model	Linear predictor	Missing covariates
M_0	$\beta_0 + W$	S, M & L
M_S	$\beta_0 + S + W$	M & L
M_M	$\beta_0 + M + W$	S & L
M_L	$\beta_0 + L + W$	S & M
M_ML	$\beta_0 + M + L + W$	S
M_SM	$\beta_0 + S + M + W$	L
M_SL	$\beta_0 + S + L + W$	M
M_SML	$\beta_0 + S + M + L + W$	–

Table 1: Summary of fitted models. W refers to a geostatistical spatial correlation term, S, M and L refer to the small, medium and large scale covariates, respectively.

		Combination of drivers							
	Model	Residual	S	M	L	S & M	S & L	M & L	S, M & L
Geostatistics (INLA)	M_0	0.62 (0.14)	0.30 (0.13)	0.27 (0.18)	0.35 (0.22)	0.11 (0.06)	0.17 (0.08)	0.12 (0.15)	**0.01 (0.02)**
	M_S	0.56 (0.18)	0.66 (0.22)	0.19 (0.16)	0.25 (0.19)	0.33 (0.17)	0.41 (0.16)	**0.01 (0.03)**	0.16 (0.12)
	M_M	0.47 (0.17)	0.19 (0.15)	0.71 (0.25)	0.26 (0.22)	0.26 (0.21)	**0.04 (0.08)**	0.37 (0.21)	0.11 (0.14)
	M_L	0.55 (0.17)	0.21 (0.14)	0.24 (0.21)	0.78 (0.35)	**0.04 (0.04)**	0.29 (0.20)	0.33 (0.24)	0.08 (0.12)
	M_SM	0.34 (0.17)	0.50 (0.24)	0.61 (0.28)	**0.07 (0.13)**	0.48 (0.26)	0.22 (0.15)	0.21 (0.18)	0.24 (0.20)
	M_SL	0.41 (0.23)	0.51 (0.21)	**0.08 (0.11)**	0.67 (0.35)	0.18 (0.11)	0.53 (0.25)	0.17 (0.20)	0.20 (0.16)
	M_ML	0.36 (0.18)	**0.06 (0.10)**	0.59 (0.24)	0.65 (0.26)	0.11 (0.11)	0.13 (0.14)	0.60 (0.24)	0.16 (0.17)
	M_SML	**0.09 (0.15)**	0.27 (0.16)	0.40 (0.22)	0.43 (0.24)	0.25 (0.16)	0.28 (0.18)	0.40 (0.24)	0.25 (0.22)
2D splines (GAM)	M_0	0.50 (0.09)	0.18 (0.10)	0.11 (0.08)	0.07 (0.07)	0.09 (0.08)	0.12 (0.08)	**0.04 (0.05)**	**0.04 (0.05)**
	M_S	0.50 (0.09)	0.38 (0.17)	0.10 (0.10)	0.08 (0.09)	0.22 (0.14)	0.28 (0.17)	**0.02 (0.04)**	0.14 (0.11)
	M_M	0.48 (0.10)	0.15 (0.11)	0.35 (0.19)	**0.03 (0.04)**	0.24 (0.16)	**0.07 (0.08)**	0.23 (0.17)	0.15 (0.14)
	M_L	0.33 (0.19)	0.13 (0.13)	**0.06 (0.08)**	0.16 (0.19)	**0.08 (0.10)**	0.16 (0.17)	0.11 (0.12)	0.11 (0.13)
	M_SM	0.49 (0.10)	0.38 (0.17)	0.36 (0.19)	**0.00 (0.02)**	0.42 (0.22)	0.23 (0.14)	0.18 (0.14)	0.28 (0.19)
	M_SL	0.35 (0.17)	0.25 (0.19)	**0.05 (0.08)**	0.16 (0.17)	0.16 (0.13)	0.26 (0.19)	0.10 (0.12)	0.18 (0.13)
	M_ML	0.33 (0.16)	**0.09 (0.10)**	0.20 (0.19)	0.14 (0.16)	0.16 (0.14)	0.13 (0.14)	0.23 (0.19)	0.18 (0.14)
	M_SML	0.34 (0.17)	0.23 (0.21)	0.20 (0.20)	0.14 (0.20)	0.27 (0.20)	0.26 (0.19)	0.22 (0.18)	0.28 (0.17)

Table 2: Resemblance between fitted spatial effects, using geostatistics and 2D smoothing splines, against all the possible combinations of covariate surfaces (per simulation). Scores must be read by row, and reflect the difference between the best SIP score and all possible combinations of drivers for each simulation and model. Therefore, lower values represent higher resemblance and have been highlighted in bold. We present the mean difference and standard deviation (in parenthesis). See Annex for a more detailed explanation of the procedure that we followed.

References

Nicole H Augustin, Verena M Trenkel, Simon N Wood, and Pascal Lorance. Space-time modelling of blue ling for fisheries stock management. *Environmetrics*, 24(2):109–119, 2013.

Daniel Borcard and Pierre Legendre. Environmental control and spatial structure in ecological communities: an example using oribatid mites (acari, oribatei). *Environmental and Ecological statistics*, 1(1):37–61, 1994.

HJ De Knegt, F van van Langevelde, MB Coughenour, AK Skidmore, WF De Boer, IMA Heitkönig, NM Knox, R Slotow, C Van der Waal, and HHT Prins. Spatial autocorrelation and the scaling of species–environment relationships. *Ecology*, 91(8):2455–2465, 2010.

José Alexandre Felizola Diniz-Filho, Luis Mauricio Bini, and Bradford A Hawkins. Spatial autocorrelation and red herrings in geographical ecology. *Global ecology and Biogeography*, 12(1):53–64, 2003.

Carsten F Dormann. Effects of incorporating spatial autocorrelation into the analysis of species distribution data. *Global ecology and biogeography*, 16(2): 129–138, 2007.

Jennifer L Dungan, JN Perry, MRT Dale, Pousty Legendre, S Citron-Pousty, M-J Fortin, A Jakomulska, M Miriti, and MS2002 Rosenberg. A balanced view of scale in spatial statistical analysis. *Ecography*, 25(5):626–640, 2002.

Jane Elith and John R Leathwick. Species distribution models: ecological explanation and prediction across space and time. *Annual review of ecology, evolution, and systematics*, 40:677–697, 2009.

Roman Flury, Florian Gerber, Bernhard Schmid, and Reinhard Furrer. Identification of dominant features in spatial data. *Spatial Statistics*, 41:100483, 2020.

Roman Flury, Florian Gerber, Bernhard Schmid, and Reinhard Furrer. Identification of dominant features in spatial data. *Spatial Statistics*, 41:100483, 2021.

Esther L Jones, Luke Rendell, Enrico Pirotta, and Jed A Long. Novel application of a quantitative spatial comparison tool to species distribution data. *Ecological Indicators*, 70:67–76, 2016.

Timothy H Keitt, Ottar N Bjørnstad, Philip M Dixon, and Steve Citron-Pousty. Accounting for spatial pattern when modeling organism-environment interactions. *Ecography*, 25(5):616–625, 2002.

Pierre Legendre. Spatial autocorrelation: trouble or new paradigm? *Ecology*, 74(6):1659–1673, 1993.

Pierre Legendre, Mark RT Dale, Marie-Josée Fortin, Jessica Gurevitch, Michael Hohn, and Donald Myers. The consequences of spatial structure for the design and analysis of ecological field surveys. *Ecography*, 25(5):601–615, 2002.

Pierre Legendre, Xiangcheng Mi, Haibao Ren, Keping Ma, Mingjian Yu, I-Fang Sun, and Fangliang He. Partitioning beta diversity in a subtropical broad-leaved forest of china. *Ecology*, 90(3):663–674, 2009.

Jack J Lennon. Red-shifts and red herrings in geographical ecology. *Ecography*, 23(1):101–113, 2000.

Finn Lindgren, Håvard Rue, et al. Bayesian spatial modelling with r-inla. *Journal of Statistical Software*, 63(19):1–25, 2015.

Leena Pasanen, Tuomas Aakala, and Lasse Holmström. A scale space approach for estimating the characteristic feature sizes in hierarchical signals. *Stat*, 7 (1):e195, 2018.

JN Perry, AM Liebhold, MS Rosenberg, J Dungan, M Miriti, A Jakomulska, and S Citron-Pousty. Illustrations and guidelines for selecting statistical methods for quantifying spatial pattern in ecological data. *Ecography*, 25(5):578–600, 2002.

Fabrizio Ungaro, Ingo Zasada, and Annette Piorr. Mapping landscape services, spatial synergies and trade-offs. a case study using variogram models and geostatistical simulations in an agrarian landscape in north-east germany. *Ecological indicators*, 46:367–378, 2014.

Simon N Wood. *Generalized additive models: an introduction with R*. CRC press, 2017.

Damaris Zurell, Janet Franklin, Christian König, Phil J Bouchet, Carsten F Dormann, Jane Elith, Guillermo Fandos, Xiao Feng, Gurutzeta Guillera-Arroita, Antoine Guisan, et al. A standard protocol for reporting species distribution models. *Ecography*, 2020.

A Annex

The aim of this annex is to explain the procedure that we followed to create Table 2. To do so we use a single simulated species distribution (as compared to 50 simulations in the study) that is also driven by three spatially structured environmental covariates acting at different spatial scales and a geographical range dispersion process.

We fitted all the models described in Table 1 and we computed SIP scores between each model's spatial effect and all the possible different combinations of covariate surfaces. By doing so, we get Table 3, which displays highest SIP scores along the diagonal, matching the combination of drivers (columns) with the covariates that are missing in the fitted models (rows).

	Combination of drivers							
Model	Residual	S	M	L	S & M	S & L	M & L	S, M & L
M_0	-0.01	0.43	0.57	0.49	0.71	0.55	0.73	**0.82**
M_S	0.12	-0.04	0.73	0.43	0.49	0.25	**0.85**	0.59
M_M	0.03	0.50	0.19	0.69	0.47	**0.72**	0.46	0.72
M_L	0.05	0.40	0.75	0.06	**0.78**	0.33	0.65	0.75
M_SM	0.16	0.03	-0.09	**0.81**	-0.03	0.48	0.46	0.33
M_SL	0.11	-0.06	**0.83**	-0.15	0.57	0.01	0.76	0.53
M_ML	0.01	**0.64**	0.13	0.03	0.57	0.54	0.05	0.57
M_SML	**0.22**	-0.05	0.17	-0.12	-0.00	-0.12	0.10	0.09

Table 3: SIP scores between fitted spatial effects and all the combinations of covariate surfaces. Scores must be read by row. Values closer to one reflect bigger resemblance between spatial fields.

Once we repeat the simulation 50 times we get 50 SIP scores for each position in the table, which could be summarised by the mean and standard deviation of these 50 values. However, we decided to use the difference between the best SIP score for each model and combinations of covariate fields because results were clearer, i.e. differences by row in the Table 3. This way Table 3 becomes Table 4, where zero values represents the best SIP score per model (by row) and the rest of the scores represent the SIP score difference with respect to the best score by row.

	Combination of drivers							
Model	Residual	S	M	L	S & M	S & L	M & L	S, M & L
M_0	0.83	0.39	0.24	0.32	0.11	0.27	0.08	**0.00**
M_S	0.74	0.89	0.13	0.43	0.36	0.60	**0.00**	0.26
M_M	0.70	0.23	0.53	0.03	0.26	**0.00**	0.26	0.01
M_L	0.74	0.38	0.03	0.72	**0.00**	0.45	0.13	0.03
M_SM	0.66	0.78	0.90	**0.00**	0.85	0.33	0.35	0.48
M_SL	0.72	0.90	**0.00**	0.99	0.26	0.82	0.07	0.30
M_ML	0.63	**0.00**	0.51	0.61	0.06	0.10	0.59	0.07
M_SML	**0.00**	0.28	0.05	0.34	0.23	0.35	0.13	0.13

Table 4: The difference in score between the best SIP score and the rest for each model (by row). Values closer to zero reflect bigger resemblance between spatial fields.