Spatial occupancy models for data collected on stream networks

Olivier Gimenez[1]*

2024-07-25

[1] CEFE, Univ Montpellier, CNRS, EPHE, IRD, Montpellier, France

* Corresponding author: olivier.gimenez@cefe.cnrs.fr

To effectively monitor biodiversity in streams and rivers, we need to quantify species distribution accurately. Occupancy models are useful for distinguishing between the non-detection of a species and its actual absence. While these models can account for spatial autocorrelation, they are not suited for streams and rivers due to their unique network spatial structure. Here, I propose spatial occupancy models specifically designed for data collected on stream and river networks. I present the statistical developments and illustrate their application using data on a semi-aquatic mammal. Overall, spatial stream network occupancy models offer a robust method for assessing biodiversity in freshwater ecosystems.

Keywords: Bayesian statistics, Spatial stream network models, Occupancy models, Spatial autocorrelation, Wildlife monitoring

Introduction

Streams and rivers provide essential habitats for numerous species of animals and plants (Reid et al. 2019). The ecological health of these freshwater ecosystems is paramount not only for the biodiversity they harbor but also for the ecosystem services they provide, which are indispensable to both wildlife and human populations (Vári et al. 2021). However, human activities are altering the natural conditions of streams, rivers, and their associated riparian habitats, jeopardizing the persistence of these ecosystems (Albert et al. 2020).

In this context, species distribution models (SDMs) are essential tools in understanding and preserving biodiversity (Elith and Leathwick 2009). SDMs predict the distribution of species, helping scientists and conservationists in identifying critical habitats and biodiversity hotspots. Additionally, SDMs inform strategies aimed at mitigating the impacts of climate and land-use changes, managing invasive species, and enhancing habitat connectivity in freshwater ecosystems (Domisch et al. 2015).

SDMs are influenced by two main issues: imperfect detection and spatial autocorrelation (Guélat and Kéry 2018). First, imperfect detection occurs when a species present in a given area is not detected during surveys due to factors such as observer experience, species behavior, and environmental conditions. Ignoring imperfect detection can lead to biased estimates of species distribution and flawed inferences about the relationship between species presence and environmental factors (e.g., Lahoz-Monfort et al. 2014). This can misinform conservation strategies and habitat management decisions. To address this issue, occupancy models are SDMs that rely on repeated visits of spatial sampling units for inferring distribution (MacKenzie et al. 2017). These models have been widely used in freshwater ecosystems for various taxa (e.g., Wedderburn et al. 2022, Couturier et al. 2023).

Second, SDMs rely on the assumption of independent residuals. This assumption may be violated if nearby sampling sites tend to have similar probabilities of species presence, leading to biased estimates of species distribution and potentially inflating the effects of environmental factors (Dormann et al. 2007). Several extensions of occupancy models have been proposed to account for spatial autocorrelation, building on the spatial statistics literature, with conditional autoregressive models (Johnson et al. 2013) and geoadditive models (Rushing et al. 2019). However, these models rely on the Euclidean distance between the spatial sampling units, which does not acknowledge the spatial structure in networks of streams and rivers.

Here I propose spatial occupancy models that account for spatial autocorrelation based on flow connectivity and river network (Peterson et al. 2013). I build on the linear mixed modelling approach developed by Ver Hoef and Peterson (2010) and Peterson and Hoef (2010), which integrates various distance-based spatial correlation structures (both Euclidean and non-Euclidean) within a single model. I plug-in this variance component approach into occupancy models using a Bayesian approach. A similar approach was recently undertaken by Lu et al. (2024) for count data to estimate abundance. Below I outline the statistical developments of this model and demonstrate its application to a semi-aquatic mammal in French streams and rivers.

Methods

Occupancy models

To address imperfect detection, I use occupancy models to estimate the true species distribution (MacKenzie et al. 2017). In these models, monitoring occurs across S spatial sampling units, or sites. If detection was perfect, the state z_i of a site i would be a Bernoulli random variable, taking value 1 with occupancy probability ψ_i if the site was occupied, and 0 otherwise with probability $1 - \psi$. However, because the ecological process of state occupancy z_i is only partially observable (since the species might be present but undetected), we must also account for the observation process, which is also modeled as a Bernoulli random variable. When the species is detected at site i, i.e. $y_i = 1$, with detection probability p, it confirms that the site is occupied. Conversely, if the species is not detected, i.e. $y_i = 0$, with probability $1 - p$, we cannot determine whether the site is occupied or not. Both parameters, ψ and p, can be modeled as functions of explanatory spatial variables, in the spirit of generalized linear models, and logistic regressions for example. To estimate occupancy and detection probabilities separately, data should be collected from at least two independent visits to each site within a short period, ensuring that sites remain in the same state.

Spatial autocorrelation for stream networks

How is spatial autocorrelation accounted for in occupancy models? The usual way is to write the probabilities of occupancy $\boldsymbol{\psi} = (\psi_1, \ldots, \psi_S)$ on some scale, say the logit scale, as a function of explanatory variables gathered in a matrix $\mathbf{X}$ with corresponding regression parameters $\boldsymbol{\beta}$ that need to be estimated. To account for spatial autocorrelation, a random effect $\boldsymbol{\epsilon}$ is added to the model, which captures the spatial dependencies among sites (Guélat and Kéry 2018):

$$\text{logit}(\boldsymbol{\psi}) = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}.$$

The random effect $\boldsymbol{\epsilon}$ can be structured using methods such as conditional autoregressive models and their extensions (Johnson et al. 2013), or geoadditive models (Rushing et al. 2019). However, these approaches typically rely on Euclidean distance to assess proximity among sites, which may not fully capture the complex spatial dependencies present in streams and rivers. Specifically, we are interested in flow connectivity and the topology of streams and rivers (Peterson et al. 2013). Following Ver Hoef and Peterson (2010) and Peterson and Hoef (2010), I define two sites as flow-connected if water flows from an upstream site to a downstream site, and as flow-unconnected if they share a common confluence downstream but do not directly share flow. Then I parameterize occupancy by rewriting the random effect as a mixture of four components as follows:

$$\text{logit}(\boldsymbol{\psi}) = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\tau}_{tu} + \boldsymbol{\tau}_{td} + \boldsymbol{\tau}_{eu} + \boldsymbol{\epsilon}$$

where $\boldsymbol{\tau}_{tu}$ is a random effect with spatial covariance between flow-connected sites that can occur in the same direction of the river flow (tail-up), which is suitable for organisms that move passively, such as mussels), $\boldsymbol{\tau}_{td}$ is a random effect with spatial covariance between flow-connected and flow-unconnected sites, which can occur with or against the direction of flow (tail-down), applicable to organisms that move actively, such as semi-aquatic mammals, $\boldsymbol{\tau}_{eu}$ is a random effect with a spatial covariance independent of the network topology, influenced by factors like air temperature or precipitation, and $\boldsymbol{\epsilon}$ is a random effect with variance, often referred to as the nugget, which accounts for additional variability. How to build these covariance components is described in details elsewhere (e.g., Ver Hoef et al. 2019), and I provide an example in the next section.

Case study

To illustrate the new approach, I investigated the impact of human disturbance on the occupancy of European otter (*Lutra lutra*) in France. The otter, a semi-aquatic mammal, faced near extinction in the 20th century in France due to extensive hunting for its fur. With hunting bans and protection efforts, the species is now recolonizing the country, and the ecological question is assessing its current distribution. Data on otter detection and non-detection were collected in 2003-2005 in the Midi-Pyrénées region (see panel a in Fig. 1). Observers searched for signs of otter presence at a small river catchment scale, which was used as the spatial sampling unit. These data were analyzed by Couturier et al. (2023), who found that human density and the proportion of cultivated area influenced occupancy. In this study, I focus on a subsample of this dataset, covering $S = 56$ sites in the Lot, Aveyron and Cantal counties, which were visited 3 times (see panel b in Fig. 1). I used human population density as a proxy for human disturbance, calculated as the number of inhabitants per km2 within a 200-m buffer around each stream (see panel c in Fig. 1). Additionally, I considered the proportion of cultivated areas as an explanatory variable. Detection was considered as constant. Regarding spatial autocorrelation, since otters can move both downstream and upstream, I used a tail-down model

$$\text{logit}(\boldsymbol{\psi}) = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\tau}_{td}$$

with an exponential covariance:

$$\text{Cov}(\tau_i, \tau_j) = \begin{cases} \sigma^2 \exp(-h/\theta), & \text{if sites } i \text{ and } j \text{ are flow-connected} \\ \sigma^2 \exp(-(a+b)/\theta), & \text{if sites } i \text{ and } j \text{ are flow-unconnected} \end{cases}$$

where I dropped the *td* notation in τ_{td} for clarity, h is the stream distance between sites i and j, b denotes the longer of the distances between sites i and j to the common downstream junction, a denotes the shorter of the two distances, σ^2 is a variance parameter usually referred to as the partial sill and θ is a range parameter (see Apendix A in Peterson and Hoef 2010).

Implementation

For all analyses, I used the statistical language `R` (R Core Team 2023). I used the `openSTARS` (Kattwinkel et al. 2020) and `SSN` (Ver Hoef et al. 2014) packages to build and characterize the network and calculate stream distances. I fitted models within the Bayesian framework, specifying weakly informative priors and using the `rstan` (Stan Development Team 2023) package. I ran two chains for a total of 15,000 iterations with a burn-in of 5,000 iterations. I summarized posterior distributions with posterior mean and 95% credible intervals. I assessed model convergence using R-hat values (< 1.1), effective sample size (> 100), and visual inspection of the trace plots. I provide additional information in the code available at https://github.com/oliviergimenez/spatial-stream-network-occupancy-model.

Results and discussion

Here, I provide the parameter estimates from the new model accommodating spatial autocorrelation, unless otherwise specified. Detection probability was less than one, estimated at 0.71 (0.59, 0.80). The proportion of cultivated area had no effect, with a slope estimated at 0.60 (-0.67, 1.96). Population density also had no effect on occupancy probability, with a slope estimated at -0.96 (-2.24, 0.17). However, I did find a negative effect when spatial autocorrelation was ignored, with a slope estimated at -1.10 (-1.99, -0.34). This latter result aligns with a previous analysis of a more extensive dataset (Couturier et al. 2023) that also ignored spatial autocorrelation.

As anticipated, the effect size of human density increased when spatial autocorrelation was ignored. The most likely explanation for this is a bias due to an omitted variable. There is spatial autocorrelation in human density (see panel c in Fig. 1), which inflates its effect size; this bias is controlled for when spatial autocorrelation is included in the model. There must be spatial variation in occupancy probabilities attributable to another variable that needs to be accounted for.

Two short-term perspectives arise from this work. From a methodological perspective, the new approach could be extended to multi-season occupancy models, enabling the modeling of colonization probability as a function of distance to habitat features that may impede species movement. This would facilitate the quantification of landscape connectivity in freshwater ecosystems. Such development requires moving to spatio-temporal models for stream and river data, which have recently become avaible (Santos-Fernandez et al. 2022). From an ecological perspective, the new approach presents significant potential for the analysis of environmental DNA (eDNA). The eDNA methodology offers substantial promise for the non-invasive monitoring of biodiversity in freshwater ecosystems (Carraro et al. 2020). While spatial stream network models have been employed to analyze eDNA data (Winkowski et al. 2024), these models have overlooked the issue of imperfect detection. Previous studies have recognized occupancy models as effective tools for eDNA data analysis (Burian et al. 2021), with some considering spatial autocorrelation (Chen and Ficetola 2019), however they have yet to integrate spatial stream networks. The new approach addresses this gap by incorporating both imperfect detection and spatial stream networks.

Acknowledgments

I would like to warmly thank Jay Ver Hoef and Edgar Santos Fernández for useful discussions on spatial stream network models. I also thank Maëlis Kervellec for sharing her code to wrangle the otter data.

Ethics and Integrity statements

Data availability statement

Data and code are available at https://github.com/oliviergimenez/spatial-stream-network-occupancy-model.

Funding statement

This research is a product of the DISCAR group funded by the French Foundation for Research on Biodiversity (FRB) through its synthesis center CESAB.

Conflict of interest disclosure

The author has no conflicts of interest to declare.

References

Albert, J. S., G. Destouni, S. M. Duke-Sylvester, A. E. Magurran, T. Oberdorff, R. E. Reis, K. O. Winemiller, and W. J. Ripple. 2020. Scientists' warning to humanity on the freshwater biodiversity crisis. Ambio 50:85–94.

Burian, A., Q. Mauvisseau, M. Bulling, S. Domisch, S. Qian, and M. Sweet. 2021. Improving the reliability of eDNA data interpretation. Molecular Ecology Resources 21:1422–1433.

Carraro, L., E. Mächler, R. Wüthrich, and F. Altermatt. 2020. Environmental DNA allows upscaling spatial patterns of biodiversity in freshwater ecosystems. Nature Communications 11:3585.

Chen, W., and G. F. Ficetola. 2019. Conditionally autoregressive models improve occupancy analyses of autocorrelated data: An example with environmental DNA. Molecular Ecology Resources 19:163–175.

Couturier, T., J. Steinmetz, P. Defos du Rau, D. Marc, E. Trichet, R. Gomes, and A. Besnard. 2023. Intensive agriculture as the main limiting factor of the otter's return in southwest france. Biological Conservation 279:109927.

Domisch, S., S. Jähnig, J. Simaika, M. Kuemmerlen, and S. Stoll. 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 186:45–61.

Dormann, F. C., J. M. McPherson, M. B. Araújo, R. Bivand, J. Bolliger, G. Carl, R. G. Davies, A. Hirzel, W. Jetz, W. Daniel Kissling, I. Kühn, R. Ohlemüller, P. R. Peres-Neto, B. Reineking, B. Schröder, F. M. Schurr, and R. Wilson. 2007. Methods to account for spatial autocorrelation in the analysis of species distributional data: A review. Ecography 30:609–628.

Elith, J., and J. R. Leathwick. 2009. Species distribution models: Ecological explanation and prediction across space and time. Annual Review of Ecology, Evolution, and Systematics 40:677–697.

Guélat, J., and M. Kéry. 2018. Effects of spatial autocorrelation and imperfect detection on species distribution models. Methods in Ecology and Evolution 9:1614–1625.

Johnson, D. S., P. B. Conn, M. B. Hooten, J. C. Ray, and B. A. Pond. 2013. Spatial occupancy models for large data sets. Ecology 94:801–808.

Kattwinkel, M., E. Szöcs, E. Peterson, and R. Schäfer. 2020. Preparing GIS data for analysis of stream monitoring data: The R package openSTARS. Plos One 15:e0239237.

Lahoz-Monfort, J. J., G. Guillera-Arroita, and B. A. Wintle. 2014. Imperfect detection impacts the performance of species distribution models. Global Ecology and Biogeography 23:504–515.

Lu, X., Y. Kanno, G. P. Valentine, J. M. Rash, and M. B. Hooten. 2024. Using multi-scale spatial models of dendritic ecosystems to infer abundance of a stream salmonid. Journal of Applied Ecology 61:1703–1715.

MacKenzie, D. I., J. D. Nichols, J. A. Royle, K. H. Pollock, L. L. Bailey, and J. E. Hines. 2017. Occupancy Estimation and Modeling: Inferring Patterns and Dynamics of Species Occurrence. Elsevier.

Peterson, E. E., and J. M. V. Hoef. 2010. A mixed-model moving-average approach to geostatistical modeling in stream networks. Ecology 91:644–651.

Peterson, E. E., J. M. Ver Hoef, D. J. Isaak, J. A. Falke, M.-J. Fortin, C. E. Jordan, K. McNyset, P. Monestiez, A. S. Ruesch, A. Sengupta, N. Som, E. A. Steel, D. M. Theobald, C. E. Torgersen, and S. J. Wenger. 2013. Modelling dendritic ecological networks in space: An integrated network perspective. Ecology Letters 16:707–719.

R Core Team. 2023. R: A language and environment for statistical computing. R Foundation for

Statistical Computing, Vienna, Austria.

Reid, A. J., A. K. Carlson, I. F. Creed, E. J. Eliason, P. A. Gell, P. T. J. Johnson, K. A. Kidd, T. J. MacCormack, J. D. Olden, S. J. Ormerod, J. P. Smol, W. W. Taylor, K. Tockner, J. C. Vermaire, D. Dudgeon, and S. J. Cooke. 2019. Emerging threats and persistent conservation challenges for freshwater biodiversity. Biological Reviews 94:849–873.

Rushing, C., J. andrew Royle, D. Ziolkowski Jr, and K. Pardieck. 2019. Modeling spatially and temporally complex range dynamics when detection is imperfect. Scientific Reports 9.

Santos-Fernandez, E., J. M. Ver Hoef, E. E. Peterson, J. McGree, D. J. Isaak, and K. Mengersen. 2022. Bayesian spatio-temporal models for stream networks. Computational Statistics & Data Analysis 170:107446.

Stan Development Team. 2023. RStan: The R interface to Stan.

Vári, Á., S. Podschun, T. Eros, T. Hein, B. Pataki, C. Ioja, C. Adamescu, A. Gerhardt, T. Gruber, A. Dedić, M. Ciric, B. Gavrilović, and A. Báldi. 2021. Freshwater systems and ecosystem services: Challenges and chances for cross-fertilization of disciplines. Ambio 51:135–151.

Ver Hoef, J., and E. Peterson. 2010. A moving average approach for spatial statistical models of stream networks. Journal of the American Statistical Association 105:6–18.

Ver Hoef, J., E. Peterson, D. Clifford, and R. Shah. 2014. SSN: An R package for spatial statistical modeling on stream networks. Journal of Statistical Software 56:1–45.

Ver Hoef, J., E. Peterson, and D. Isaak. 2019. Spatial statistical models for stream networks. Pages 421–441 *in* A. E. Gelfand, M. Fuentes, J. A. Hoeting, and S. R. Lyttleton, editors. Handbook of environmental and ecological statistics. Chapman; Hall/CRC.

Wedderburn, S. D., N. S. Whiterod, and L. Vilizzi. 2022. Occupancy modelling confirms the first extirpation of a freshwater fish from one of the world's largest river systems. Aquatic Conservation: Marine and Freshwater Ecosystems 32:258–268.

Winkowski, J. J., J. D. Olden, and S. Brown. 2024. Integrating spatial stream network models and environmental DNA to estimate current and future distributions of nonnative smallmouth bass. Transactions of the American Fisheries Society 153:180–199.

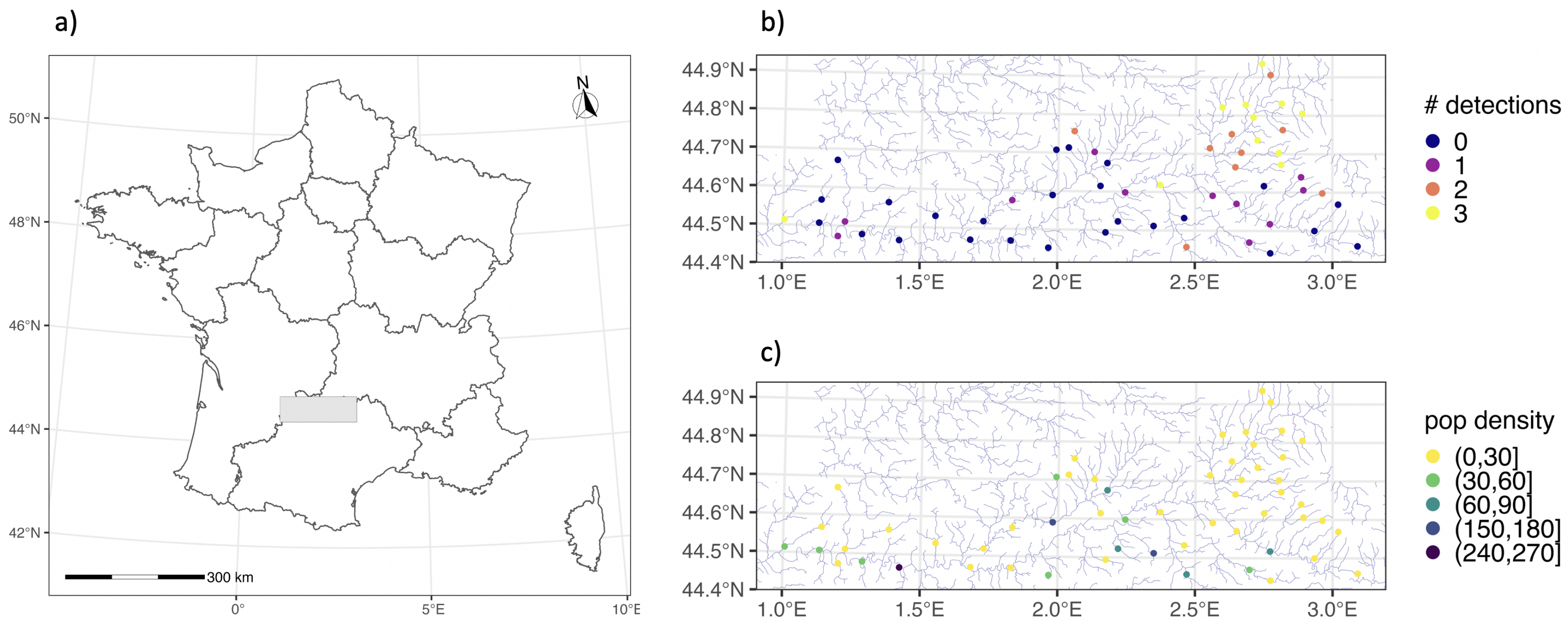

Figure 1: Information on the otter data. In panel a), the study area is given in a grey rectable on a map of France. In panel b) the number of detections is given on a map of the study area. In panel c) the human population density is represented on a map of the study area, in number of inhabitants per km2.