Comparing alternative Bayesian structures for semi-continuous and spatio-temporally correlated data: an application to fisheries

Iosu Paradinas[1], *David Conesa*[1], *Antonio López-Quílez*[1], *José María Bellido*[2]

[1]Departament d'Estadística i Investigació Operativa. Universitat de Valéncia.
C/ Dr. Moliner 50. Burjassot. 46100. Valencia. Spain.
[2]Instituto Español de Oceanografía. Centro Oceanográfico de Murcia.
C/ Varadero 1. San Pedro del Pinatar. 30740. Murcia. Spain.

Abstract

Spatial modelling is an essential tool for the management of living natural resources such as fish stocks. This study considers two classic ecological problems, zero-inflated continuous data and temporal correlation between sampling events. We fit a two-part spatio-temporal regression with shared components and we also compare different model structures to infer the ecological behaviour of the underlying process.

Keywords: Joint modelling; Spatio-temporal; Species distribution

1. Introduction

First, species distribution maps were simple representations of sighting sites using points or shaded areas. Statistical developments brought species distribution models, that started to relate observations to environmental variables and later on, include spatial correlation terms. Nowadays, we are still facing a handful of restrictions to express species ecology in mathematical equations and interactions, some of which may be key requirements for certain conservation management tools.

In fisheries, the latest management directives, e.g. Ecosystem Approach to Fisheries Management (EAFM), require the understanding of marine biological processes at a spatial scale [1]. However, quantifying the ecosystemic importance of an area is a challenging task because of the inherent constrains of sampling at sea. Furthermore, fish distribution can vary with time, thus, understanding that distributional pattern changes of key fish species could be an outstanding proxy to the EAFM.

Model based geostatistics allows fine scale spatial models and they have been widely used in fisheries. However biological spatial processes not only tend to be spatially correlated but also to evolve in time. Consequently, the inclusion of a temporal structure in the model not only improves the spatial interpolation but also allows us to predict forward in time at different locations. Temporal correlation lays basically on the same principle as spatial correlation but since temporal and spatial scales are different, such spatio-temporal analysis is far more complicated than the simple addition of an extra dimension to the spatial domain. Moreover, most fishery survey designs sample the target study area during a fixed time window of the year, which forces scientists to both assume static populations during each time window and to discretize time.

Another important issue when modelling fish species distribution is the usual semi-continuous nature of the response variable. Fish abundance is commonly measured in continuous Catch Per Unit Effort (CPUE) values, where zero observations are quite common at adverse environmental conditions. A straightforward approach is to model this data as a result of two processes, where one process determines whether the response is zero and the other determines the abundance when it is non-zero. This way both models are qualitatively independent and the fitted relationships with the covariates can be different. Nevertheless, it is natural to think that both processes are related; low abundance sites may be linked to low probabilities of occurrence and vice versa.

2. Hierarchical spatio-temporal model

As mentioned before, spatially sampled biological data tend to be spatially and temporally correlated processes. Consequently, a correct model requires the inclusion of spatial and temporal correlation terms. The way to do so is considering the variable of interest (the CPUE in fisheries) as a spatio-temporally correlated process $Y(w,t)$:

$$Y(w,t) \equiv \{y(w,t) : (w,t) \in D \subseteq \mathbb{R}^2 \times \mathbb{R}\}, \tag{1}$$

where $\mathbb{R}^2$ represents the two dimensional continuous spatial domain and $\mathbb{R}$ the temporal domain. This can also be written in regression terms as:

$$\begin{aligned} Y(w,t) &= \mu(w,t)\beta + Z(w,t) + \epsilon(w,t) \\ Z(w,t) &= \alpha Z(w,t-1) + W(w,t) \end{aligned} \tag{2}$$

where $\epsilon(w,t)$ is a spatially and temporarily uncorrelated Gaussian measurement error (fine scale variability), $\mu(w,t)$ is a deterministic trend surface specified by the model covariates (large-scale variability) and $Z(w,t)$ is an unobserved spatio-temporal Gaussian field (small-scale varibility) composed by an autoregressive temporal effect and a spatially correlated field with covariance function:

$$\begin{aligned} Cov(Z(w,t), Z(w',t')) &= \sigma^2 C(h) \quad \text{for} \quad t = t' \\ Cov(Z(w,t), Z(w',t')) &= 0 \quad \text{for} \quad t \neq t' \end{aligned} \tag{3}$$

where $h = \parallel w - w' \parallel$ is the Euclidean distance between points and $l = \mid t - t' \mid$ is the temporal lag. The covariance function is null for two points observed at different times.

3. Model for semi-continuous data

Another technical issue when dealing with the CPUE is the usual semi-continuous nature of fisheries data that combines a continuous distribution with a point mass at zero. Typically, this type of data has been decomposed into two independent processes [3, 4], the first one being an occurrence process where presence is assigned to observations greater than 0 and absence to observations equal to 0. The second one is an abundance process conditional to presence. Note that using this modelling is like deleting zero observations.

However, a major concern when fitting this sort of data is the fact that the created sub-processes do not necessarily have to be independent. The shared component analysis, introduced in spatial statistics by Knorr-Held and Best [2], allows the joint modelling of related processes.

This way, both predictors can share information by jointly modelling different terms of both linear predictors. The way to do so is now by decomposing Y_{wt} in U_{wt} and V_{wt}, indistinctly the occurrence and abundance processes, using:

$$\begin{array}{rcl} U_{wt} & = & \mu(w,t) + Z(w,t) + \epsilon(w,t) \\ V_{wt} & = & \theta_1\mu(w,t) + \theta_2 Z(w,t) + \theta_3\epsilon(w,t) \end{array} \tag{7}$$

where fitted effects are shared and multiplied by some unknown parameters (θ_1, θ_2 and θ_3) in V_{wt} to scale the effects between both processes. Note that it is not necessary for all effects to be shared, thus there are a number of model structures to compare.

4. Comparing model structures

As previously introduced, the complexity of ecological processes is far beyond the use of covariates to explain our observations. The spatio-temporal correlation and the incapacity to plug a semi-continuous dataset into a single statistical distribution are two examples of such complexity. Considering this and based on the described models, we propose the comparison of a number of model structures (Table 1) and apply it to hake recruitment in the Western Mediterranean as a case study. Model 0 represents the reference model, which assumes fully independent occurrence and abundance models including spatial correlation, an unstructured random effect for time and a previously selected set of covariates. Model 4, conversely, fits shared components for all the effects in the linear predictor including the spatio-temporal correlation. All the sub-models in between model 0 and model 4 have been indexed according to model complexity (see Table 1). All these models have been fitted using the INLA package for R that implements Bayesian analysis based on the Integrated Nested Laplace Approximations (INLA, [5]).

Hake recruitment results show a substantial model fit improvement when adding a temporal correlation term to the spatial field. Moreover, fitted environmental relations also improved by combining information from the occurrence and abundance processes, thus the selected model structure in this case study is Model 2.3 (see Table 1).

5. Bibliography

[1] FAO. *Fisheries management. 2. The ecosystem approach to fisheries*. FAO, 2008.

[2] L. Knorr-Held and N.G. Best. A shared component model for detecting joint and selective clustering of two diseases. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 164(1):73–85, 2001.

[3] D.C. Maravelias. Habitat selection and clustering of a pelagic fish: effects of topography and bathymetry on species dynamics. *Canadian Journal of Fisheries and Aquatic Sciences*, 56(3):437–450, 1999.

[4] I. Paradinas, D.V. Conesa, M.G. Pennino, F. Muñoz, A. Fernández, A. López-Quílez, and J.M. Bellido. A Bayesian spatio-temporal approach to identifying fish nurseries by validating persistence areas. *Marine Ecology Progress Series*, 2015. In press.

	Model structure
Model 0	Independent spatial models + unstructured random effect for time + smooth environmental effects
Model 1	Independent models with different unrelated spatial realizations per year + smooth environmental effects
Model 2.1	Independent spatial models + temporally correlated between consecutive years + smooth environmental effects
Model 3.1	Joint spatial effect + unstructured random effect for time + smooth environmental effects
Model 3.2	Joint environmental effects + unstructured random effect for time + independent spatial effects
Model 3.3	Joint unstructured random effect for time + unstructured random effect for year + smooth environmental effects
Model 3.4	Full joint model without temporal structure
Model 2.2	Joint spatial effect + temporally correlated between consecutive years + smooth environmental effects
Model 2.3	Joint environmental effects + temporally correlated between consecutive years + independent spatial effects
Model 4	Full joint and spatio-temporal model

Table 1: Proposed model structure comparison for ecological processes. Note that all joint effects have a scaling parameter to scale the effect between the occurrence and abundance linear predictors.

[5] H. Rue, S. Martino, and N. Chopin. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the royal statistical society: Series B (statistical methodology)*, 71(2):319–392, 2009.