Environ Ecol Stat (2010) 17:347–376
DOI 10.1007/s10651-009-0111-6

# Modelling spatial zero-inflated continuous data with an exponentially compound Poisson process

**Sophie Ancelet · Marie-Pierre Etienne · Hugues Benoît · Eric Parent**

Received: 8 May 2008 / Revised: 2 March 2009 / Published online: 3 April 2009


**Abstract** A parsimonious model is presented as an alternative to *delta* approaches to modelling zero-inflated continuous data. The data model relies on an exponentially compound Poisson process, also called the law of leaks (LOL). It represents the process of sampling resources that are spatially distributed as Poisson distributed patches, each containing a certain quantity of biomass drawn from an exponential distribution. In an application of the LOL, two latent structures are proposed to account for spatial dependencies between zero values at different scales within a hierarchical Bayesian framework. The LOL is compared to the *delta*-gamma ($\Delta\Gamma$) distribution using bottom-trawl survey data. Results of this case study emphasize that the LOL provides slightly better fits to learning samples with a very high proportion of zero values and small strictly positive abundance data. Additionally, it offers better predictions of validation samples.

**Keywords** Bayes factor · Bayesian hierarchical modelling · Excess zeros · Intrinsic AutoRegressive spatial model · MCMC algorithms · Posterior predictive loss criterion

S. Ancelet (✉)
INSERM-INED U822, Equipe "Epidémiologie de la reproduction et du développement de l'enfant", Hôpital de Bicêtre, 82 rue du Général Leclerc, 94276 Le Kremlin-Bicêtre Cedex, France
e-mail: sophie.ancelet@orange.fr

M.-P. Etienne · E. Parent
AgroParisTech-INRA UMR518, 16 rue Claude Bernard, 75231 Paris Cedex 05, France

H. Benoît
Fisheries and Oceans Canada, 343 Université Avenue, P.O. Box 5030, Moncton, NB E1C 9B6, Canada

## 1 Introduction

Abundance survey data (e.g., counts, biomass) typically contain a large proportion of zeros accompanied by a skewed distribution of the remaining values, including extremes. Martin et al. (2005) recently argued that the analysis of such zero-inflated data (Heilbron 1994) should begin by distinguishing the sources of excess zeros. Three general sources can affect the choice of a model. The most trivial is the inclusion in a data set of observations from areas where there are no chances of making a non-zero observation, such as areas outside the possible environmental range of a species. These cases are easily addressed by excluding them from analysis. Remaining true zero values can occur as a direct result of the effect under study (e.g., suitability of a given intertidal habitat) or as a stochastic result of sampling from areas of low density. On the other hand, false zeros can occur as a result of detection limits, observer effects (e.g., hiding behavior in the presence of observers) or a mistiming of observation (e.g., sampling a site that is normally occupied by a species, just not at the time of observation).

Zero-inflated data create many problems for statistical analysis. First, the sample mean computed from overdispersed data can be an imprecise indicator of stock abundance (Pennington 1996). Second, designing realistic models for such data remains a challenge. For instance, the Poisson and the negative binomial distributions (if the data are discrete) or the lognormal and gamma distributions (in the continuous case) are common models proposed for marine surveys samples. But the spike at zero in the empirical histograms often contains many more zeros than would be expected from these standard distributions. Consequently, these common sense approaches often lead to poor fits (Welsh et al. 1996) since the underlying distributional assumptions (e.g. variance-mean relationship for the Poisson distribution, null probability of zero value in the continuous case) are violated.

Model properties are commonly exploited to define efficient indicators of stock abundance useful to monitor species over time and guide management decisions. Consequently, a fair amount of statistical effort has recently been devoted to dealing with zero-inflated data sets. However, these developments have focused mainly on the modelling of discrete zero-inflated data (Martin et al. 2005; Ridout et al. 1998). Counts with extra zeros can be modelled with the well-known Zero-Inflated Poisson and Negative Binomial regression models (ZIP and ZINB, respectively) (Lambert 1992) or with their respective Random-Effects version (RE-ZIP and RE-ZINB) (Hall 2000). In the continuous case, *delta* models (i.e., conditional or two part-models) are routinely used by ecologists to analyze abundance data for fish and plankton surveys with many zero values (Pennington 1983; Chyan-Huei Lo et al. 1992; Fletcher et al. 2005). These models assume that the presence–absence of a species at a site and its abundance when present result from separate ecological mechanisms. Formally, these mixture models separately define the occurrence of a zero value as a Bernoulli random trial and the positive abundances using either a gamma (Stefansson 1996) or a lognormal (Aitchison and Brown 1957) random variable. They have the attractive advantage of an orthogonal parametrization, which makes them easy to fit and interpret. They offer a powerful framework to analyze and predict the spatial distribution of organisms as they can incorporate a parametric regression to relate the sources of zero observations

and species abundance when present to environmental characteristics. However, the break between zero and non-zero values presents a particularly unnatural discontinuity in species abundance or density data, where many zeros are an important component of decreasing gradients. In addition, there is no change of parameters of *delta* models that coherently matches a change of sampling effort.

In this paper, we propose an alternative model for the analysis of continuous zero-inflated data. Its parsimonious stochastic structure is based on a mixture of distributions: a Poisson random sum of exponential variables. This compound Poisson process, originally coined *la loi des fuites* (*the law of leaks*—LOL) by Bernier and Fandeux (1970), represents the process of sampling resources that have a latent patchy spatial distribution in an homogeneous areal unit (e.g., with environmental conditions). Abundance data result from an appealing representation of data collected from a hidden Poisson sampling process, ensuring spatial coherence with regards to a change of sampling effort. If covariates were known to explain the latent data model variability, the occurrence of zeros could then be interpreted as a natural endpoint of a progression from high to low intensities.

The LOL model is then embedded as a data submodel within a general hierarchical framework to take into account possible between-unit biological or environmental heterogeneity. Indeed, it is well known that physical features and processes (e.g., water depth, currents, winds,. . .) create broad-scale spatial structure in the environment and biological systems such as gradients and large patchy structures separated by discontinuities (Legendre and Legendre 1998). Hierarchical Bayesian modelling distinguishes between the three different stages of a model's hierarchy: a data submodel describing how the data $y$ are collected, a process submodel accounting for spatial covariations (or the dynamics) of the phenomenon through latent variables $z$, and a top structure that quantifies the partial knowledge about the unknown parameters $\theta$. We compared two different process submodels to represent spatial variations at areal unit level. The first one, which we term *regionalized*, consists of a same distribution for area-specific random effects. The idea underlying this submodel is that information is shared among exchangeable areal units. In the second approach, we considered the model proposed by Besag, York and Mollié (called BYM model) (Besag et al. 1991) to introduce local spatial dependencies between neighboring areal units. For a fair and realistic comparison, the competing *delta* models were embedded as a data submodel within similar hierarchical structures.

We computed Bayes factors from a bottom-trawl survey dataset to compare the fitting abilities of competing models. Bayes factors are known to be sensitive to prior choice, and informative priors are recommended to get a meaningful comparison between models. We adapted the frequentist approach of split test sample analysis to easily verify Bayes factor sensitivity to priors in our case study. We also computed the posterior predictive loss criterion proposed by Gelfand and Ghosh (1998) to compare the predictive abilities of competing models. In this paper, we argue that:

1. The LOL conveniently accommodates the presence of a large number of zeros and a skewed distribution of non-zero values.

2. In our case study, the LOL has slightly better fitting abilities than the *delta* models when the abundance data contain a very large proportion of zeros and small strictly positive abundances.
3. In our case study, the LOL predicts the data much better than the commonly used *delta* models.
4. Thanks to Markov Chain Monte-Carlo inferential techniques (MCMC), only few additional computational costs are required when dealing with the two additional latent layers of the LOL (The LOL can itself be viewed as a hierarchical data model).
5. The introduction of a spatial structure on the top of the hierarchy is straightforward. It can improve the robustness of marine resource survey analyses which are frequently characterized by relatively small sample sizes due to the high cost of sampling at sea.

In the following section, we detail the available bottom-trawl survey dataset then, the conceptual and mathematical hypotheses underpinning the LOL model used to represent zero-inflated data. Two possible underlying spatial patterns of abundance data for the process submodel are proposed. In Sect. 3, we briefly describe the Bayesian inference techniques used. Section 4 compares the fitting and predictive abilities of the LOL model with a *delta* model using our case study. Finally, we discuss the advantages and limits of the LOL.

## 2 Hierarchical modelling of spatially structured zero-inflated data

### 2.1 Data description

Scientific bottom-trawl surveys have been conducted by Fisheries and Oceans Canada in the southern Gulf of St. Lawrence (sGSL) (NW Atlantic) each September since 1971 (Hurlbut and Clay 1990). The main objective is to quantify the abundance and the distribution of various marine species.

The survey follows a stratified random design, with stratification based on depth and geographic area. Each year, the number of sampling sites chosen is generally proportional to the size of each stratum. The total number of sites sampled annually has varied from about 65–75 during the 1970s to 140–200 in all but one year since 1989.

The target fishing procedure is a 30-min straight-line tow at 3.5 knots (i.e., 3.21 km trawled distance). However, the actual distance trawled can vary because some tows are shortened to avoid tearing the net on rough bottom and because of variations in vessel speed resulting from prevailing winds and currents. Consequently, it is recorded after the catch as the difference between starting and ending positions of tow.

Collected species are identified, sorted, weighted (in kilogrammes per tow) and some are counted and measured. While the survey was initially targetted at fish, data on the abundance of epibenthic invertebrates such as urchins, starfishes, whelks and anemones have also been collected since the mid 1980s.

Certain environmental covariates are measured at each sampling site: depth, water temperature. Moreover, the type of bottom sediment at each sampling site can be inferred from a geological map of the sGSL, made in 1973 (Loring and Nota 1973).

These covariates are not considered in this study though they were used to refine the sampling stratification of the sGSL in 38 (instead of 27) areal units containing approximately uniform habitats (see Ancelet 2008 for details).

We used data on the abundance of sea urchins (*Strongylocentrotus* sp.) and sea anemones (order Actiniaria) collected from 1999 to 2001, which corresponds to 540 bottom-trawl surveys. The time period was chosen so as to minimize the impact of interannual changes in abundance on our analyses. These species were chosen for two main reasons. First, they are relatively sessile with very negligible interannual movements on the scale of the survey. Secondly, they have clearly different spatial distributions in the sGSL (Fig. 1). Urchins are present in most of the sGSL except where the grazing opportunities for this herbivore are limited by depth (e.g., in the northern part of the survey area called Laurentian channel, >200 m depth) or sediment type (e.g., silty areas in south and east of Prince Edward Island). In contrast, anemones have a more restricted distribution in the sGSL. Large catches mainly occur in the Laurentian Channel. The empirical distribution of the abundances of anemones contains many more zeros and is characterized by smaller strictly positive abundances compared to the distribution for urchins (Fig. 2). Like the majority of marine organisms, both species are distributed in patches of localized variable abundance, interspersed by numerous and relatively large areas of absence (Fig. 1). The data can be obtained on request from the third author (email: hugues.benoit@dfo-mpo.gc.ca).

## 2.2 The traditional *delta* models

Let $\{y_k; k = 1, 2, 3, \ldots, r\}$ denote quantities of biomass measured in $r$ sampling sites located in a survey area D and $\{S_k; k = 1, 2, 3, \ldots, r\}$ the corresponding sampling effort (i.e., swept area as in our case study but, in other contexts, could also be volume filtered, observation time, ...). The records of $r$ independent sampling events contain strictly positive continuous values but zero values can also occur. Zero-inflated continuous data sets are often modelled using *delta* models. *Delta* models are

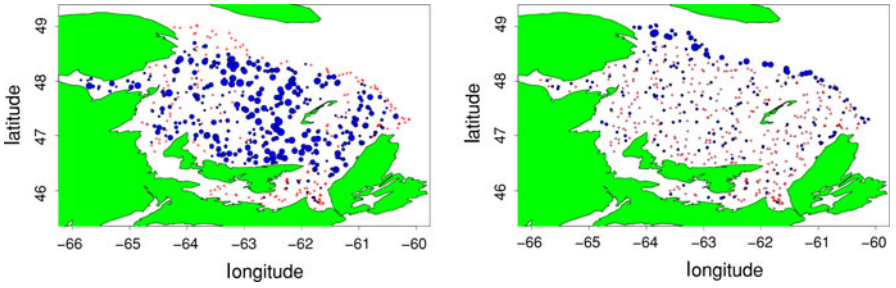


**Fig. 1** The spatial distribution of sea urchins (*left*) and sea anemones (*right*) biomass from individual tows in the sGSL bottom-trawl survey, 1999–2000–2001. The "*" denotes the sites of null catch whereas the radii of the circles are proportional to the biomass caught (in kg/tow)

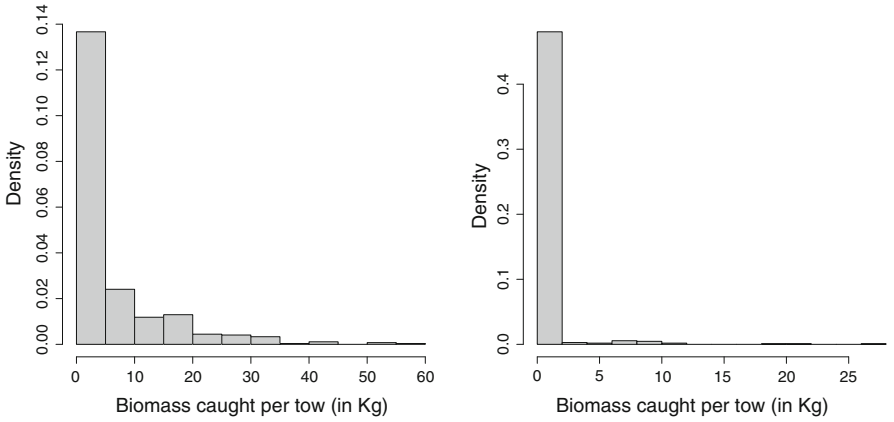


**Fig. 2** Two examples of zero-inflated datasets: the biomass (kg/tow) of sea urchins (*left*) and sea anemones (*right*) from individual tows in the sGSL bottom-trawl survey, 1999–2000–2001

named after the Dirac function at zero, modelling the occurrence of a zero value with a Bernoulli random variable. The other component gives the strictly positive abundances using either a gamma or a lognormal distribution after rescaling the data $y_k$ by the catch effort $S_k$ ($k = 1, 2, \ldots, r$). The cumulative distribution function (c.d.f) of the abundance $Y_k$ ($Y_k \geq 0$) is:

$$[Y_k \leq y_k] = \delta + (1 - \delta) G_\lambda \left( \frac{y_k}{S_k} \right)$$

where:

- $\delta$ ($0 \leq \delta \leq 1$) denotes the probability of null catch (i.e., of zero data) per unit of area.
- $G_\lambda$ is a continuous c.d.f. (gamma or lognormal) defined by a set of unknown parameters $\lambda$ and describing the abundance of strictly positive values per unit of area.

In this paper, the notation [ ] means either a distribution function for discrete variable or a density for continuous ones (Gelfand and Smith 1990).

In what follows, $G_\lambda$ denotes a gamma cumulative distribution function. Actually, we followed the results of Myers and Pepin (1990) which suggested that the use of a gamma density is preferable to the use of a lognormal density for fisheries data especially when there is a considerable probability of small observations as this is the case for the urchins and anemones of the sGSL. A shape parameter $\alpha > 0$ and a rate parameter $\beta > 0$ are added by the gamma density $G_\lambda$ (i.e., $\lambda = (\alpha, \beta)$):

$$[Y_k = 0] = \delta$$
$$[Y_k = y_k] = (1 - \delta) \frac{\beta^\alpha}{\Gamma(\alpha)} \left( \frac{y_k}{S_k} \right)^{\alpha - 1} \exp\left( -\beta \left( \frac{y_k}{S_k} \right) \right) \quad \text{with } y_k > 0$$

Consequently, three unknown quantities $(\alpha, \beta, \delta)$ characterize the random mechanism of data occurrence for this particular *delta* model called *delta*-gamma model ($\Delta\Gamma$ model).

A property of towing a net is that the same overall distribution would be expected to hold for a long tow as for a short tow. Unfortunately, it is not the case with the $\Delta\Gamma$ distribution. This can be simply seen by considering the characteristic function of the $\Delta\Gamma$ distribution, given by:

$$\varphi^{\Delta\Gamma}(t) = \mathbb{E}(e^{itY_k}|\alpha, \beta, \delta) = \delta + (1-\delta)\frac{1}{\left(1 - it\frac{S_k}{\beta}\right)^{\alpha}}$$

It is clear that the probability distribution of the sum of two $\Delta\Gamma$ variables is not a $\Delta\Gamma$ distribution but a quite complex one for which analytical results are not trivial to obtain. The lack of such additivity property is a major drawback when working of the $\Delta\Gamma$ distribution (Stefansson 1996). A preliminary standardization of data is needed to consider biomass abundances rescaled to the same sampling effort. Therefore, when tow durations are very dispersed, this standardization produces an artificial increase of the number of zero values and indicators of stock abundance are then biased.

## 2.3 A new competing data model: the LOL

The LOL belongs to the class of compound Poisson processes until now used to model continuous-time stochastic processes. It was initially designed to model losses from French gas pipelines (Bernier and Fandeux 1970). Exponential gas intensities would flow out from Poisson distributed holes all along a pipe and only the sum of the local contributions was measured. It is also commonly used in the insurance industry, assuming exponentially-distributed damage events occurring as a temporal Poisson process. In what follows, we propose to extend the LOL model to represent spatial stochastic processes by pointing out an ecological analogy: a latent Poisson sampling process collecting patches of organisms that each have an associated exponential mass. Formally, the LOL can be presented under a hierarchical setting.

### *2.3.1 The underlying process submodel with latent patches*

Conceptually, the LOL describes the process involved in sampling many of the living organisms that are the focus of ecological analysis. For example, imagine one bottom-trawl survey that consists in sweeping the sea floor with a large fishing net to collect organisms of a given species. The model assumes there are patches of organisms to be collected. Patches are drawn from an homogeneous Poisson process so that the random number $N_k$ of patches collected during the sampling event k is given by:

$$N_k \sim Poisson(S_k\mu)$$

where $\mu$ is the expected number of patches for a unit of sampling effort. The latent $N_k$ are independent but not identically distributed Poisson variables depending on the corresponding sampling effort $S_k$.

Each patch $p_k = 1, \ldots, N_k$ contains a certain biomass $M_{p_k}$. We assumes that the $M_{p_k}$ are independent and exponentially distributed with parameter $\rho$ such that $\mathbb{E}(M_{p_k}) = \frac{1}{\rho}$:

$$M_{p_k} \overset{i.i.d}{\sim} Exp(\rho), \quad p_k = 1, \ldots, N_k$$

The exponential distribution is chosen for both reasons of parsimony and because of a conjugate property that simplifies Bayesian inference of the model.

### *2.3.2 The data submodel*

The sum of the individual patches captured by the trawl yields the total observed sample biomass $Y_k = \sum_{p_k=1}^{N_k} M_{p_k}$ (see Fig. 3). By definition, an absence of patches (i.e., $N_k = 0$) offers a zero value and the occurrence of at least one patch (i.e., $N_k \geq 1$) produces a strictly positive outcome, distributed as the random sum of independent exponential variables (i.e., gamma pdf). As defined, the LOL model belongs to the class of compound Poisson processes.

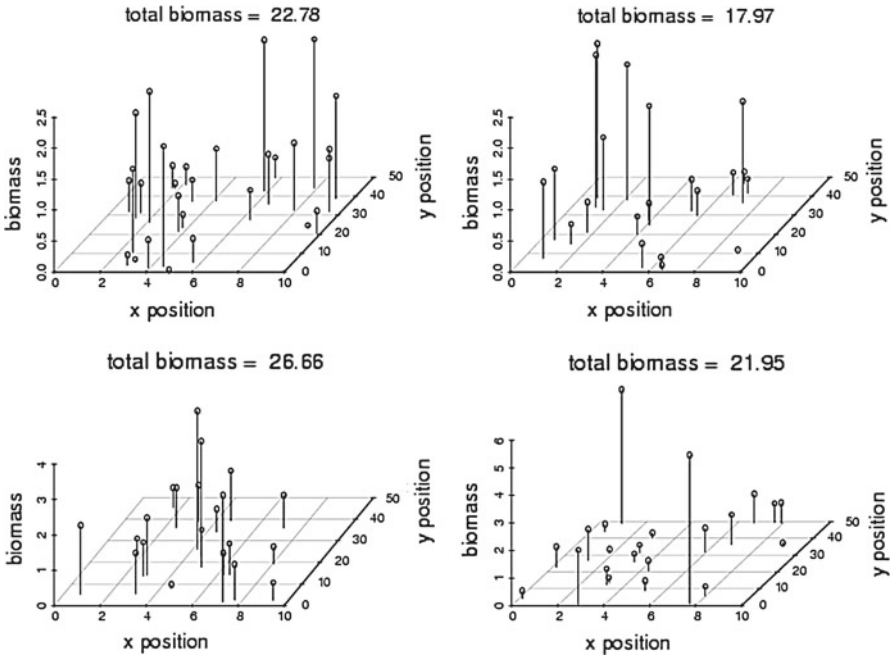


**Fig. 3** Four example of individual bottom-trawl survey tows. The LOL is obtained by harvesting a marked Poisson point process

Conditionally on $\rho$ and the unknown number of patches $N_k$ ($k = 1, 2, \ldots, r$), we see that:

$$Y_k|N_k, \rho \sim \begin{cases} \Gamma(N_k, \rho) & \text{if } N_k > 0 \\ \delta_0 & \text{if } N_k = 0 \end{cases}$$

where the scale parameter $\rho$ is interpreted as the inverse of average biomass in each patch. The above equation shows the explicitation of the latent variables $M_{p_k}$ makes easier the conditional description of the LOL model. However, their knowledge is not necessary to define the distribution of the abundance data $Y_k$ ($k = 1, 2, \ldots, r$) because, working conditionally on $\rho$ and $\mu$ only, we integrate out the latent $N_k$ to obtain the pdf for the LOL:

$$[Y_k = y_k|\mu, \rho] = \begin{cases} \sum_{n=1}^{\infty} \left(e^{-S_k\mu} \frac{(\mu S_k)^n}{n!}\right) \frac{\rho^n}{\Gamma(n)} y_k^{n-1} e^{-\rho y_k} & \text{if } y_k > 0 \\ e^{-S_k\mu} & \text{if } y_k = 0 \end{cases} \quad (1)$$

The zero occurrence (no patch) is only a function of the unknown parameter $\mu$. In contrast to the $\Delta\Gamma$ model, only two unknown quantities $(\rho, \mu)$ are necessary to describe the random mechanism of data occurrence.

### *2.3.3 Model properties*

In this subsection, we briefly sum up some interesting properties of the LOL model. The characteristic function of the LOL is:

$$\varphi^{\text{LOL}}(t) = \mathbb{E}(e^{itY_k}|\mu, \rho) = e^{\frac{it\mu S_k}{\rho - it}} \quad (2)$$

Common quantities of interest for ecologists are easily derived. The probability of getting a zero value at site $k$ is given by $e^{-S_k\mu}$ and the pointwise mean and variance of the abundance linearly depends on the sampling effort $S_k$:

$$\mathbb{E}(Y_k|\rho, \mu) = \frac{\mu S_k}{\rho}$$
$$\mathbb{V}ar(Y_k|\rho, \mu) = \frac{2\mu S_k}{\rho^2}$$

The coefficient of variation $\sqrt{\frac{2}{\mu S_k}}$ will desirably increase towards infinity as the expected number of patches tends to zero.

To represent the variability of each capture event, the LOL assumes that, within a survey area, the occurrence of individuals of a species follows an homogeneous Poisson point process (with exponential marks). As a consequence, the model inherits a Poisson spatial coherence with regards to a change of sampling effort. Imagine that the sampling effort goes from $S_1$ to $\widetilde{S} = S_1 \cup S_2$, with $S_1 \cap S_2 = \varnothing$. The events $Y_1$ and $Y_2$ made in a same survey area occur independently on $S_1$ and $S_2$, and Eq. 2 shows

the total catch event $Y' = Y_1 + Y_2$ also follows a LOL distribution. More generally, the LOL model is stable by addition and is a well known member, as a compound Poisson process, of the class of distributions with infinite divisibility property (Feller 1971). Therefore, contrary to the $\Delta\Gamma$ distribution, the LOL enables to work with raw data directly. Each survey set corresponds to a sampling event according to an homogeneous Poisson point process: the greater the sampling effort, the greater the number of collected patches.

Moreover, the LOL offers a natural representation of data collected along a gradient, where zeros are often a natural endpoint of a progression from high to low abundances. Contrary to *delta* models with a lognormal or a gamma non-zero component, the probability density function of Eq. 1 tends to a strictly positive value at zero (namely $\mu\rho e^{-\mu}$). Hence, some interplay between parameters $\mu$ (controlling the probability of absence) and $\rho$ ensures a much smoother link between the zero and non-zero values than the *delta* models characterized by an abrupt shift modelled by a Dirac function.

## 2.4 Accounting for spatial dependencies between areal units

As defined in the previous sections, both the LOL and $\Delta\Gamma$ models assume that sampling events are independent and identically distributed (i.i.d) random experiments. This strong hypothesis means that the survey area D must be small enough to encompass a rather uniform habitat, suitable for the studied species and characterized by a non-structured spatial distribution of patches at the (micro) scale of a tow. In practice, this hypothesis is often violated. Additionally, the models ignore the ubiquitous observation that the distribution of organisms is structured at numerous spatial scales (Legendre and Legendre 1998).

We now deal with the situation where the survey area is split into I known areal units, each small enough to encompass a rather homogeneous habitat. For instance, in our case study, the sGSL has been divided into 38 homogeneous strata (Sect. 2.1).

A first basic idea consists in leading inference of the model on each areal unit $i$ ($i = 1, 2, \ldots, I$) separately. The weakness of this method is obvious: in small areal units with few samples only, the quality of the estimation will be very poor.

But all areal units share a common feature: they are located in the same ecosystem. We might therefore expect that their properties present generally similar behavior. In the following sections, we develop a hierarchical model aimed at sharing information among units. Two cases are considered, a regionalized version where information is shared between all units and one in which information is shared only amongst neighboring units. These structures are applied to both LOL and $\Delta\Gamma$ models.

### *2.4.1 A regionalized structure to borrow strength from exchangeable areal units*

The LOL and $\Delta\Gamma$ submodels yield the total amount of biomass collected, $Y_{i,k}$, at site k located in areal unit $i$ ($i = 1, 2, \ldots, I$). Both the LOL and $\Delta\Gamma$ distribution are defined by area specific sets of parameters: $(\mu_i, \rho_i)$ and $(\delta_i, \alpha_i, \beta_i)$, respectively.

A simple structure to take into account a similar behavior between areal units is a regionalized version of the data submodels in which an upper level is added to the basic models. The parameters of the data submodels called $z_i$ $(i = 1, 2, \ldots, I)$ are assumed to stem from an i.i.d common *regional* distribution $H$ describing the various degrees of resemblance between sites. We obtain a non linear mixed model:

$$[z_i|\theta] \sim^{i.i.d} H(\theta)$$

For instance, the LOL with $z_i = (\mu_i, \rho_i)$ can be regionalized (model called $R_{\mu,\rho}$-LOL) via a gamma structure as a convenient regional distribution $H$ with $\theta = (a, b, c, d)$. For areal unit $i$ taken at random in the survey area **D**, the expected number of patches $\mu_i$ and the expected biomass contained in each patch $1/\rho_i$ are random effects following a gamma distribution of parameters $(a, b)$ and $(c, d)$, respectively. The four letters are first level parameters. The Directed Acyclic Graph in Fig. 4 suggests a simple representation of the $R_{\mu,\rho}$-LOL model. For each areal unit $i$, the expected number of patches to be collected is $a/b$ and the inverse of the expected biomass contained in one patch is $c/d$. Dissimilarity between all areal units is determined by the dispersion among the $z_i$. The smaller the variances $a/b^2$ and $c/d^2$ will be, the more similar the areal units will become.

For the $\Delta\Gamma$ distribution of Sect. 2.2, a regional distribution has to be put on a three dimensional vector $z_i = (\alpha_i, \beta_i, \delta_i)$. We chose to draw $logit(\delta_i) = \log(\frac{\delta_i}{1-\delta_i})$ from a normal distribution and $\beta_i$ from a gamma distribution (model called $R_{\delta,\beta}$-$\Delta\Gamma$). For a fair comparison keeping the same number of parameters among models, we decided in what follows to set the coefficients of variation of the gamma distributions to a constant $\alpha_i = \tilde{\alpha}$ for all $i = 1, 2, \ldots, I$, which is not an unrealistic assumption from an ecological point of view. It means that the coefficient of variation of strictly positive abundances is constant whatever the areal unit.

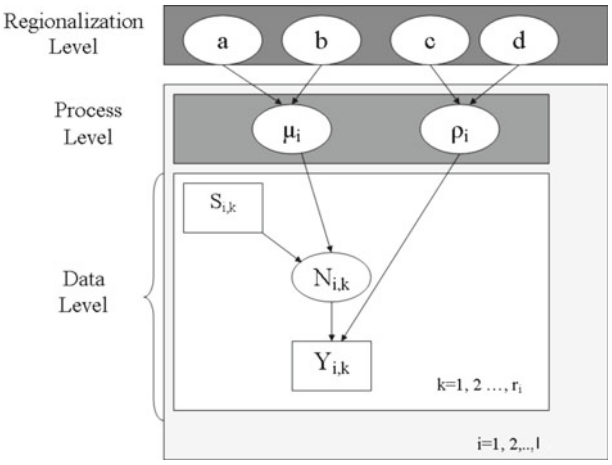


**Fig. 4** Directed Acyclic Graph for the $R_{\mu,\rho}$-LOL model with regionalization structure. The square nodes represent observed data and the circles refer to latent variables or parameters. Arrows depict stochastic dependence between nodes

Regionalized model structures take advantage of global similarity between areal units regardless of the distance that separates them. However, one could argue that spatially more proximate units should have a more similar behavior. The idea that vicinity matters is developed in the following section using a BYM model which includes an intrinsic Gaussian conditional autoregressive component to impose a spatial structure.

### *2.4.2 A BYM model to link neighboring areal units*

Another approach to model the spatial variations between areal units is to assume that mainly neighboring units have a similar behavior. To represent spatially correlated random effects $z_i$ ($i = 1, 2, \ldots, I$), we used the following model, originally proposed by Besag et al. (1991) and usually referred as BYM model:

$$g(z_i) = m_0 + \Phi_i + \varepsilon_i$$

In the above expression:

- $g$ is an univariate link function of the latent vector $z = (z_1, z_2, \ldots, z_I)$
- $m_0$ is a constant term. It represents an overall average effect.
- $\Phi_i$ is the spatially structured component of the model through an intrinsic Gaussian conditional autoregressive model (IAR) (Banerjee et al. 2004) on $\Phi = (\phi_1, \phi_2, \ldots, \phi_I)$:

$$\left[\Phi_i | \Phi_j,\, j \sim i\right] \propto \exp\left\{ -\frac{\left(\Phi_i - \frac{1}{n_i}\Sigma_{j\sim i}\Phi_j\right)^2}{2\frac{s^2_{\text{IAR}}}{n_i}} \right\}.$$

  where $i \sim j$ means areal unit $i$ and areal unit $j$ are neighboring, $n_i$ denotes the number of adjacent areal units to unit $i$ and $s^2_{\text{IAR}}$ is the local standard deviation. We suppose two areal units sharing a common boundary are neighbors.
- $\varepsilon_i$ captures a residual unstructured heterogeneity using a normal pdf:

$$(\varepsilon_i)_{1\le i\le I} \overset{i.i.d}{\sim} \mathcal{N}(0, s^2_\varepsilon).$$

The IAR structure is improper (the sum of the $\Phi_i$ must be centered so that the conditional distributions make sense to define a joint pdf on $\mathbb{R}^{n-1}$), but we will see in the next section that, since it is used only as a top level structure (prior) in a hierarchical Bayesian setting, the inference does not present any difficulty and the posterior will be proper. Each time a component of $z$ is modelled by a BYM structure through a univariate link function $g$, three top-level parameters ($m_0, s^2_\varepsilon, s^2_{\text{IAR}}$) have to be inferred. It is worth noting that when $s^2_{\text{IAR}} = 0$, the BYM structure yields a regionalized model with the normal pdf as a regional distribution $H$ on the $g(z_i)$.

If the LOL model is used as data submodel, one may wish that the expected numbers of patches in areal unit i, given by $\mu_i$, ressembles those in neighboring units. Through the univariate link function $g(z_i) = \log(\mu_i)$, a BYM structure can be added on the

**Table 1** Ten competing hierarchical models based on the LOL and the $\Delta\Gamma$ distribution

| Spatial structure/data submodel | LOL $z_i = (\mu_i, \rho_i)$ | $\Delta\Gamma$ $z_i = (\delta_i, \alpha_i, \beta_i)$ |
|---|---|---|
| No spatial effect (I) | $(\text{LOL})^{\otimes I}$ | $(\Delta\Gamma)^{\otimes I}$ |
| Neighbors are exchangeable (R) | $R_{\mu,\rho}$-LOL | $R_{\delta,\beta}$-$\Delta\Gamma$ |
| Neighbors are partially exchangeable (PR) | $R_{\mu}$-LOL | $R_{\delta}$-$\Delta\Gamma$ |
| Neighbors are spatially correlated but partially (PS) | $BYM_{\mu}$-LOL | $BYM_{\delta}$-$\Delta\Gamma$ |
| Neighbors are spatially correlated and regionalized (S + R) | $BYM_{\mu}R_{\rho}$-LOL | $BYM_{\delta}R_{\beta}$-$\Delta\Gamma$ |

latent vector $\mu = (\mu_1, \mu_2, \ldots, \mu_i)$. Modelling $\rho$ with a spatial structure instead of $\mu$ or modelling both $\rho$ and $\mu$ as a multivariate BYM is discussed later.

A BYM layer can also be added to a *delta* model. For the $\Delta\Gamma$ distribution with the three dimensional vector $z_i = (\alpha_i, \beta_i, \delta_i)$ (Sect. 2.2), spatial structure can be added on the probability of a zero e.g., using the simple link function $g(z_i) = logit(\delta_i)$ or it could be added on the expected non-zero biomass in areal unit $i$ e.g., writing $\log\left(\frac{\alpha_i}{\beta_i}\right) = m_0 + \Phi_i + \varepsilon_i$. Additionally, more sophisticated models could also rely on a multivariate BYM structure.

## 2.5 Competing hierarchical constructions

Depending on the number of parameters in each data submodel to be spatialized, the chosen link functions and the hypothetized regional distribution, many hierarchical constructions can be specified. Table 1 sums up the 10 combinations for the LOL and $\Delta\Gamma$ submodels, with or without spatial structure, which are compared in Sect. 4.

Five variants of the LOL were considered (Table 1): the LOL independently applied to each areal unit (called $(\text{LOL})^{\otimes I}$), the $R_{\mu,\rho}$-LOL model (see Sect. 2.4.1), a partially regionalized version of LOL (called $R_{\mu}$-LOL) in which the latent variables $\mu_i$ (i =1, ..., $I$) are i.i.d following a gamma distribution and $\rho_i = \widetilde{\rho}$ from 1 to I, a $BYM_{\mu}$-LOL model in which the latent variables $\mu_i$ follow a BYM model (Sect. 2.4.2) and $\rho_i = \widetilde{\rho}$ from 1 to I and a $BYM_{\mu}R_{\rho}$-LOL model in which the latent variables $\mu_i$ follow a BYM model and the $\rho_i$ are i.i.d following a gamma regional distribution.

Five models based on similar structures have been considered for the $\Delta\Gamma$ distribution (Table 1): the $\Delta\Gamma$ distribution independently applied to each stratum (called $(\Delta G)^{\otimes I}$), the $R_{\delta,\beta}$-$\Delta\Gamma$ model (Sect. 2.4.1), a partially regionalized version (called $R_{\delta}$-$\Delta\Gamma$) in which logit($\delta_i$) are i.i.d following a normal distribution and $\beta_i = \tilde{\beta}$ for all i from 1 to I, a spatialized version (called $BYM_{\delta}$-$\Delta\Gamma$) in which the logit($\delta_i$) follows a BYM model and $\beta_i = \tilde{\beta}$ from 1 to I and a second spatialized version (called $BYM_{\delta}R_{\beta}$-$\Delta\Gamma$) in which the logit($\delta_i$) follows a BYM model and $\beta_i$ are i.i.d following a gamma regional distribution.

## 3 Bayesian inference

The inference for the family of models relying on the LOL or on the $\Delta\Gamma$ distribution has been developed under the Bayesian paradigm. This setting offers the possibility of accounting for external qualitative information through the prior, the commonsense intuitive interpretation of posterior statements in terms of probabilistic bets and the coherence with probability theory when deriving predictive judgments by integrating out nuisance parameters. Thanks to Markov Chain Monte Carlo (MCMC) sampling algorithms—the Gibbs sampler and/or the Metropolis Hastings algorithm (Robert and Casella 2004), the Bayesian framework is also particularly appropriate for working with hierarchical models such as the ones we presented here (Banerjee et al. 2004).

The Bayesian model specification requires prior distributions. In our case, it is difficult to produce informative priors for *theoretical* quantities such as an expected number of collected patches during a sampling event given that the definition of what consists of a patch depends on scale. Consequently, a simple approach was to consider flat priors. When the LOL was applied independently to each stratum, $\mu_i$ and $\rho_i$ $(i = 1, 2, \ldots, 38)$ were independent with a flat normal prior truncated at 0. For the regionalized versions of the LOL (Sect. 2.4.1), the unknown parameters a, b, c and d all followed a gamma prior with 0.01 as shape and rate parameters. Finally, for the BYM$_\mu$-LOL and BYM$_\mu$R$_\rho$-LOL models, we used the priors recommended in Banerjee et al. (2004): the precision $\frac{1}{s_\epsilon^2}$ of the non-spatialized residual component (Sect. 2.4.2) with a gamma prior with 0.001 as shape and rate parameters and the local precision of the IAR model $\tau_{\text{IAR}} = \frac{1}{s_{\text{IAR}}^2}$ following a gamma prior with 0.1 as shape and rate parameters.

Models fitting was performed using the software OpenBUGS and the BRugs package in R (R-Project for Statistical Computing, Version 2.4.1). OpenBUGS is a useful and well documented tool (Spiegelhalter et al. 2007; Congdon 2001) for Bayesian analysis of complex statistical models using MCMC techniques. It is worth mentioning that the introduction of an additional spatial prior for the expected numbers of patches $\mu_i$ $(i = 1, 2, \ldots, I)$ has not to be paid much when inferring the unknowns of the LOL model. In particular, the BYM prior can be easily implemented thanks to GeoBUGS (Thomas et al. 2007), an add-on to OpenBUGS that fits spatial models.

We ran each model of Table 1 for 200,000 cycles with a burn-in period of 50,000 cycles and a thinning of 100 cycles. We checked the convergence of MCMC algorithms by computing Brooks and Gelman statistics (Brooks and Gelman 1998) thanks to the R package Coda.

## 4 Case study: bottom-trawl survey data

In this section, we report results from the analyses of the data from the bottom-trawl survey of the sGSL. As detailed in Sect. 2.1, we used zero-inflated data collected from 1999 to 2001. Additionally, we assumed that each of the 38 strata defined in the sGSL encompasses approximately uniform habitat.

The goals of our analysis were:

1. to select a model by comparing the fitting and predictive abilities of the $(\text{LOL})^{\otimes I}$, $\text{R}_{\mu,\rho}$-LOL, $\text{R}_{\mu}$-LOL, $\text{BYM}_{\mu}$-LOL and $\text{BYM}_{\mu}\text{R}_{\rho}$-LOL models with the corresponding versions of the $\Delta\Gamma$ model (see Table 1).
2. to analyse more precisely the results obtained with the best structure: the $\text{BYM}_{\mu}\text{R}_{\rho}$-LOL model.

### 4.1 Comparing the fitting abilities

Generally speaking, it is difficult to define meaningful criteria to compare hierarchical Bayesian models. Even if MCMC sampling techniques make it easy to compute, the Deviance Information Criterion (DIC) proposed by Spiegelhalter et al. (2002) has no clear statistical interpretation (except for normal linear models) and depends on the choice of parametrization. Instead we have focused on the Bayes factors which do not have these major drawbacks. Bayes factors were computed from asymptotic approximations as proposed by Kass and Raftery (1994). More details are given in Appendix 1.

To get a meaningful comparison between models, we computed partial Bayes factors from informative priors by performing a split sample test analysis. Appendix 2 details an ingenious method to perform quickly a Bayes factor sensitivity analysis to priors. The sequence of survey years was ignored since we were also ignoring temporal trends in abundance. Consequently, there were seven ways to split the sample of 3 years, so as to define a learning sample to get a proper pdf for the unknowns, to serve as a prior on the remaining sample:

– empty learning sample (non-informative priors as defined in the previous section),
– one year of data to learn (3 possibilities: either 1999 or 2000 or 2001), yielding an informative state of knowledge for the unknows and 2 years to compare models,
– two years of data to learn (3 possibilities: 2 years to be taken among 3) and the remaining year as a validation data set. This case corresponds to very informative priors.

In order to obtain meaningful partial Bayes factors, we used 2 years of data (about 66% of the datasets) to learn and define the prior distributions. The diagonal lines of Tables 2 and 3 show the relative credibility of similar versions of each model applied to urchins and anemones abundance data, respectively, based on validation samples consisting of single years (either 1999 or 2000 or 2001). Following Kass and Raftery's ideas (Kass and Raftery 1994), the Bayes factors are indicated at $(2 \times \log)$ scale and a value greater than 2 means that the tested LOL version has significantly better fitting performances than the similar $\Delta\Gamma$ version. On the contrary, a negative Bayes factor, lower than $-2$, means that the tested $\Delta\Gamma$ version has significantly better fitting performances than the similar LOL version.

For both urchins and anemones, the partially regionalized version $R_{\mu}$-LOL and the spatial version $\text{BYM}_{\mu}$-LOL fitted to the data better than the analogous $\Delta\Gamma$ models (Tables 2 and 3). Basically, the hypothesis $\beta_i = \widetilde{\beta}$ $(i = 1, 2, \ldots, 38)$ of the $\text{R}_{\delta}$-$\Delta\Gamma$ and $\text{BYM}_{\delta}$-$\Delta\Gamma$ models (Sect. 2.5) resulted in overly strong smoothing of the average

**Table 2** Urchins: partial Bayes factors (at (2× log) scale) computed from the validation samples consisting of single years (1999, 2000 or 2001)

| ↱ | $(\Delta\Gamma)^{\otimes 38}$ | $R_{\delta,\beta}$-$\Delta\Gamma$ | $R_{\delta}$-$\Delta\Gamma$ | $BYM_{\delta}$-$\Delta\Gamma$ | $BYM_{\delta}R_{\beta}$-$\Delta\Gamma$ |
|---|---|---|---|---|---|
| | 3.88 | | | | |
| $(LOL)^{\otimes 38}$ | −4.93 | | | | |
| | −5.53 | | | | |
| | | −3.66 | | | |
| $R_{\mu,\rho}$-LOL | | −2.51 | | | |
| | | −7.08 | | | |
| | | | 3.49 | | |
| $R_{\mu} - LOL$ | | | 13.05 | | |
| | | | 4.64 | | |
| | | | | 1.42 | |
| $BYM_{\mu}$-LOL | | | | 11.36 | |
| | | | | 4.78 | |
| | | | | | −4.41 |
| $BYM_{\mu}R_{\rho}$-LOL | | | | | −3.79 |
| | | | | | −7.01 |

**Table 3** Anemones: partial Bayes factors (at (2× log) scale) computed from the validation samples consisting of single years (1999, 2000 or 2001)

| ↱ | $(\Delta\Gamma)^{\otimes 38}$ | $R_{\delta,\beta}$-$\Delta\Gamma$ | $R_{\delta}$-$\Delta\Gamma$ | $BYM_{\delta}$-$\Delta\Gamma$ | $BYM_{\delta}R_{\beta}$-$\Delta\Gamma$ |
|---|---|---|---|---|---|
| | −11.04 | | | | |
| $(LOL)^{\otimes 38}$ | 3.28 | | | | |
| | −4.98 | | | | |
| | | −0.66 | | | |
| $R_{\mu,\rho}$-LOL | | 3.29 | | | |
| | | 0.23 | | | |
| | | | −50.04 | | |
| $R_{\mu}$-LOL | | | 36.63 | | |
| | | | 16.82 | | |
| | | | | −51.51 | |
| $BYM_{\mu}$-LOL | | | | 26.85 | |
| | | | | 19.22 | |
| | | | | | −0.42 |
| $BYM_{\mu}R_{\rho}$-LOL | | | | | 8.41 |
| | | | | | 0.49 |

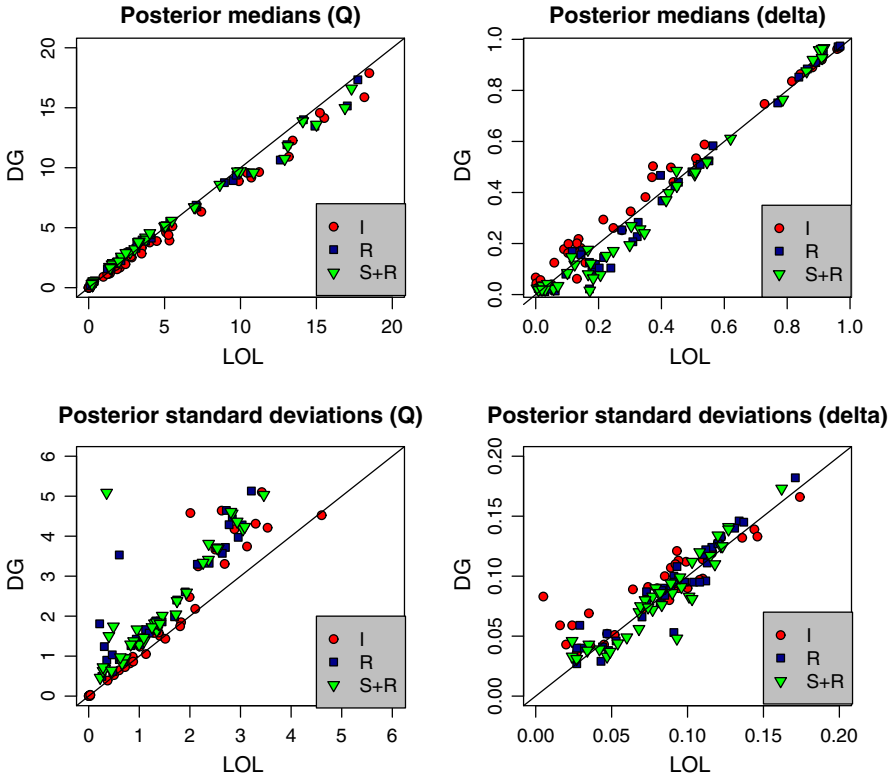


**Fig. 5** Urchins: Posterior medians (*top*) and posterior standard deviations (*bottom*) of the average biomass estimates $Q_i$ ($i = 1, 2, \ldots, 38$) and proportion of zero values estimates $\delta_i$ ($i = 1, 2, \ldots, 38$) from modelling the zero-inflated abundance data with similar hierarchical versions of LOL and $\Delta\Gamma$ (here referred as DG) model. I: no spatial effect, R: Exchangeable neighbors, S+R: Neighbors are spatially correlated and regionalized

abundance estimates compared to the hypothesis $\rho_i = \widetilde{\rho}$ for the analogous LOL versions. Consequently, the LOL versions provided average biomass estimates that were closer to the empirical average biomass compared to $\Delta\Gamma$ versions (results not shown).

The independent version of the LOL model globally provided a poorer fit than the independent version of the $\Delta\Gamma$. For both urchins and anemones, the partial Bayes factors were significantly negative for two validation samples among three. Additionally, Figs. 5 and 6 indicate that the $(\mathrm{LOL})^{\otimes 38}$ model tended to slightly underestimate the proportions of zero values, given by $\delta_i$ ($i = 1, 2, \ldots, 38$), and to overestimate the average biomass, given by $Q_i$, compared to the $(\Delta\Gamma)^{\otimes 38}$ model. Moreover, the posterior standard deviations often were greater for the $(\mathrm{LOL})^{\otimes 38}$ model meaning that it provided less precise estimations compared to the $(\Delta\Gamma)^{\otimes 38}$ model.

The $\mathrm{R}_{\mu,\rho}$-LOL and $\mathrm{BYM}_{\mu}\mathrm{R}_{\rho}$-LOL models provided poorer fits to the urchin abundance data compared to the analogous $\Delta\Gamma$ models (Table 2). The partial Bayes factors were significantly negative for all validation samples. The LOL versions tended to overestimate both the empirical small proportions of zero values and the high

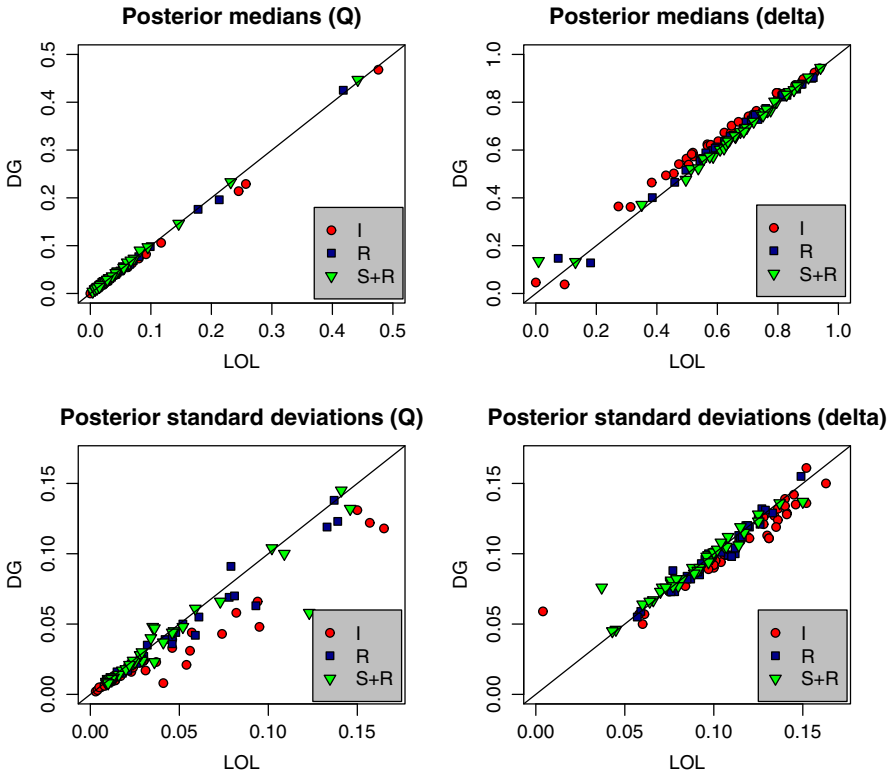


**Fig. 6** Anemones: Posterior medians (*top*) and posterior standard deviations (*bottom*) of the average biomass estimates $Q_i$ $(i = 1, 2, \ldots, 38)$ and proportion of zero values estimates $\delta_i$ $(i = 1, 2, \ldots, 38)$ from modelling the zero-inflated abundance data with similar hierarchical versions of LOL and $\Delta\Gamma$ model (here referred as DG). I: no spatial effect, R: Exchangeable neighbors, S+R: Neighbors are spatially correlated and regionalized

empirical average abundances compared to the $\Delta\Gamma$ models (results not shown). On the contrary, the $\mathrm{R}_{\mu,\rho}$-LOL and $\mathrm{BYM}_{\mu}\mathrm{R}_{\rho}$-LOL models provided superior fits to data on the abundance of anemones than the analogous $\Delta\Gamma$ models (Table 3). The partial Bayes factors were close to 0 for two validation samples among three and greater than 2 for the validation sample based on data collected in 2000. Figure 6 shows that the posterior medians and posterior standard deviations globally matched for the average abundances and proportions of zero for analogous versions of LOL and $\Delta\Gamma$ models.

Computed DIC values associated with each competing model (not shown) globally confirmed all our results.

To check Bayes factors sensitivity to priors, we performed a partial robustness analysis to prior specification. In order to smooth the numerical instability due to the computation of harmonic means (see Appendix 1 Eq. 11), we monitored the evolution of the Bayes factors from an initial sample of 200,000 iterations to a sample of 500,000 iterations by successively increasing the former by 1,500 additional iterations. The sensitivity of Bayes factors both to prior choice and to the corresponding

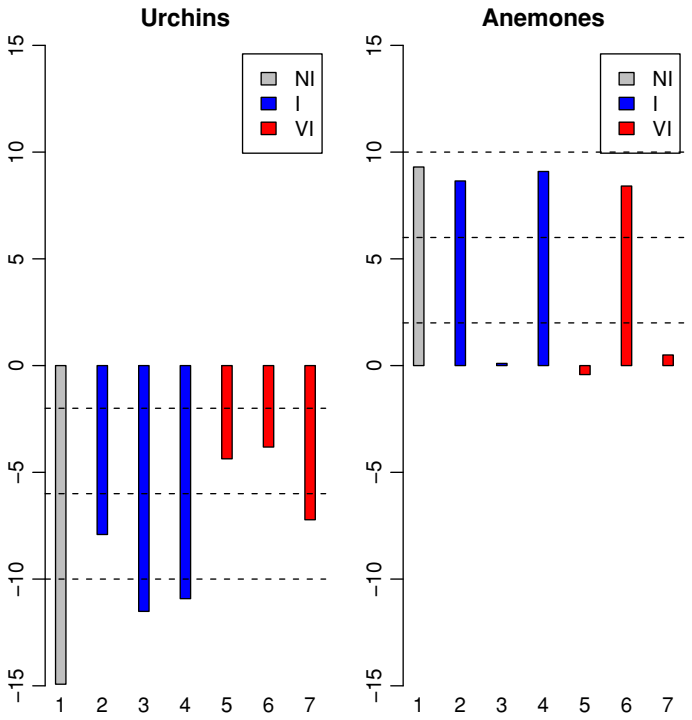


**Fig. 7** Evolution of the partial Bayes factors (at $2\times$ log) scale) comparing the $\mathrm{BYM}_{\mu}\mathrm{R}_{\rho}$-LOL model to the $\mathrm{BYM}_{\delta}\mathrm{R}_{\beta}$-$\Delta\Gamma$ model according different priors specifications and different partitions of the 3 years sample. "NI"= Non-informative prior (partition 1), "I"=Informative prior (partitions 2, 3, 4), "VI"=Very Informative prior (partitions 5, 6, 7)

validation sample is clearly observed in Fig. 7. This is true for all the models compared. Although the evidence in favor of the $\mathrm{BYM}_{\delta}\mathrm{R}_{\beta}$-$\Delta\Gamma$ model decreased when the level of prior information increased, the same conclusion remained in the case of the urchin abundance data: the $\mathrm{BYM}_{\delta}\mathrm{R}_{\beta}$-$\Delta\Gamma$ fitted the data better than the $\mathrm{BYM}_{\mu}\mathrm{R}_{\rho}$-LOL (i.e., Bayes factors were always strictly smaller than $-2$). Moreover, for a given level of prior information, this analysis showed the degree of evidence in favor of the $\mathrm{BYM}_{\mu}\mathrm{R}_{\rho}$-LOL model could vary enormously depending on the data used to compute Bayes factors. This was clearly the case for the anemone abundance data: a very strong evidence in favor of the LOL versions based on the data collected in 2000 but no evidence for the ones collected in 1999 and 2001. This sensitivity of Bayes factors to the validation sample may indicate interannual changes in the distribution of abundance data.

## 4.2 Comparing the predictive abilities

When the interannual changes in the abundance data are small enough to be ignored, such statistical models could be used as predictive tools, for instance, to help to plan future surveys in the sGSL. Therefore, comparing models can also be performed within

a predictive setting with a loss function to evaluate both the discrepancy between the test data and the predictive structure and the variance of predictions.

By dividing again the data y into a learning sample $y_{(-t)}$ and a test sample $y_{(t)}$ of quantities we want to re-evaluate, we computed the Posterior Predictive Loss Criterion (PPLC) proposed by Gelfand and Ghosh (1998) for each model $M$ of Table 1:

$$C^{\omega}_{(y_{(-t)},y_{(t)})}(M) = \sum_{l=t}^{n}(\hat{\sigma}_l^2) + \frac{\omega}{\omega+1}\sum_{l=t}^{n}\left(\hat{\mu}_l - y_l\right)^2 \tag{3}$$

where $\hat{\mu}_l = \mathbb{E}(\hat{Y}_l)$ and $\hat{\sigma}_l^2 = \mathbb{V}ar(\hat{Y}_l)$, i.e., the mean and variance of the predictive distribution of $\hat{Y}_l$ (the predictive variable for $y_l$) given the learning sample $y_{(-t)}$. For a given model, a smaller criterion indicates better predictive abilities.

Starting with a non-informative prior, we processed the abundance data from 1999 to 2001 and then tried to predict data collected in 2002. For each tow made in 2002, we generated a 2000-sample of abundance values according to the related predictive distribution. Summary statistics were calculated from this. For instance, Table 4 summarizes the PPLC computed when $\omega = 1$ (see Eq. 3) for the proportion of null catches (no organisms caught) and the average biomass expected to be collected for urchins and anemones, respectively.

For both species, the LOL provided better predictions of the average expected biomass and proportion of zero values collected compared to the analogous $\Delta\Gamma$ models (Table 4). The largest improvements were obtained for the predictions of the average biomass with PPLC clearly smaller for the LOL model compared to analogous versions of the $\Delta\Gamma$ model.

Looking at each component of the PPLC (Eq. 3), we noted that not only the precision of predictions was clearly improved from the LOL models but also that the discrepancy between the observed average abundances and the predicted values was

**Table 4** PPLC computed from the abundance data collected in 2002 for $\omega = 1$

| | Urchins | | Anemones | |
|---|---|---|---|---|
| | Proportion of zero values | Average biomass | Proportion of zero values | Average biomass |
| $(\text{LOL})^{\otimes 38}$ | 1.30 | 566.49 | 2.20 | 26.37 |
| $\text{R}_{\mu,\rho}$-LOL | 1.32 | 492.35 | 2.11 | 27.26 |
| $\text{R}_{\mu}$-LOL | 1.25 | 414.97 | 2.16 | 29.20 |
| $\text{BYM}_{\mu}$-LOL | 1.28 | 411.81 | 2.11 | 29.76 |
| $\text{BYM}_{\mu}\text{R}_{\rho}$-LOL | 1.34 | 473.67 | 2.03 | 26.89 |
| $(\Delta\Gamma)^{\otimes 38}$ | 1.45 | 758.89 | 2.17 | 34.53 |
| $\text{R}_{\delta,\beta}$-$\Delta\Gamma$ | 1.36 | 1235.47 | 2.14 | 28.18 |
| $\text{R}_{\delta}$-$\Delta\Gamma$ | 1.36 | 646.15 | 2.13 | 50.34 |
| $\text{BYM}_{\delta}$-$\Delta\Gamma$ | 1.32 | 644.17 | 2.05 | 50.46 |
| $\text{BYM}_{\delta}\text{R}_{\beta}$-$\Delta\Gamma$ | 1.33 | 810.32 | 2.05 | 31.97 |

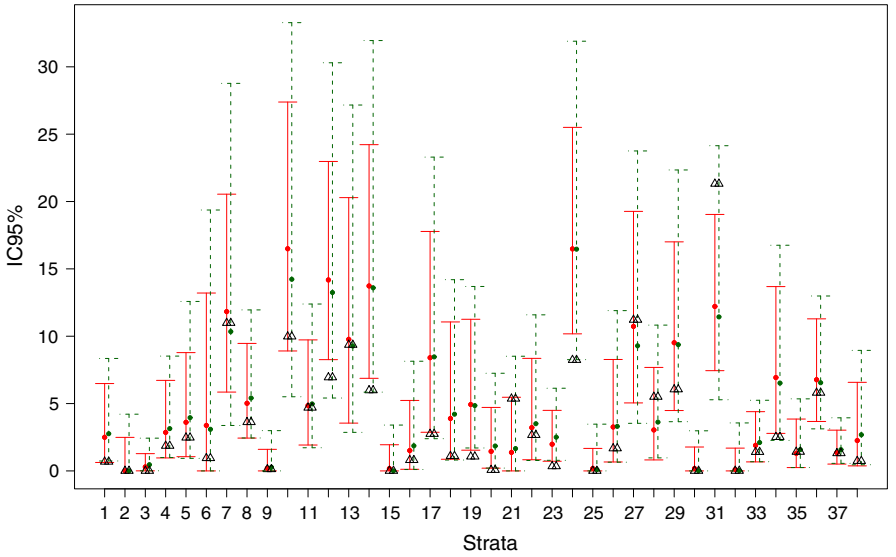


**Fig. 8** Urchins: 95% credible intervals computed from the predictive distributions for the average biomass expected to be caught from the $\mathrm{BYM}_{\mu}\mathrm{R}_{\rho}$-LOL model (solid lines) and $\mathrm{BYM}_{\delta}\mathrm{R}_{\beta}$-$\Delta\Gamma$ model (dashed lines). The triangles correspond to the average biomass observed in 2002. The points correspond to the medians of the predictive distributions.

globally smaller (results not shown). Figure 8 illustrates these results. We note that the $\mathrm{BYM}_{\mu}\mathrm{R}_{\rho}$-LOL model always provided smaller 95% predictive intervals while the medians of the predictive distributions were either similar or slightly closer to the empirical values compared to the $\mathrm{BYM}_{\delta}\mathrm{R}_{\beta}$-$\Delta\Gamma$ model.

Independent versions $(\mathrm{LOL})^{\otimes 38}$ and $(\Delta\Gamma)^{\otimes 38}$ are overfitted models which explains why they generally provided the worst predictions (Table 4).

It is worth mentioning that we also computed the PPLC when $\omega = 3$ and $\omega = +\infty$. The same conclusions remained concerning the predictive performances of competing models when the weight of the component measuring the discrepancy between observations and predictions, given by $\frac{\omega}{1+\omega}$, increased (results not shown).

## 4.3 The $\mathrm{BYM}_{\mu}\mathrm{R}_{\rho}$-LOL model

Among the competing hierarchical versions based on LOL (see Table 1), the $\mathrm{R}_{\mu}$-LOL and $\mathrm{BYM}_{\mu}$-LOL are the poorest members of the LOL family in terms of fitting ability and the independent version $(\mathrm{LOL})^{\otimes 38}$ is over-fitted. Globally speaking, the $\mathrm{R}_{\mu,\rho}$-LOL and $\mathrm{BYM}_{\mu}\mathrm{R}_{\rho}$-LOL models have similar fitting abilities for both species (partial Bayes factors computed but not shown). However, the spatial version $\mathrm{BYM}_{\mu}\mathrm{R}_{\rho}$-LOL is clearly superior to the regionalized version for predictive purposes (Table 4). That is why, we favor this model and we give details on the results obtained for this model.

**Table 5** Posterior means, coefficients of variation and 95% credible intervals for the $\mathrm{BYM}_{\mu}\mathrm{R}_{\rho}$-LOL model parameters (see Sect. 2.4.2).

| Parameters | Posterior mean | Posterior CV | $\mathrm{IC}_{95\%}$ |
|---|---|---|---|
| Urchins | $s^2_{\mathrm{IAR}}$ | 2.23 | [0.90, 4.30] |
| | $s^2_{\epsilon}$ | 0.12 | [0.002, 0.53] |
| | $m_0$ | −0.03 | [−0.22, 0.15] |
| | $\kappa$ | 0.93 | [0.68, 0.99] |
| Anemones | $s^2_{\mathrm{IAR}}$ | 1.15 | [0.45, 2.36] |
| | $s^2_{\epsilon}$ | 0.05 | [0.001, 0.29] |
| | $m_0$ | −1.015 | [−1.21, −0.83] |
| | $\kappa$ | 0.94 | [0.65, 0.99] |

$\kappa$ indicates the part of variability explained by the IAR structure

Table 5 contains the posterior means, coefficients of variation and 95% credible intervals for the parameters of the $\mathrm{BYM}_{\mu}\mathrm{R}_{\rho}$-LOL model from MCMC runs performed on abundance data of urchins and anemones collected in 1999, 2000 and 2001. The proportion of variability explained by the IAR component of the BYM structure was measured by the ratio $\kappa$ of the IAR model and the global variability.

In the absence of spatial effects, the expected number of patches of urchins is rather variable around 1 (0 is contained within the 95% posterior credible interval for $m_0$, the global mean for the $\log(\mu_i)$ $i = 1, 2, \ldots, 38$). For the anemones, this expected number is close to zero with a 95% posterior credible interval for $m_0$ strictly inferior to zero.

For both species, a strong spatial autocorrelation between neighboring strata for the average number of patches collected was estimated (Table 5). Indeed, the IAR structure accounted for about 93% of the global heterogeneity for each species. The variance of the residual noise was therefore smaller than that of the IAR structure. This seems to indicate that the urchins and anemones have a broad-scale patchy distribution in the sGSL. The local variance estimate of the IAR structure was smaller for the anemones (1.15) than for the urchins (2.23). Basically, this means that the spatial distribution of urchins displays more heterogeneity between neighboring strata than the anemones.

The maps displayed in Fig. 9 facilitate a comparison of the average biomass of urchins and anemones collected in 2002 to the values predicted with the $\mathrm{BYM}_{\mu}\mathrm{R}_{\rho}$-LOL model. Globally speaking, the maps of observed and predicted average biomass agree (empirical correlation: around 0.78 for both species). The main visual differences appear for the anemones. This may be explained by major interannual changes in the abundances of anemones as previously emphasized by the high sensitivity of Bayes factors to the year of data used to compute them (Fig. 7). As expected, the $\mathrm{BYM}_{\mu}\mathrm{R}_{\rho}$-LOL model induced a local smoothing of the average biomass between neighboring strata. Hence, the predicted maps emphasize less heterogeneity in the spatial variations of the average biomass compared to the observed maps.

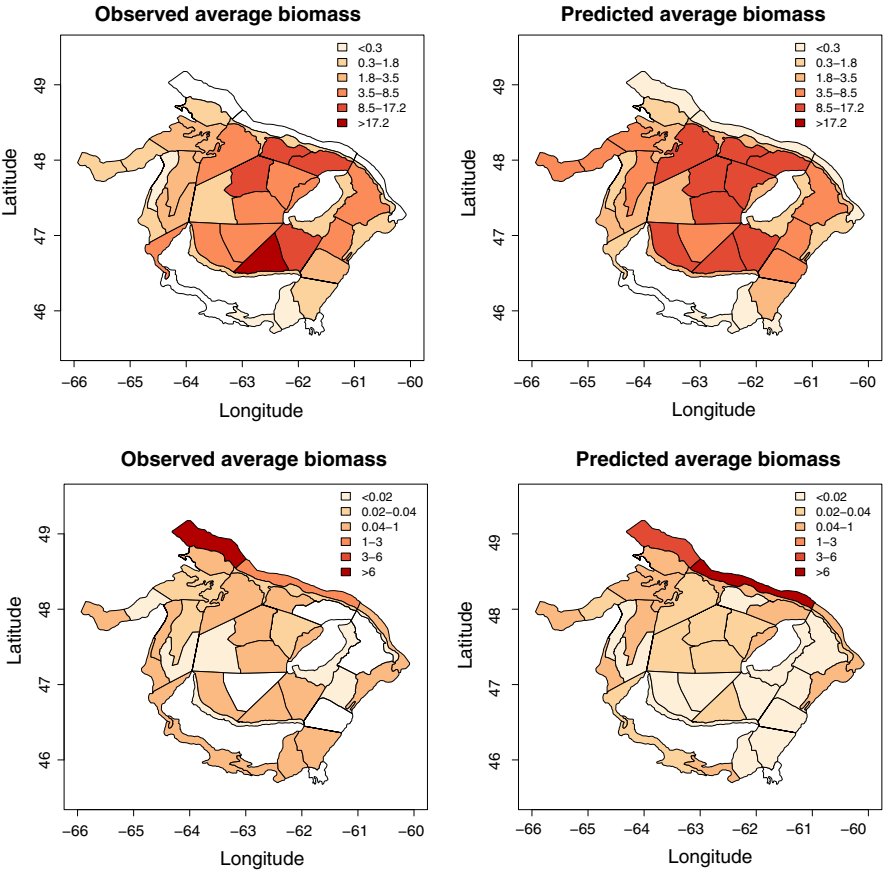


**Fig. 9** Average biomass in urchins (*top*) and anemones (*bottom*) observed in 2002 and predicted with the $\mathrm{BYM}_{\mu}\mathrm{R}_{\rho}$-LOL model, given the learning sample of data collected from 1999 to 2001

## 5 Discussion and perspectives

The LOL has considerable potential for the general analysis of zero-inflated abundance survey data. First, its latent variables $\mu$ and $\rho$ can easily incorporate several relevant properties of organisms distribution. For example, environmental covariates can be used in determining the prior distributions for $\mu$ and $\rho$ via generalized linear models. Another approach could be to explain the average biomass $\frac{\mu}{\rho}$ by environmental covariates instead of $\mu$ and $\rho$. Secondly, many hierarchical constructions can be imagined from this conceptual data submodel. In this paper, we proposed five possible hierarchical variants. We excluded multivariate $\mathrm{BYM}_{\mu,\rho}$-LOL to favor parsimony and to avoid mathematical intricacy. We did not consider spatial effects on $\rho$ ($\mathrm{BYM}_{\rho}$-LOL, $\mathrm{R}_{\rho}$-LOL or $\mathrm{BYM}_{\rho}\mathrm{R}_{\mu}$-LOL) on the basis of a preliminary descriptive analysis of the sGSL dataset. We computed Moran's statistics for $\rho = (\rho_1, \rho_2, \ldots, \rho_{38})$ and $\mu = (\mu_1, \mu_2, \ldots, \mu_{38})$, respectively. In our case, as it must refer to latent variables $z$,

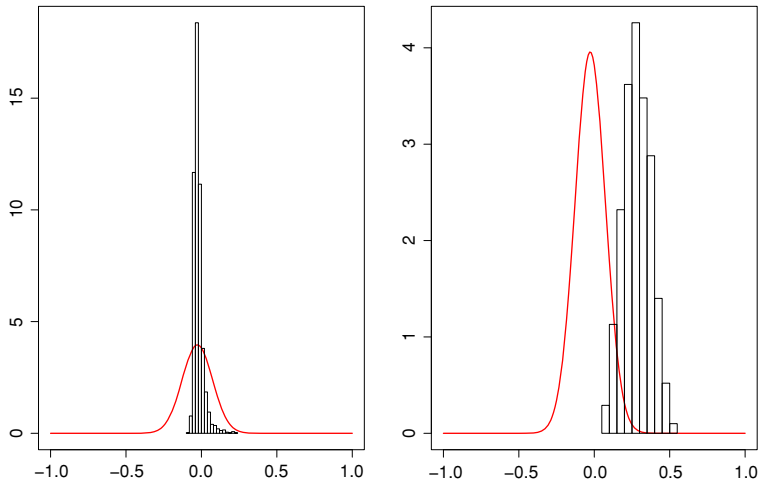


**Fig. 10** Posterior distributions of Moran's index for the latent variables $\mu_i$ (*right*) and $\rho_i$ (*left*) when the $(\text{LOL})^{\otimes 38}$ model is applied to urchin abundance data. The solid line represents the distribution of Moran's index under the independence hypothesis

we used the MCMC samples obtained from the independent case $(\text{LOL})^{\otimes 38}$ to compute a posterior sample of Moran's statistics for $\rho$ and $\mu$, respectively. These are plotted in Fig. 10. Moran (Banerjee et al. 2004) showed that if the $\mu_i$ (respectively $\rho_i$) are i.i.d., Moran's index is asymptotically distributed with mean $-\frac{1}{\text{Number of strata}-1} = -\frac{1}{37}$ and an explicit variance. Figure 10 shows a larger departure from the independence hypothesis for the average number of patches collected during a standard survey $\mu_i$ ($p_{\text{value}} = 0.0005$) than for the inverse of the average quantities of biomass in one patch $\rho_i$ ($p_{\text{value}} = 0.0049$). Consequently, we put a BYM structure on $\log(\mu_i)$ and assumed no spatial effect on $\rho_i$.

The analytical p.d.f for the LOL is:

$$[Y = y|\mu, \rho] = 1_{y=0}e^{-S\mu} + 1_{y>0}\frac{\mu S\rho}{\sqrt{\mu S\rho y}}\exp(-S\mu - \rho y)I_1(2\sqrt{\mu S\rho y}) \quad (4)$$

where $S$ denotes the catch effort, and $I_1$ the modified Bessel function of the first kind linked with the occurrence of a strictly positive event. We did not present Eq. 4 earlier as we preferred to highlight its conceptual properties of compound Poisson process. The hierarchical Bayesian approach makes clear the role of different processes by specifying them at different stages of the model's hierarchy and confers a conceptual sense on parameters: $\mu$, the expected number of patches to be collected during a standard sampling event and $1/\rho$ the average biomass contained in any patch in a given habitat.

Although the LOL is easily interpretable in its hierarchical form (see Eq. 1), great care must be given not to over-interpret the significance of the LOL parameters $\mu$ and $\rho$. These unknowns are not directly observed when sweeping a net through water to capture organisms. Furthermore, the definition of patchiness depends on the scale at

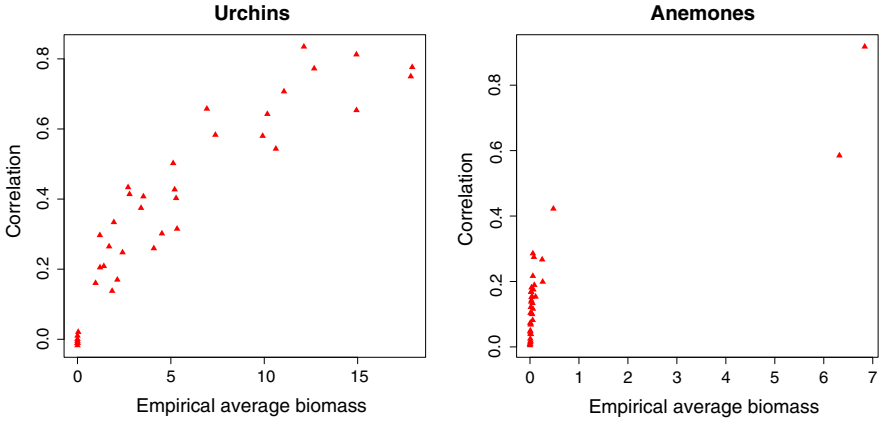


**Fig. 11** Posterior correlations between $\rho_i$ and $\mu_i$ ($i = 1, 2, \ldots, 38$) according to the empirical average biomass collected in September from 1999 to 2001. They were computed from MCMC samples obtained by fitting the $(\text{LOL})^{\otimes 38}$ model. Each red point corresponds to one stratum

which it is measured. While it might be tempting to use the two variables to make statements about the patchiness of organism distribution, the possible high degree of correlation between the two confounds their interpretation (Fig. 11). Basically, this linear correlation increases with the average biomass collected in a given habitat. In our case study, this degree of correlation is clearly higher for urchin compared to anemone abundance data (Fig. 11). More large and fewer null catches for urchins compared to anemones explain this difference. In other words, when a lot of biomass is caught, the LOL model does not properly resolve the differences between sampling a single large patch or numerous small single-individual patches. One drawback of a high degree of correlation is that successive values of MCMC chains are highly autocorrelated. Consequently, the convergence of MCMC algorithms is reached more slowly. This correlation should not however be viewed necessarily as a weakness of the model. Indeed, when the LOL is used to estimate or predict the biomass in a given habitat, the correlation between $\mu$ and $\rho$ does not matter since the only quantity of interest $\frac{\mu S}{\rho}$ combines the two unknowns of Eq. 1. In particular, our results show that the predictive abilities of the LOL models are globally better than $\Delta\Gamma$ models whatever degree of correlations between $\mu$ and $\rho$. It is only jointly however that these two parameters can be used to describe the distribution of observed biomass.

Our comparative study emphasizes that the fitting abilities of the LOL models are slightly superior to analogous versions of the $\Delta\Gamma$ models when the considered zero-inflated data have an empirical distribution with a strong peak at zero and small quantities of strictly positive abundances (e.g., anemones, Fig. 2). However, as soon as departures from these characteristics are observed (e.g., urchins), the fitting abilities of the LOL models tend to worsen relative to $\Delta\Gamma$ models for the small proportions of zero values. The expected number of patches is often under-estimated, possibly due to a high level of correlations between $\mu$ and $\rho$.

In this paper, a BYM model has been proposed to create local spatial dependencies between neighboring areal units through a IAR component. It is worth mentioning

that this choice is only relevant when we assume that the survey area is split into permanent areal units. It allows to coherently match a change of sampling effort through the additivity property provided by the LOL model for abundance data collected in homogeneous sites. However, as soon as embedded partitions of this survey area are considered, the conditional dependency relationships implied by the IAR structure are modified. Consequently, the additivity property provided by the LOL model is not verified any more between two (or more) observations located in areal units defined as sub-divisions of a same homogeneous original unit.

There are numerous possibilities for extending the use of the LOL in future studies. For example, a multivariate LOL structure could be used to describe interspecies relationships in an analysis of ecological communities. Also, a discrete version of the LOL, obtained by changing the exponential into a geometric p.d.f, may provide an alternate model to the traditional ZIP and ZINB models used for counts. In this paper, we have focused on underlying areal-based structures to model spatial correlations. However, other models exist to represent spatially structured processes (Cressie 1993). Hence, it would also be worth exploring structures that directly model spatial correlations between abundance data depending on the geographical distances between sampling sites. A geostatistical version of the LOL would likely be well suited for analyzing zero-inflated continuous data at a finer spatial scale.

## Appendix 1: MCMC Computation of Bayes factor

The Bayes factor $BF_{i,j}$ is a measure of the relative credibility of the model $M_i$ compared to the model $M_j$ given the data $y$ and appears as the ratio of the posterior probability of the competing models when the prior probabilities are equal. Consequently, it is simply the ratio of the marginal likelihood of the data under the model $M_i$ to the marginal likelihood under the model $M_j$:

$$BF_{i,j} = \frac{\frac{[M_i|y]}{[M_j|y]}}{\frac{[M_i]}{[M_j]}} = \frac{[y|M_i]}{[y|M_j]} \tag{5}$$

Computing the marginal likelihood can be done by the following standard decomposition for a given model $M$:

$$[y|M] = \int [y|M, \phi][\phi|M]d\phi \tag{6}$$

In this section, $\phi$ is a generic notation employed for *any* subset of the unknowns $(z, \theta)$. To emphasize that prior, likelihood and predictive pdfs are model specific, we write in this section $[\phi|M]$, $[y|M, \phi]$, $[y|M]$ instead of $[\phi]$, $[y|\phi]$, $[y]$.

A traditional stochastic approximation by drawing a sample $(\phi^{(1)}, \phi^{(2)}, \ldots, \phi^{(k-1)}, \phi^{(k)})$ from the prior distribution $[\phi|M]$ and performing the prior arithmetic mean of the $[y|M, \phi^{(k)}]'s$ is often not very efficient since probable values from

the sampling prior distribution and from the likelihood function may lie far apart. We used the method proposed by Kass and Raftery (1994):

$$[y\,|M\,]^{-1} = \int [y\,|M, \phi\,]^{-1}[\phi\,|M, y\,]d\phi$$

A more robust approximation of the marginal likelihood of model $M$ can be computed by the posterior harmonic mean:

$$[y\,|M\,] \approx \left(\frac{1}{H}\sum_{h=1}^{H}[y\,\big|\phi^{(h)}, M\,]^{-1}\right)^{-1} \tag{7}$$

where $\phi^{(h)}$ is the $g$th set of unknowns of the posterior sample (from $[\phi\,|y, M\,]$), $[y\,|\phi^{(h)}, M\,]$ is the value of the sampling density of the observed data in $M$ calculated at $\phi^{(h)}$, and $H$ is the sample size. This estimate of the marginal likelihood is known to be unstable, because the harmonic mean of the likelihood is highly sensitive to the very small values that may appear during the sampling. This may hinder the comparison of two models when their difference in credibility is low. However, in practice, if the contrast between the credibility of the different models is high, the approximation from (7) gives results which are accurate enough for identifying the most credible model(s). An additional trick to increase numerical stability of (6) or (7) is to ensure $\phi$ as the smallest subset of $(z, \theta)$ such that $[y\,|M, \phi\,]$ is explicitly known. As hierarchical models often exhibit partial conjugacy between their successive layers, dimension reduction can be obtained by performing integration whenever possible, to shrink $[y\,|M, z, \theta\,]$ into $[y\,|M, \phi\,]$.

In our case study, a 500, 000 MCMC replicates were used to compute stable approximations of Bayes factors.

## Appendix 2: Testing Bayes factor sensitivity to priors

Equation 6 shows that the marginal likelihood is not defined when an improper distribution $[\theta\,|M\,]$ is chosen as a prior. More generally, Bayes factors are known to be sensitive to prior choice (Kass and Raftery 1994; Sinharay and Stern 2002). A traditional approach to testing Bayes factor sensitivity to priors consists in computing Bayes factors from different prior distributions. Its major drawback is computational intensity, as one MCMC algorithm has to be run for each tested prior.

Informative priors are recommended to get a meaningful comparison between models relying on Bayes factors (O'Hagan 1995). However, in many hierarchical structures, parameters have purely conceptual interpretations (e.g., $\mu$ and $\rho$ for the LOL model). Defining a prior knowledge on these unknown quantities through informative probability distributions is tricky. That is why, non-informative priors are mainly chosen. This is the case in this paper.

Now, consider a split sample test analysis. Suppose that the data $y$ can be divided into two independent samples blocks $y = (y_{(-t)}, y_{(t)})$. $t$ is a partition index such that:

$$[y\,|M,\theta\,] = [y_{(-t)}\,|M,\theta\,] \times [y_{(t)}\,|M,\theta\,] \tag{8}$$

The subsample $y_{(-t)}$ is a *learning* sample: starting with the non-informative prior $[\theta\,|M\,]$, even improper, information conveyed in $y_{(-t)}$ is processed into an informative pdf $[\theta\,\big|M,\, y_{(-t)}\,]$. This latter posterior pdf is used as a proper prior for the study with the *test* sample $y_{(t)}$. Using definition (5), the following partial Bayes factor can be straightforwardly derived as a measure of the relative credibility of the model $M_i$ compared to the model $M_j$ for the test data $y_{(t)}$ after *assimilating* $y_{(-t)}$ during a learning stage:

$$BF_{ij,(t)} = \frac{[y_{(t)}|M_i,\, y_{(-t)}]}{[y_{(t)}|M_j,\, y_{(-t)}]} \tag{9}$$

The learning sample $y_{(-t)}$ must be made large enough such that $[\theta\,\big|M,\, y_{(-t)}\,]$ is proper and sufficiently informative, and small enough so that $y_{(t)}$ remains a representative sample of the whole dataset.

When using an informative pdf $[\theta\,\big|M,\, y_{(-t)}\,]$ as a prior for the study with the validation sample $y_{(t)}$, provided that the likelihood can be decomposed using Eqs. 8, 7 gives the marginal likelihood:

$$\begin{aligned}[y_{(t)}\,\big|M,\, y_{(-t)}\,]^{-1} &= \int [y_{(t)}\,|M,\phi\,]^{-1}[\phi\,\big|M,\, y_{(-t)},\, y_{(t)}\,]d\phi \\ &= \int [y_{(t)}\,|M,\phi\,]^{-1}[\phi\,|M,\, y\,]d\phi \end{aligned} \tag{10}$$

Varying the partition index $t$ in the computation of the partial Bayes factor (see definition (9)) does not induce any additional numerical cost: indeed Eq. 10 can be approximated for various $t$, by the following expression analogous to 7, but with the same replicates $\phi^{(h)}$ drawn once from the complete posterior $[\phi\,|M,\, y\,]$:

$$[y_{(t)}\,|M\,] \approx \left( \frac{1}{H} \sum_{h=1}^{H} [y_{(t)}\,\Big|\phi^{(h)},\, M\,]^{-1} \right)^{-1} \tag{11}$$

This split test sample method, based on an original idea proposed by Lempers (1971), allows to easily conduct a robustness analysis to prior specification by varying the partition index $t$, possibly after rearrangering the ordering of the data.

**Acknowledgements** We are grateful to Etienne Rivot and Liliane Bel for their constructive suggestions and to Jacques Bernier for pointing out the possible ecological interpretation of the LOL. Financial support was provided by AgroParisTech for Sophie Ancelet to work overseas with ecologists from Fisheries and Oceans Canada in Moncton, N.B.

## References


Aitchison J, Brown JAC (1957) The lognormal distribution with special reference to its uses in econometrics. Cambridge University Press, Cambridge

Ancelet S (2008) Exploiter l'approche hiérarchique bayésienne pour la modélisation statistique de structures spatiales. Application en écologie des populations. Dissertation, Agro Paris Tech, Paris Institute of Life and Environmental Science and Technology

Banerjee S, Carlin B, Gelfand A (2004) Hierarchical modeling and analysis for spatial data. Chapman and Hall/CRC, Boca Raton

Bernier J, Fandeux D (1970) Théorie du renouvellement—application à l'étude statistique des précipitations mensuelles. Rev Stat Appl 18(2):75–87

Besag J, York J, Mollié A (1991) Bayesian image restoration with two applications in spatial statistics. Ann I Stat Math 43(1):1–51

Brooks SP, Gelman A (1998) Alternative methods for monitoring convergence of iterative simulations. J Comput Graph Stat 7(4):434–455

Chyan-Huei Lo N, Jacobson LD, Squire JL (1992) Indices of relative abundance from fish spotter data based on delta-lognormal models. Can J Fish Aquat Sci 49(12):2515–2526

Congdon P (2001) Bayesian statistical modelling. John Wiley & Sons Inc., New York

Cressie NAC (1993) Statistics for spatial data. John Wiley & Sons Inc., New York

Feller W (1971) An introduction to probability theory and its applications. John Wiley & Sons Inc., New York

Fletcher D, MacKenzie D, Villouta E (2005) Modelling skewed data with many zeros: a simple approach combining ordinary and logistic regression. Environ Ecol Stat 12(1):45–54

Gelfand AE, Ghosh S (1998) Model choice: a minimum posterior predictive loss approach. Biometrika 85(1):1–11

Gelfand AE, Smith AFM (1990) Sampling based approach to calculating marginal densities. J Am Stat Assoc 85(410):398–409

Hall DB (2000) Zero-inflated Poisson and binomial regression with random dffects: a case study. Biometrics 56(4):1030–1039

Heilbron DC (1994) Zero-altered and other regression models for count data with added zeros. Biometrical J 36(5):531–547

Hurlbut T, Clay D (1990) Protocols for research vessel cruises within the Gulf Region (demersal fish)(1970–1987). Can Manus Rep Fish Aquat Sci 2082:2143

Kass RE, Raftery AE (1994) Bayes factors. J Am Stat Assoc 90(430):773–795

Lambert D (1992) Zero-inflated Poisson regression, with an application to defects in manufacturing. Technometrics 34(1):1–4

Legendre P, Legendre L (1998) Numerical ecology. Elsevier Science, Amsterdam

Lempers FB (1971) Posterior probabilities of alternative linear models. Rotterdam University Press, Rotterdam

Loring DH, Nota DJG (1973) Morphology and sediments of the Gulf of St. Lawrence. B Fish Res Board Can 182:147

Martin TG et al (2005) Zero tolerance ecology: improving ecological inference by modelling the source of zero observations. Ecol Lett 8:1235–1246

Myers RA, Pepin P (1990) The robustness of lognormal based estimators of abundance. Biometrics 46(4):1185–1192

O'Hagan A (1995) Fractional bayes factors for model comparison. J Roy Stat Soc B 57(1):99–138

Pennington M (1983) Efficient estimators of abundance for fish and planktons survey. Biometrics 39(1):281–286

Pennington M (1996) Estimating the mean and variance from highly skewed marine data. Fish B 94:498–505

Ridout M, Demétrio CGB, Hinde J (1998) Models for count data with many zeros, International biometric conference, Cape Town

Robert CP, Casella G (2004) Monte Carlo statistical methods. Springer, New York

Sinharay S, Stern HS (2002) On the sensitivity of Bayes factors to the prior distributions. Am Stat 56(6):196–201

Spiegelhalter D, Best N, Carlin BP, VanDer Linde A (2002) Bayesian measures of model complexity and fit. J Roy Stat Soc B 64(4):583–639

Spiegelhalter D, Thomas A, Best N, Lunn D (2007) OpenBUGS user manual, Version 3.0.2, MRC Biostatistics Unit, Cambridge, Available via http://www.mathstat.helsinki.fi/openbugs/

Stefansson G (1996) Analysis of groundfish survey abundance data: combining the GLM and delta approaches. ICES J Mar Sci 53(3):577–588

Thomas A, Best N, Lunn D, Arnold R, Spiegelhalter D (2007) GeoBUGS user manual, Version 1.3, Rolf Nevanlinna Institute, Available via http://www.mathstat.helsinki.fi/openbugs/

Welsh AH, Cunningham RB, Donnelly CF, Lindenmayer DB (1996) Modelling the abundance of rare species: statistical counts with extra zeros. Ecol Model 88(1):297–308

## Author Biographies

**Sophie Ancelet** is a post-doctoral researcher in Biostatistics at the French National Institute of Health and Medical Research (INSERM). Her main research interests are Bayesian hierarchical modelling, with a particular interest for latent spatial structures and multivariate processes commonly explained by latent random effects, and Bayesian statistics. Her methodological developments are applied to environmental management and population biology and, more recently, disease mapping and reproductive epidemiology.

**Marie-Pierre Etienne** is a Senior Lecturer in Applied Statistics for Ecology, Biodiversity and Environment at AgroParisTech. She belongs to the research team MORSE. Her main research subjects are environmental spatial statistics and genomics with a particular interest for hierarchical modelling and Bayesian inference.

**Hugues Benoît** is a fisheries ecologist working for the Federal Department of Fisheries and Oceans Canada at the Gulf Fisheries Centre in Moncton, NB. His research interests include trying to understand the direct and indirect impacts of harvesting and environmental change on marine communities.

**Eric Parent** is a Professor in Applied Statistics and Probabilistic Modelling for Environmental Engineering at AgroParisTech. He is the director of the research group MORSE. His broader interests include "Bayesian Statistics at work" and statistical modelling for environmental sciences with a particular interest for spatial statistics, extremes and their decisional aspects. He has co-authored two books (in French), one on Bayesian statistics for environmental engineering and the other on theoretical and algorithmic aspects of Bayesian theory.