Ecological Modelling xxx (2009) xxx–xxx

Contents lists available at ScienceDirect

Ecological Modelling

journal homepage: www.elsevier.com/locate/ecolmodel

Autocorrelation offsets zero-inflation in models of tropical saplings density

O. Flores [a,*], V. Rossi [c], F. Mortier [b]

[a] Centre d'Écologie Fonctionnelle et Évolutive, CNRS – UMR 5175, 1919, route de Mende, 34293 Montpellier Cedex 5, France
[b] CIRAD - UPR Génétique forestière, TA 10/C, Campus international de Baillarguet, 34398 Montpellier Cedex 5, France
[c] CIRAD - UPR Dynamique des forêts naturelles, TA 10/D, Campus international de Baillarguet, 34398 Montpellier Cedex 5, France

ARTICLE INFO

Article history:
Received 10 March 2008
Received in revised form 13 January 2009
Accepted 14 January 2009
Available online xxx

Keywords:
Hierarchical Bayesian Modelling
Conditional Auto-Regressive model
Variable selection
Zero-Inflated Poisson
Posterior predictive
Paracou
French Guiana

ABSTRACT

Modelling the local density of tropical saplings can provide insights into the ecological processes that drive species regeneration and thereby help predict population recovery after disturbance. Yet, few studies have addressed the challenging issues in autocorrelation and zero-inflation of local density. This paper presents Hierarchical Bayesian Modelling (HBM) of sapling density that includes these two features. Special attention is devoted to variable selection, model estimation and comparison.

We developed a Zero-Inflated Poisson (ZIP) model with a latent correlated spatial structure and compared it with non-spatial ZIP and Poisson models that were either autocorrelated (Spatial Generalized Linear Mixed, SGLM) or not (generalized linear models, GLM). In our spatial models, local density autocorrelation was modeled by a Conditional Auto-Regressive (CAR) process. 13 explicative variables described ecological conditions with respect to topography, disturbance, stand structure and intraspecific processes. Models were applied to six tropical tree species with differing biological attributes: *Oxandra asbeckii*, *Eperua falcata*, *Eperua grandiflora*, *Dicorynia guianensis*, *Qualea rosea*, and *Tachigali melinonii*. We built species-specific models using a simple method of variable selection based on a latent binary indicator.

Our spatial models showed a close correlation between observed and estimated densities with site spatial structure being correctly reproduced. By contrast, the non-spatial models showed poor fits. Variable selection highlighted species-specific requirements and susceptibility to local conditions. Model comparison overall showed that the SGLM was the most accurate explanatory and predictive model. Surprisingly, zero-inflated models performed less well.

Although the SZIP model was relevant with respect to data distribution, and more flexible with respect to response curves, its model complexity caused marked variability in parameter estimates. In the SGLM, the spatial process alone accounted for zero-inflation in the data. A refinement of the hypotheses employed at the process level could compensate for distribution flaws at the data level. This study emphasized the importance of the HBM framework in improving the modelling of density–environment relationships.

1. Introduction

The population dynamics of tropical tree species involves multiple and heterogeneous processes. These biotic and abiotic processes, such as competition and disturbance, are of a particular impact on the spatial patterns of early life-stages. These patterns integrate not only species preferences, but also some dispersal signal which blurs as mortality filters come into operation (Wang and Smith, 2002). Because of this complexity, and particularly in early life-stages, spatial patterns constitute the subject of studies used to draw ecological inference (Austin, 2002). The analysis of these spatial patterns can be valuably conducted in a modelling approach (Guisan and Zimmermann, 2000).

The modelling approach developed here follows the general framework proposed by Austin (2002) which integrates three interacting conceptual components. First, the *ecological* model addresses ecological theory in a given system. When species distributions are concerned, the individualistic community scheme sets a relevant model in which each species interacts with its environment through intrinsic rules (Guisan and Zimmermann, 2000). Second, the *data* model describes the studied system through designed response and explicative variables. In most studies of tree species distribution, the response variable is presence/absence. Fewer studies tackle the local density of conspecifics, especially in tropical rainforests (but see Svenning et al., 2006). This kind of response variable induces zero-inflation which occurs when the frequency

* Corresponding author.
E-mail address: olivierflores@free.fr (O. Flores).

0304-3800/$ – see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.ecolmodel.2009.01.030

of zero observations exceeds that expected in a classical distribution. Also, in tropical forests, marked heterogeneity in space and time makes it difficult to define and measure relevant explicative variables. Indirect explicative variables often serve as proxies that quantify ecological processes and direct (physiological) or resource gradients (Guisan and Zimmermann, 2000). The third and final component, the *statistical* model, defines the relationships between data model variables and the methods used for their analysis.

In this contribution we focused on the *statistical* model of Austin's framework in order to develop models of sapling density that include the issues raised by the *data model*: zero-inflated count data, numerous explicative variables, and spatial autocorrelation. Zero-inflation is a common feature of data in many domains and has recently received particular attention (Martin et al., 2005) in ecology. Null observations have different causes: (i) "structural" zeros relate to the absence of a species in unsuitable habitats or because it is scarce (Welsh et al., 1996), whereas (ii) "random" zeros arise by chance from ecological processes (e.g. dispersal limitation), or sampling or observer error (Martin et al., 2005). True zeros (structural or random) arise from ecological processes, whereas false zeros stem from sampling. True zeros are particularly likely to arise in tropical forests, due to vegetation features: extreme species richness implies low specific densities, even in abundant species, and a high frequency of rare species. Focusing on a particular life-stage may also induce zero-inflation because of low abundance.

Zero-inflated (ZI) models are a special case of finite-mixture models that mix two distributions to account for dispersion in data. ZI models offer statistical robustness and flexibility in the shape of response curves (Flores et al., 2006), a central issue in modelling studies (Guisan and Zimmermann, 2000; Oksanen and Minchin, 2002; Austin, 2007). However, they come at the cost of additional complexity over Poisson models. In the conditional ZI (Hurdle) model, structural and random zeros are modeled together as derived from a binomial process (Ridout et al., 1998). Non-zero data are modeled separately through a truncated Poisson (or negative binomial) distribution (Welsh et al., 1996; Barry and Welsh, 2002; Kuhnert et al., 2005). In the mixture ZI model, structural and random zeros are considered separately (Martin et al., 2005; Flores et al., 2006) in a two-stage process. A binary (Bernoulli) process first determines whether, in a second stage, an observation proceeds from a degenerated null process (leading to structural zeros) or from a Poisson process (possibly leading to random zeros). Finite-mixture models generalize parametric methods in allowing specification of non-classical data distributions (Richardson and Green, 1997). However, various alternative parametric and non-parametric methods are also available for empirical modellers to tackle these statistical issues. Recently applied methods include generalized linear models (GLM, Guisan et al., 2002; Miller and Franklin, 2002; Stephenson et al., 2006), generalized additive models (GAM, Barry and Welsh, 2002; Guisan et al., 2002; Moisen and Frescino, 2002), classification and regression trees (CART, Moisen and Frescino, 2002; Miller and Franklin, 2002), multivariate adaptive regression splines (MARS, Moisen and Frescino, 2002) and artificial neural networks (ANN, Moisen and Frescino, 2002).

Spatial autocorrelation has for many years been recognized as ubiquitous in ecological field data (Legendre, 1993). It challenges the classical statistical hypothesis of observations being independent. At the same time, explicit modelling of autocorrelation may provide insight into unobserved processes at various scales (Svenning et al., 2006; Miller et al., 2007). It is in tropical forests that local density is most likely to be autocorrelated. Tree species often display clumped spatial patterns at a local scale (Condit et al., 2000), because of limited dispersal, facilitation by conspecifics or patchy habitat requirement. It is noteworthy that such clumping may also induce zero-inflation, for instance in a regular sampling design. Autocorrelation can also be handled in various ways. At a local scale, a random variable often accounts for some dependence between neighboring observations (Lichstein et al., 2002; Miller et al., 2007; Svenning et al., 2006). Alternatives, for instance auto-regressive (AR) models, and particularly the Conditional Auto-Regressive (CAR) model, can account for spatial dependence arising from ecological processes (Lichstein et al., 2002).

Spatial statistical models can become complex when a mixed data distribution and/or mixed effects are addressed. The Hierarchical Bayesian Modelling (HBM) approach is particularly suited to such cases (Clark, 2005). The main advantage of HBM over other approaches is that it accommodates biological complexity into a series of simple conditional models (Wikle, 2003; Clark, 2005) and provides robust parameter estimates (Angers and Biswas, 2003). The classical hypothesis of independence between observations is replaced by conditional independence, given hypotheses on the structure of data covariance. At the same time, the Bayesian paradigm offers attractive advantages by its ability to integrate prior knowledge into a model, through prior distributions (Banerjee et al., 2003), and to provide a posterior parameters distribution instead of estimated values (Clark, 2005).

When multiple processes are likely to influence the response, the selection of variables becomes paramount. Explicative variables can be selected on a subjective basis or with respect to statistical criteria. Most studies dealing with variable selection use stepwise procedures based on the Akaike Information Criterion (AIC). Parameter estimation then lead to difficulties when variables are numerous, collinear, and when effects are low. Again, an HBM approach may be an effective method for dealing with the selection of variables (Clark, 2005). Several methods have been proposed and may be implemented with varying degrees of difficulty (Dellaportas et al., 2002). Here, we adopt a simple method based on a binary latent indicator. Its estimation provides the posterior probability that an explicative variable improves fitting when included in a model (Dellaportas et al., 2000; Ntzoufras et al., 2000).

This paper describes the building of a density model based on a latent CAR layer that drives a spatially structured behavior to a ZI Poisson data layer. It also compares simple and autocorrelated versions of Poisson and Zero-Inflated Poisson models based on selected explicative variables, and addresses model performance and complexity. The issues of zero-inflation, autocorrelation and variable selection are considered within the HBM framework. The models used are applied to six tropical tree species differing in shade-tolerance and dispersal modes in permanent sample plots (PSP) located in French Guiana. Specific emphasis is placed on investigating the effects of the local environment and intraspecific processes on sapling density.

2. Materials and methods

2.1. Study site and focal species

The study was conducted at the Paracou experimental site (5°18′N, 52°23′W) in a *terra firme* rain forest. The site lies in the coastal part of French Guiana and is subject to an under equatorial climate with a wet season and a dry season. A short drier period interrupts the rainy season from March to April.

The site consists of 300 m × 300 m PSP with a 25 m inner buffer zone. In each central 250 m × 250 m square, all trees ≥ 10 cm diameter at breast height (DBH) were identified and georeferenced. Girth at breast height, tree mortality (standing deaths and treefalls) and recruitment over 10 cm DBH have been monitored annually since 1984. Three treatments were applied over the 1986–1988 period combining selective logging of increasing intensity and additional

Please cite this article in press as: Flores, O., et al., Autocorrelation offsets zero-inflation in models of tropical saplings density. Ecol. Model. (2009), doi:10.1016/j.ecolmodel.2009.01.030

poison-girdling. The study described here focused on four adjacent PSP (an undisturbed control plot and one treated plot in each treatment) and on the period 1986–2003.

Six focal species were studied: one shade-loving species *Oxandra asbeckii* Pulle, R.E.Fr. (Annonaceae), three shade tolerant to mid-tolerant species (*Eperua falcata* Aublet, Caesalpiniaceae, *Eperua grandiflora* Aublet, Benth., Caesalpiniaceae, *Dicorynia guianensis* Amshoff, Caesalpiniaceae), and two light-demanding species (*Qualea rosea* Aublet, Vochysiaceae, *Tachigali melinonii*, Harms, Caesalpiniaceae). *O. asbeckii* is a bird-dispersed species of the understorey, with maximal height of 15 m. *E. falcata* is self-dispersed and *E. grandiflora* is gravity-dispersed; both species occur in the top canopy at a maximal height of 30–35 m (Sabatier, 1983). *D. guianensis*, *Q. rosea* and *T. melinonii* are wind-dispersed species of the top canopy with emergent trees reaching 40 m. *T. melinonii* is the fastest-growing and most light-demanding of the six species.

2.2. Data model: ecological descriptors

In 2002–2003, all plants in the four plots with 1 cm ≤ DBH ≤ 10 cm were sampled and georeferenced. DBH were recorded in 1-cm classes. Because of large differences in growth potential, tropical trees spend varying periods of time in early life-stages. Here, we allowed the sapling stage to be specifically defined in the data model. The sapling stage was limited by a species-specific upper DBH limit accounting for average growth during the post-logging period. Sapling DBH classes corresponded to 1–2 cm for *O. asbeckii*, 1–3 cm for *E. grandiflora*, 1–4 cm for *E. falcata*, 1–5 cm for *D. guianensis*, 1–6 cm for *Q. rosea* and 1–9 cm for *T. melinonii*. Saplings were counted on an exhaustive and regular basis in 10 m × 10 m cells (625 cells per PSP). The observed sapling density in the cells constituted the studied response variable ($n = 2500$).

Explicative variables constituted of 13 descriptors of ecological conditions that control tropical tree species density (Table 1). These variables were derived either from a Digital Elevation Model (DEM) of the site (elevation and slope), or from census data for trees (≥ 10 cm DBH, stand variables), calculated on 20-m radius plots centered on sampling cells. Two static variables described local forest structure in 2002: total basal area and basal area of pioneer taxa. Five variables characterized stand dynamics during both logging (1986–1988) and the following recovery period (1988–2003): four disturbance variables (Table 1) and a variable quantifying gross change in total basal area over the recovery period. The local disturbance regime was characterized by the mean and standard deviation of treefall age during the recovery period (Table 1).

Finally, two population variables estimated interactions with surrounding conspecific trees (Table 1) to account for intra-population and inter-life-stage autocorrelation. First, the distance from cell center to the nearest adult estimated saplings potential dispersal distance. Second, the basal area of living conspecific trees (≥ 10 cm DBH) on the 20-m radius plots accounted for intraspecific competition. Adults included mature trees, i.e. trees with a DBH greater than a threshold. DBH at maturity was defined with respect to species status and confirmed by literature data when possible: 10 cm for *O. asbeckii*, 25 cm for *D. guianensis*, and 35 cm for *E. falcata*, *E. grandiflora*, *Q. rosea* and *T. melinonii*. Adults included living trees in 2002 and trees either logged during treatment application or that died naturally during the *recovery* period.

2.3. Spatial HBM using latent CAR

The HBM approach accommodates complexity in a high-dimension model through decomposition into a series of simpler conditional hierarchically defined models (Banerjee et al., 2003; Clark, 2005): at a given level, inference conditionally relies on lower-level hypotheses. Three basic levels are mandatory. First, a *data* level specifies the conditional distribution of the data Z given parameters and underlying processes. The hypothesis of conditional independence between observations replaces the classical hypothesis of complete independence. Second, a *process* level specifies the conditional distribution of processes given their own parameters. Third, a *parameter* level specifies the prior distributions of remaining parameters (Wikle, 2003). The purpose of the Bayesian analysis is then to estimate the posterior distribution of the parameters conditional on the data.

A major issue in spatial modelling is to describe correctly the covariance structure of the data. In the sections below, we present a Zero-Inflated Poisson (ZIP) model and its spatial version. The spatial ZIP (SZIP) includes fixed effects and a spatially structured random effect (Fig. 1 a) which models autocorrelation in the response that cannot be explained by fixed effects only. We then briefly describe spatial Poisson models. Finally, we focus on the selection of variables (Fig. 1b), and model calibration and comparison using four criteria.

We modelled the distribution of sapling density as a special case of finite mixture distribution, i.e. the ZIP distribution (Lambert, 1992). In the mixture ZIP model, the distribution of observed data Z follows a mixture of a zero-point mass distribution (modelling structural zeros) and a Poisson distribution $\mathcal{P}(\lambda)$. The model assigns an unknown mass of ω ($0 \leq \omega \leq 1$) to structural zeros and a mass of $(1-\omega)$ to the Poisson distribution. The probability function of the model is

$$\mathbb{P}(Z = z_i|\omega, \lambda) = \begin{cases} \omega + (1-\omega)\mathcal{P}(Z = 0|\lambda) & \text{if } z_i = 0 \\ (1-\omega)\mathcal{P}(Z \neq 0|\lambda) & \text{if } z_i > 0, i = 1, \ldots, n \end{cases}$$

Table 1
Explicative variables derived from a Digitalized Elevation Model (DEM) of Paracou or from census data of trees ≥ 10 cm DBH (*units* in brackets).

Type	Label	Description	Period
Topography	Ele	Elevation (m)	–
	Slo	Slope (°)	
Structure	G_{pio}	Basal area of pioneer taxa (m^2)	2002
	G_{tot}	Total basal area (m^2)	
Logging disturbance	M_{tfL}	Basal area lost in treefalls (m^2)	1986–1988
	M_{sdL}	Basal area lost in standing deaths (m^2)	
Post-logging dynamics	M_{tfR}	Basal area lost in treefalls (m^2)	1989–2002
	A_{tf}	Mean age of treefalls (year)	
	SD_{tfR}	Standard deviation of treefalls ages (year)	
	M_{sdR}	Basal area lost in standing deaths (m^2)	
	dG	Change in basal area (m^2)	
Population variables	dna	Distance to nearest adult (m)	2002
	G_{con}	Basal area of conspecific trees ≥ 10 cm DBH (m^2)	

The period indicates calculus years: 1986–1988 (*logging*) or 1989–2002 (*recovery*). Structure and population variables were calculated in 2002.

Please cite this article in press as: Flores, O., et al., Autocorrelation offsets zero-inflation in models of tropical saplings density. Ecol. Model. (2009), doi:10.1016/j.ecolmodel.2009.01.030

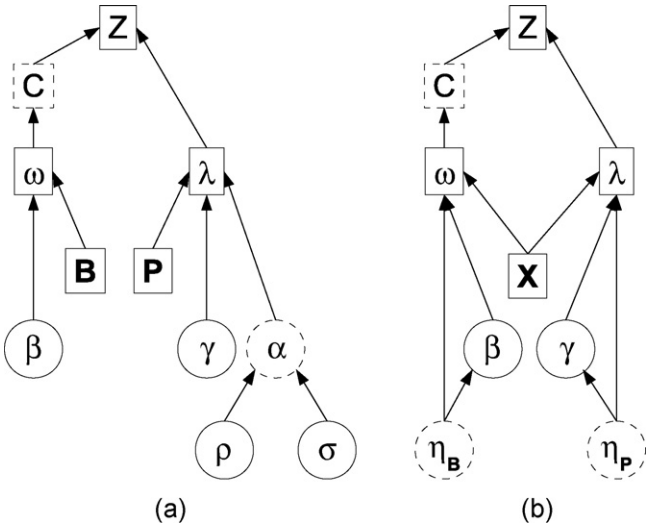


Fig. 1. Directed acyclic graphs of the most complete models: (a) Zero-Inflated Poisson model with random spatial effect, (b) Zero-Inflated Poisson model with binary indicator for variable selection. The models are presented in the four-level HBM scheme including data, process, parameter and hyperparameter levels (Wikle, 2003). Observed or deterministic (defined through an equation, not a distribution) variables are in rectangles. Unknown variables and unknown parameters are in circles. Dashed lines indicate latent variables. Z, observed local sapling density; C, latent binary variable; matrices of explicative variables: $\mathbf{X}$, complete matrices used in variable selection (see Table 1), $\mathbf{X_P}$ and $\mathbf{X_B}$, matrices of selected variables respectively for the Poisson distribution with intensity λ and the binomial distribution with probability ω; γ and β, regression coefficients; $\eta_\mathbf{B}$and $\eta_\mathbf{P}$, latent binary indicators used in variable selection; α, random spatial effect assigned a CAR prior with parameters (ρ, σ; see text for details).

where n is the number of sampling cells, or, using the mixture formulation

$$\mathbb{P}(Z|\omega,\lambda) = \omega \times \delta_0(Z) + (1-\omega)\mathcal{P}(Z|\lambda)$$

where $\delta_0(Z)$ is the Dirac distribution at zero. Here, we introduce a latent (unobserved) random binary variable, C, indicating whether the response Z is structurally null or not. C is modelled as the outcome of a Bernoulli process: $C = 1$ leads to structural zeros (i.e. structural absence), and $C = 0$ indicates that Z follows a Poisson distribution. From Bayes' theorem, the mixture distribution can be expressed as the joint distribution of (Z, C):

$$\mathbb{P}(Z, C|\omega,\lambda) = \mathbb{P}(Z|C=\mathbf{c},\omega,\lambda)\mathbb{P}(C=\mathbf{c}|\omega) = \omega^{\mathbf{c}}[(1-\omega)\mathcal{P}(Z|\lambda)]^{1-\mathbf{c}}$$

At the process level, ω, the probability of a zero being structural, and λ, the intensity of the Poisson process, depends on fixed effects measured by explicative variables through canonical link functions (McCullagh and Nelder, 1989):

$$\text{logit}(\omega) = \mathbf{B}\gamma + \mu \quad (1)$$

$$\log(\lambda) = \mathbf{P}\beta + \alpha \quad (2)$$

where μ and α are two intercepts, $\mathbf{B}$ and $\mathbf{P}$ are two matrices of selected explicative variables (with variables in common or not), and γ and β are two unknown vectors of regression parameters.

In our HBM approach, we extended this ZIP formulation to account for autocorrelation between neighboring observations. We considered that the response variable Z is spatialized, and measured at locations $\mathbf{s}$: $Z = Z(\mathbf{s})$. We assumed that $\alpha(\mathbf{s})$ is a random spatial effect resulting from a spatially structured but unobserved process (see Wikle, 2003). In the SZIP model, at the process level, the Poisson process intensity $\lambda(s)$ thus depended on fixed effects and a random spatial effect:

$$\log[\lambda(\mathbf{s})|\beta,\alpha(\mathbf{s})] = \mathbf{P}\beta + \alpha(\mathbf{s})$$

The intensity of the spatial process, $\alpha(\mathbf{s})$, can be viewed as the spatial component of λ when fixed environmental effects are taken apart: $\alpha = \log(E[y]) - \mathbf{P}\beta$. It is modeled here as a Gaussian random field over a lattice. We used a CAR model (Besag, 1974) for $\alpha(\mathbf{s})$ because observations were sampled on a regular grid and we wanted to account for local autocorrelation. Given a focal location and its neighborhood, the CAR model is interpreted as follows: if the response in the neighborhood gives higher than expected values based on explicative variables, then the focal response will also be locally higher than the expected value. $\alpha(\mathbf{s})$ followed a Gaussian distribution given intensities in a neighborhood:

$$\alpha(s_i)|\alpha(s_j)_{j\in\nu_i} \sim \mathcal{N}\left(\rho\sum_{j\in\nu_i} w_{ij}\alpha(s_j), \sigma^2\right), \quad i = 1,\ldots,n \quad (3)$$

where ρ and τ are two unknown parameters, and (w_{ij}) is a set of spatial weights defining neighborhood relationships (see Banerjee et al., 2003; Wall, 2004 for definition). ρ is a spatial dependence parameter measuring the strength of the relationship between the value of α in a focal cell s_i and in its neighborhood ν_i. σ^2 is the conditional variance. For each cell, we used a Moore neighborhood (the chess king's move).

Finally, the HBM structure of the SZIP model is (Fig. 1a)

data level :	$Z(s)\|\lambda(s) \sim \mathcal{ZIP}[\lambda(s), \omega(s)]$
process level :	$\text{logit}[\omega(\mathbf{s})\|\mu,\gamma] = \mathbf{B}\gamma + \mu$
	$\log[\lambda(\mathbf{s})\|\beta,\alpha(\mathbf{s})] = \mathbf{P}\beta + \alpha(\mathbf{s})$
parameter level :	priors for γ, β and α,
hyperparameter level :	priors for ρ and σ

It is straightforward to obtain spatial and non-spatial Poisson models from this structure.

2.4. Selection of variables

Our selection method, based on that presented in Dellaportas et al. (2002) and Ntzoufras et al. (2000), uses a binary latent variable that indicates which explicative variables are included or not in the model. Let η be the binary latent variable of length p, the number of candidate variables ($p = 13$, Table 1), so that $\eta_j = 1$ indicates that the jth variable is included in the model ($j \in 1,\ldots,p$), whereas $\eta_j = 0$ excludes the variable. A given model is thus characterized by an associated vector η, an additional parameter. The linear predictor $\mathbf{B}\gamma$ in Eq. (1) becomes

$$\sum_{j=1}^{p} X_{ij}\gamma_j\eta_{\mathbf{B}j}, \quad i \in 1,\ldots,n$$

or in matrix form $\mathbf{X}(\gamma\cdot\eta_\mathbf{B})$ where $\cdot$ indicates the dot product, $\mathbf{X}$ is the complete matrix of explicative variables of dimensions (n, p) and the subscript $\mathbf{B}$ refers to the binomial distribution (Fig. 1b). A similar modification applies to the linear predictor $\mathbf{P}\beta$ in Eq. (2). A major benefit of this approach is that the variables space dimension remains constant during the selection unlike in Reversible Jump approaches (Richardson and Green, 1997).

In theory, it is possible to select fixed effects in models with spatial autocorrelation. In practice, several difficulties are encountered. First, parameter inference requires to develop a complex algorithm whose convergence can be difficult to assess. Second, the inclusion of a spatial effect raises identifiability issues: the random effect could counterbalance fixed effects. Third, the selected variables, together with the associated fixed effects, are generally different in spatial models and their non-spatial counterpart (Kneib et al., 2008). In this contribution, we compared fixed effects with and without a random spatial effect, which requires explicative vari-

Please cite this article in press as: Flores, O., et al., Autocorrelation offsets zero-inflation in models of tropical saplings density. Ecol. Model. (2009), doi:10.1016/j.ecolmodel.2009.01.030

ables to be the same in both cases. For these reasons, our variable selection was performed without spatial effect (see Fig. 1b).

2.5. Prior choice

Let $\theta = (\eta, \beta, \gamma, \mathbf{c}, \alpha, \rho, \sigma)$, the complete set of unknown parameters and latent variables in the most complete model. At the parameter level, the definition of weakly informative priors for θ components finalizes the definition of the different models.

- For η, we retained a p-binomial distribution

$$\pi(\eta) = \prod_{j=1}^{p} \tau_j^{\eta_j}(1-\tau_j)^{1-\eta_j}$$

  where τ_j is the probability that the jth variable is present in the model. When no a priori information is available, $\tau_j = (1/2), \forall j \in \{1, \ldots, p\}$, and then $\pi(\eta) = 2^{-p}$.
- With regard to regression parameters (γ, β), two cases were possible. In the selection case, we considered a partition of γ for instance into $(\gamma_\eta, \gamma_{\setminus\eta})$, where γ_η and $\gamma_{\setminus\eta}$ correspond to variables that respectively are included in and excluded from the model. The prior of $\gamma|\eta$ was partitioned into a model prior $\pi(\gamma_\eta|\eta)$ and pseudoprior $\pi(\gamma_{\setminus\eta}|\eta)$ (see Dellaportas et al., 2002). A symmetrical definition followed for β. Without selection, Gaussian priors $\mathcal{N}(0, 100)$ were assumed for γ and β.
- The spatial random effect, α, was assigned a CAR prior as defined previously. The prior for the spatial association coefficient, ρ, was uniform. To ensure that the CAR model has a proper distribution, the ρ parameter needs to be constrained to the interval $[1/\lambda_{min}, 1/\lambda_{max}]$ where λ_{min} and λ_{max} are the minimal and maximal eigenvalues of $\mathbf{D}_w^{-(1/2)}\mathbf{W}\mathbf{D}_w^{-(1/2)}$ (see Banerjee et al., 2003 for details). For $1/\sigma^2$, we used a weakly informative Inverse Gamma distribution $\mathcal{IG}(0.1, 0.1)$.
- In our ZIP models, we used a n-binomial distribution for the latent class variable, C.

2.6. Model estimation and comparison

Four models were retained for each species: a simple GLM, a Spatial Generalized Linear Mixed (SGLM) model, a non-spatial ZIP model, and a SZIP model. We inferred the posterior distribution of θ, $\pi(\theta|\mathbf{z})$ using a Monte-Carlo Markov Chain (MCMC) algorithm. Simulations consisted in sampling θ components along a Markov Chain through a hybrid sampling algorithm of Metropolis-Hastings-within-Gibbs steps (see Agarwal et al., 2002 for a parallel approach).

Model fitting in each species–model combination consisted in two stages. Explicative variables were first selected for each species separately without a spatial effect. We retained variables for which the posterior mean of the corresponding components in $\hat{\eta}$ was greater than 0.75. This indicated that these variables had been retained at least three times out of four along the Markov Chain. In a second stage, regression and spatial parameters were estimated in another MCMC run that included only the selected variables. Each stage consisted of 250,000 iterations from which we discarded a 50,000-iterations burn-in sample. Routines were implemented in C language and run under R (R Development Core Team, 2008). Other analyses were also performed with R.

Predictive power was assessed by simulating independent datasets, which bypasses the need for calibrative and predictive datasets. In HBM, a common problem with model comparisons is the number of degrees of freedom (or effective parameters). Spiegelhalter et al. (2002) suggested comparing hierarchical models by means of a Deviance Information Criterion (DIC) based on deviance moments. However, DIC is not invariant to model parameterization (Spiegelhalter et al., 2002; Celeux et al., 2006; Raftery et al., 2007). In this work, we used the classical Spearman's correlation coefficient, and three Bayesian comparison approaches that are independent of model parameterization (see Appendix B for details about criteria).

The first Bayesian criterion, AICM, is an extension of AIC to Monte-Carlo inference based on the Bayes Factor (Raftery et al., 2007). Second, we computed the posterior predictive loss described by Gelfand and Ghosh (1998), D_1, using replicate data conditional on the posterior distribution of observations (see Appendix B for a detailed definition of criterion). Third, we calculated a posterior predictive p-value (p_{ppc}, see Appendix B) based on the posterior predictive check described by Gelman et al. (1996), which also requires simulated replicates of the data. Values of p_{ppc} that are close to 0 or 1 tend to indicate model rejection (Gelman et al., 1996). The best model should give a p_{ppc} of 0.5. Spearman's coefficient and AICM reflect model goodness-of-fit, whereas D_1 and p_{ppc} address model predictive power (Guisan and Zimmermann, 2000; Banerjee et al., 2003).

3. Results

3.1. Observed densities

Zero-inflation varied across species, with zero-frequencies between 58% for *O. asbeckii* and 87% for *T. melinonii*, compared with 40.3% and 78.2% expected for a Poisson distribution with intensity equal to the average observed density (Table 2). *O. asbeckii* was the most abundant species with 2271 identified saplings. By contrast, *D. guianensis* and *T. melinonii* were the lowest in total numbers (615 and 616) and showed the lowest maximal densities (8 and 11). *Q. rosea* was the most abundant species locally (max.: 34, tot.: 1197) and also the most variable in density. *E. falcata* and *E. grandiflora* occurred in 17% and 20% of the cells, respectively, with 17 and 11 saplings at maximal densities (tot.: 807 and 861).

3.2. Model comparison

Regarding the models' explicative power, the spatial models (SGLM, SZIP) showed a closer agreement between observations and fitted densities that did the non-spatial models (ZIP and GLM) with respect to Spearman's correlation coefficient (ρ_s, Table 3). The spa-

Table 2
Outline of sapling density data for the six focal species at the site.

	O. asbeckii	*E. falcata*	*E. grandiflora*	*D. guianensis*	*Q. rosea*	*T. melinonii*
Σ	2271	807	861	615	1197	616
Max	15	17	11	8	34	11
λ_{obs}	0.908	0.323	0.344	0.246	0.479	0.246
V_{obs}	1.570	1.005	0.923	0.735	1.915	0.870
f_0	58.3	83.3	80.4	84.5	84.6	86.9
$\mathcal{P}_0$	40.3	72.4	70.9	78.2	61.9	78.2

Σ: total number of saplings, Max: maximal observed sapling density in a 10 m × 10 m cell, λ_{obs}, V_{obs}: observed mean and variance of sapling density, f_0: observed frequency of zero counts in sapling density, $\mathcal{P}_0$: expected frequency of zero counts in a Poisson distribution with intensity λ_{obs}.

Please cite this article in press as: Flores, O., et al., Autocorrelation offsets zero-inflation in models of tropical saplings density. Ecol. Model. (2009), doi:10.1016/j.ecolmodel.2009.01.030

Table 3
Comparison statistics of estimated models.

	α	ρ_s		|AICM|		D_1		p_{ppc}	
		P	ZIP	P	ZIP	P	ZIP	P	ZIP
O. asbeckii	∅	0.43	0.44	6556	6135	4756	6000	$\simeq 0$	$\simeq 0$
	CAR	**0.76**	0.74	**6028**	6077	**3379**	3696	**0.73**	0.76
E. falcata	∅	0.48	0.49	3072	2742	1896	2249	$\simeq 0$	$\simeq 0$
	CAR	**0.62**	0.60	2776	**2699**	**1203**	1396	**0.59**	0.74
E. grandiflora	∅	0.41	0.42	3403	**3184**	1743	2097	$\simeq 0$	$\simeq 0$
	CAR	**0.64**	**0.63**	3248	3309	**1274**	1329	**0.67**	**0.66**
D. guianensis	∅	0.28	0.29	3187	2797	1267	1691	$\simeq 0$	$< 10^{-1}$
	CAR	**0.60**	0.57	**2672**	3073	**905**	1153	**0.75**	**0.74**
Q. rosea	∅	0.32	0.32	4523	3390	4675	8757	$\simeq 0$	$\simeq 0$
	CAR	**0.69**	**0.69**	**2591**	2626	**1768**	1858	**0.64**	0.68
T. melinonii	∅	0.20	0.19	3408	2820	1546	2306	$\simeq 0$	$\simeq 0$
	CAR	**0.55**	**0.55**	**2498**	2531	**918**	**919**	**0.60**	**0.61**

ρ_s: Spearman correlation coefficient between observations and fitted values, |AICM|: absolute value of the Akaike Information Criterion Monte-Carlo (Raftery et al., 2007); all computed values were negative, D_1: variance-orientated value of the posterior predictive loss (Appendix B with $k = 1$, Gelman et al., 2004). p_{ppc}: posterior predictive p-value; the closer to 0.5, the better the predictive power of the model (Gelman et al., 1996). The α column differentiates results from non-spatial (∅) and spatial models with a CAR-prior random effect. P indicates Poisson-distributed models, and ZIP indicates Zero-Inflated Poisson models. For each species and each statistics, bold numbers indicate the best models. $\simeq 0$ indicates values $< 10^{-4}$.

tial models gave lower absolute AICM values than the non-spatial models, except in *E. grandiflora*. In the non-spatial models, the marked variability of the log-likelihood along the Markov Chain induced high values for |AICM|. Absolute values of AICM were lowest for SGLM in four species, for the SZIP model in *E. falcata*, and for the ZIP in *E. grandiflora* (Table 3). All ZIP models performed better than the Poisson models (GLM) in the non-spatial case, whereas for spatial models, SZIP gave higher values than SGLM, except in *E. falcata* (Table 3).

Regarding the models' predictive power, the spatial models also performed better than their non-spatial counterpart with regard to the posterior predictive loss function, D_1, and the posterior predictive check, p_{ppc} (Table 3). The Poisson models performed better than the Zero-inflated models, but in all species, the non-spatial models gave values of p_{ppc} that were close to 0, indicating model rejection. Overall, SGLM performed best across all models, and this with respect to both D_1 and p_{ppc} (Table 3).

3.3. Comparison of fitted vs. observed patterns

Moran's I (I_M) was calculated as an indicator of local dependence in sapling density. Here, we used the same neighborhood definition as in the CAR model. All observed spatial patterns showed positive I_M values (Fig. 2) with low variance ($< 10^{-3}$, not shown) indicating positive autocorrelation. Overall, I_M values were higher for fitted than for observed patterns. This finding shows that the modelling process tended to smooth fitted distributions. Still, I_M values in the spatial models (SGLM and SZIP) were closer to observed values than to GLM and ZIP values (Fig. 2). In the non-spatial case, the models also failed to account for local maxima in sapling density (Appendix C).

Empirical variograms were calculated in order to analyze spatial patterns at the site scale. Major change in variograms slopes indicated clumps at various scales (Fig. 3, solid lines). A steep increase was observed up to about 50 m for *D. guianensis* and *E. grandiflora*, and up to about 100 m for *O. asbeckii*. Variograms for *Q. rosea* and *E. falcata* showed a slow increase up to 200 m, with higher variability for *Q. rosea*. The variograms for both species increased after 400 m due to isolated clumps (see maps in Appendix C). In *T. melinonii*, the variogram showed a steep increased in the first 30 m, but the overall spatial structure was less marked in this species.

Variograms calculated on model residuals showed how the models accounted for the spatial structure of sapling density. They were all close to zero and flat in spatial models, indicating no residual autocorrelation (Fig. 3). Overall, the spatial models were able to reproduce the spatial structure of sapling density at the local scale. The spatial structure was also well reproduced at the site scale (see maps in Appendix C). With regard to the non-spatial models (GLM and ZIP), the poor agreement between observed and fitted values induced highly autocorrelated residuals.

3.4. Variables selected and effects

The number of explicative variables chosen during the selection phase ranged from 5 in *D. guianensis* (Fig. 4) to 11 (Poisson models in *O. asbeckii*). Most of the variables selected were common across Poisson and zero-inflated models although differences arose in all species (Fig. 4). Overall, the selection procedure retained fewer explicative variables in zero-inflated than in Poisson models. Some variables retained in Poisson models had no influence on sapling

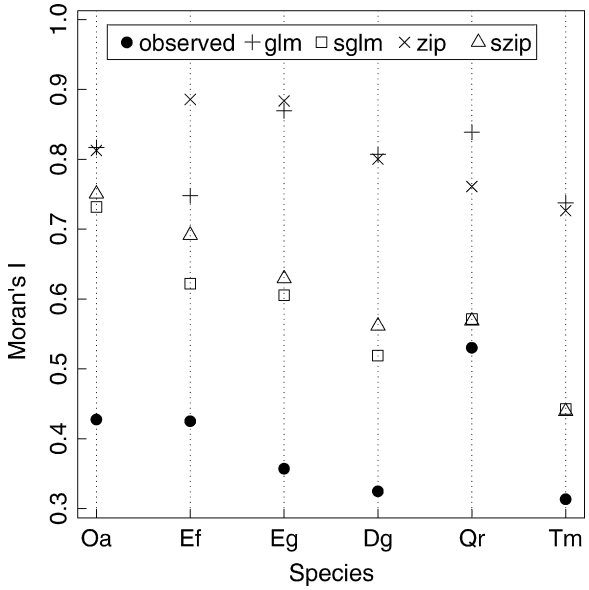


Fig. 2. Moran's I of observed and fitted sapling patterns.

Please cite this article in press as: Flores, O., et al., Autocorrelation offsets zero-inflation in models of tropical saplings density. Ecol. Model. (2009), doi:10.1016/j.ecolmodel.2009.01.030

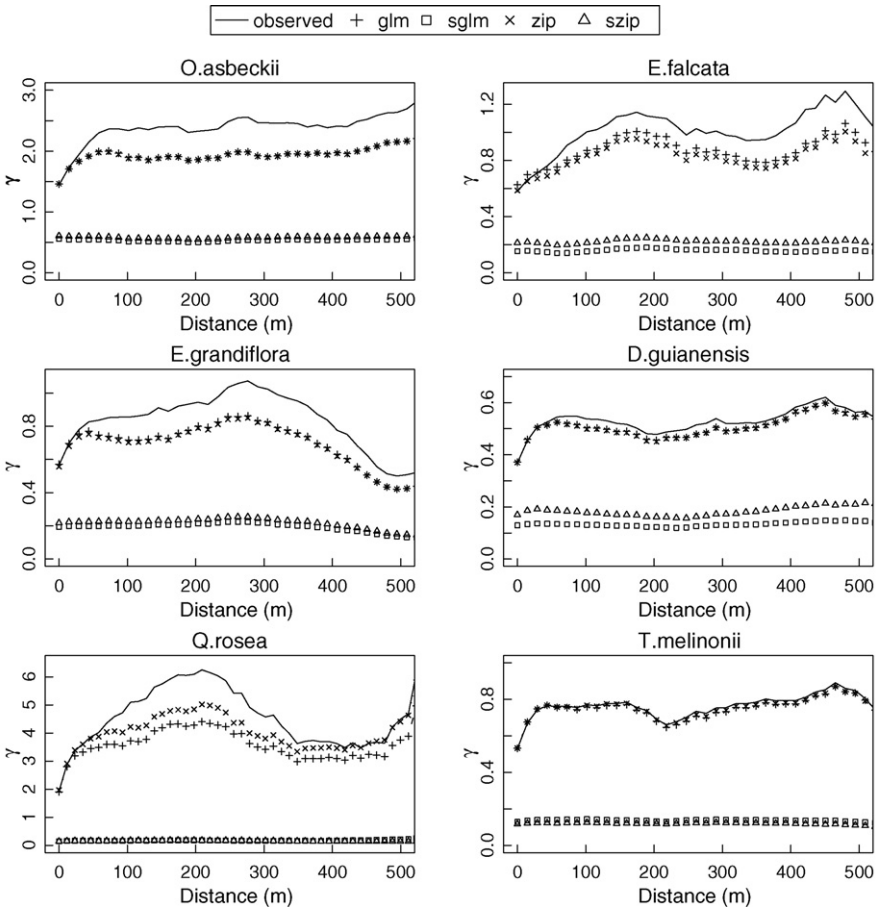


Fig. 3. Spatial structure at the site scale. For each species, the solid line shows the empirical variogram of sapling density (observed), while symbols show variograms calculated on the residuals of the four models (GLM, SGLM, ZIP, SZIP).

density when the zero-inflated distribution was used (e.g. G_{pio} in *E. falcata*, Fig. 4). In fewer cases, variables not retained in the Poisson models were retained in the zero-inflated models (e.g. G_{con} in *E. grandiflora* and *D. guianensis*, G_{con} in *D. guianensis*, G_{pio} in *Q. rosea* and SD_{tfR} in *T. melinonii*).

Each of the 13 explicative variables was retained at least once during the selection phase, and thus partly explained sapling density. Topographic variables, elevation, slope or both, were retained in *O. asbeckii*, *E. grandiflora*, *D. guianensis* and *Q. rosea* (Fig. 4). Structural variables (G_{con} and G_{pio}) were retained in all models except zero-inflated models in *E. grandiflora* and Poisson models in *Q. rosea*. At least one variable characterizing disturbance was selected in all models. Population variables were also retained in all models except in the SZIP in *T. melinonii* (Fig. 4). dna was not retained only in *T. melinonii*.

The comparison of spatial and non-spatial models with similar distributions showed that parameter estimation was substantially altered when autocorrelation was included. The effects of variables, measured here by the posterior mean of the associated regression parameters, generally decreased or reached zero (Fig. 4). Here, we focus on *D. guianensis* which had the fewest selected variables. In the best model for this species (SGLM), the most influent variables were elevation (*Ele*) and distance to nearest adult (*dna*). These had respectively a positive and a negative (decrease with increasing distance) influence on sapling density. These findings indicate a preference for an upper-slope/plateau position and limited dispersal around adults. The other variables retained were SD_{tfR}, G_{tot}, and G_{pio}. The sign of effects indicated that sapling density was more elevated in cells where treefalls were scattered over time, with a low total basal area and a low basal area of pioneer taxa, suggesting conditions of intermediate disturbance intensity.

4. Discussion

This study compared spatial Zero-Inflated Poisson models of sapling local density in a tropical forest with classical GLMs. Overall, model performance was enhanced when accounting for spatial dependence. The CAR model proved well-suited to account for autocorrelation between adjacent cells. The conditional nature of the CAR model makes it relevant for HBM, and HBM does appear to be particularly well adapted in our context. In the spatial models, the residuals appeared to be uncorrelated, showing that the spatial structure of sapling density was relevantly addressed at the local scale. Posterior estimates of the dependence parameter (ρ) were close to its space boundary, suggesting that alternative models could be used. For instance, the Simultaneous Auto-Regressive (SAR) model is formally equivalent to a CAR, but with a different covariance structure (see Keitt et al., 2002; Wall, 2004 for comparisons). Other possibilities include geostatistical models which primarily rely on a continuous description of space. However, auto-regressive models are better suited for the study of area-based data, especially on a regular lattice (Banerjee et al., 2003).

The SGLM showed greater explicative and predictive power than the other models in our case study. Adding a latent spatial effect at the process level of HBM was sufficient to handle both spatial autocorrelation and zero-inflation. We assumed that this was possible because the zero observations were autocorrelated in space (see maps in Appendix C). We suspect that a zero-inflated model would

Please cite this article in press as: Flores, O., et al., Autocorrelation offsets zero-inflation in models of tropical saplings density. Ecol. Model. (2009), doi:10.1016/j.ecolmodel.2009.01.030

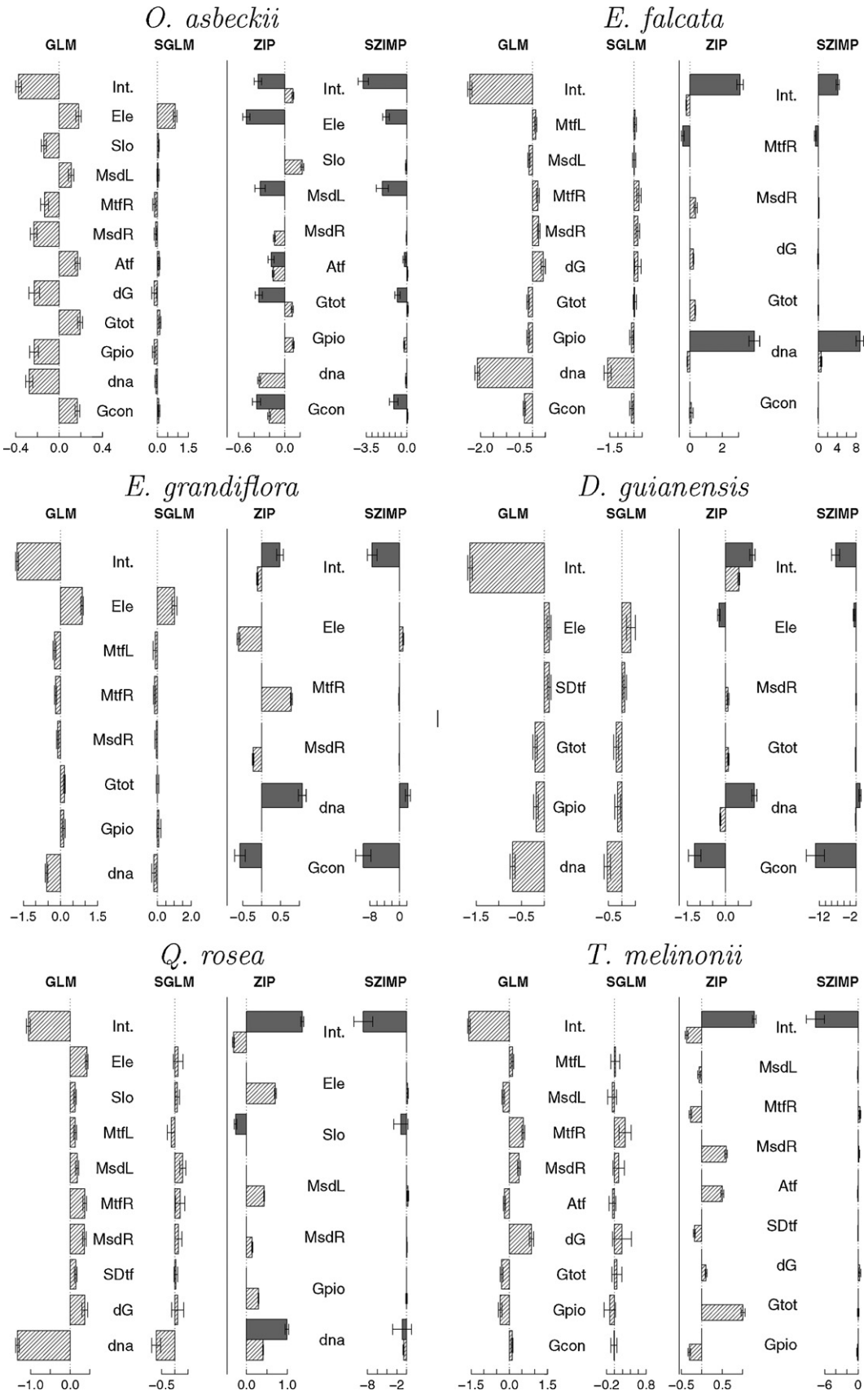


Fig. 4. Explicative variables: selection and effects. The figure shows, for each species and each of the four studied models, the posterior means and standard deviation intervals of regression parameters associated with selected variables (see text for the variable selection procedure and Table 1 for labels, Int.: intercepts). Shaded bars relate to variables included in the Poisson distribution (matrix **X**), filled bars relate to coefficients of variables included in the binomial distribution (matrix **B**).

Please cite this article in press as: Flores, O., et al., Autocorrelation offsets zero-inflation in models of tropical saplings density. Ecol. Model. (2009), doi:10.1016/j.ecolmodel.2009.01.030

be more efficient in cases of data with uncorrelated structural zeros.

In the SZIP model, we included autocorrelation in the Poisson process, which produces random zeros and non-zero counts. An alternative approach could account for dependence in the probability of observations to be structural zeros (ω). This may be justified in species with strong habitat specificity, for instance in species exclusively found in waterlogged areas. However, such model structure leads to instability and poor parameter estimates (Agarwal et al., 2002). Specific sophisticated algorithms are required to address this issue.

Zero-inflation may be induced by a number of causes in vegetation data. These include scarcity of the studied plants and sampling variability, but also ecological constraints such as habitat unsuitability or marked clumping. ZIP models have the advantage at accounting for these processes and they also allow flexibility in the shape of the response curve, a critical issue in studies of species patterns (Guisan and Zimmermann, 2000; Oksanen and Minchin, 2002). The two-component or Hurdle model is often advocated when facing the mixture specification because parameter interpretation is easier in this case. We preferred the mixture specification for three reasons. First, in the mixture case, the response curve to a given predictor can be easily calculated (Flores et al., 2006). Second, assuming that the processes leading to zero and non-zero data are independent may not be relevant to the ecological model. For instance, habitat suitability is not a binary factor: individuals surviving in transient habitats imply non-null density. Likewise, dispersal may induce structural zeros beyond a limiting distance, though infrequent long distance dispersal events occur. Dispersal also implies random zeros as seeds do not saturate a tree's influence area, because of stochasticity. Third, the mixture specification separates effects leading to structural and random zeros. Selected variables can influence either the binary, or the Poisson process, or both.

The saplings in our study appeared to be clumped, which may be due to limited dispersal around adults (Svenning, 2001), clumped seed dispersal (Howe, 1989; Russo and Augspurger, 2004), or a survival response to patchy resources (Dalling et al., 1998). Svenning et al. (2006) interpreted the high contribution made by the CAR component to local density as evidence of strong local dispersal. Clearly, aggregate dispersal is likely to induce local autocorrelation in species patterns. However, we would expect the CAR component to contribute differently across species that display different dispersal modes. No such findings were observed. In light-demanding species (*e.g. T. melinonii*), autocorrelation may reflect unobserved environmental heterogeneity induced by unobserved disturbance events (Nicotra et al., 1999).

In our study, we characterized the environment by means of continuous descriptors of ecological processes such as disturbance. This approach addressed a common issue in modelling, i.e. a disequilibrium between observed patterns and current environmental conditions (Guisan and Zimmermann, 2000; Austin, 2002). Overall, non-null effects were detected in each model-species combination, and selected variables changed across species. Designed variables thus all quantified some aspect of environmental heterogeneity or population process that partly explained sapling density and indicated specific processes. The position of adults influenced sapling patterns in five species. Whereas no such effect was seen for the anemochorous and most light-demanding species *T. melinonii*. Despite mortality filters on earlier stages, a dispersal signal persisted in sapling patterns (Clark et al., 1999; Wang and Smith, 2002). The studied species are known to be rather poor dispersers, like the bird-dispersed *O. asbeckii* (Ulft, 2004), the autochorous and barochorous *Eperua* and *D. guianensis* despite wind dispersal. Rodents secondarily dispersing seeds can increase dispersal distances (Forget, 1992).

Variables of past disturbance patterns were particularly informative, and this is consistent with previous studies of the spatial heterogeneity of light in tropical forests (Nicotra et al., 1999). Overall, disturbance effects were consistent with species shade-tolerance. Species recruitment was differentially affected by disturbance-induced opening of the canopy. Regarding topography, *D. guianensis* and *E. grandiflora* are known to mainly settle on the upper part of slopes and *E. falcata* on bottomlands. Here, *E. falcata* was weakly affected by topography. In this species, population variables were sufficiently informative to mask the effects of physical conditions because of the marked clumping of saplings around adults. This finding raises the issue of covariance between explicative variables. Here, we selected explicative variables from candidates based on an efficient and stable selection method. We used a prior that favors models with $p/2$ variables. Other choices are possible that would take account of covariance between variables. The influence of such priors on statistical and ecological inference remains to be tested.

In modelling studies, the trade-off between model complexity and relevance is a well-known issue. In our case, the SZIP model appeared to be conceptually relevant as it could account for two critical features of data, i.e. autocorrelation and zero-inflation. Empirically, the SGLM model appeared to show the best performance. This finding shows that refining processes addressed at the process level of HBM could compensate for statistical dispersion observed in the data. In other words, a priori required complexity at the data level was not necessary when accurate specification occurred at process level. In a predictive context, statistical simplicity may be preferred. Sophisticated models may nevertheless be required to evidence hidden biological processes.

Acknowledgements

We wish to thank Lilian Blanc, Jean-Gaël Jourget, Pascal Pétronelli (CIRAD, Kourou, French Guiana) and the Paracou field workers who participated in collecting the data. We also extend our thanks to Sylvie Gourlet-Fleury and Hélène Dessard for constructive discussions and helpful comments.

Appendix A. Site map

See Fig. 5.

Appendix B. Model comparison

The Bayes Factor (BF) is among the common approaches used for model comparison. It is based on the integrated posterior harmonic mean of the likelihood:

$$\pi(Z) = \int f(Z|\theta)\pi(\theta)\,d\theta.$$

which can be approximated by the harmonic mean of the likelihood along a standard Markov Chain Monte-Carlo run (Raftery et al., 2007). Although $\pi(Z)$ is consistent as the simulation size increases, its precision is not guaranteed. Raftery et al. (2007) proposed the use of a shifted gamma estimator which leads to modified versions of AIC and BIC. We retained the AICM (M for Monte-Carlo) which addresses model explicative power, and is defined as

$$\mathrm{AICM} = 2(\hat{l} - s_l^2)$$

where $\hat{l}$ and s_l^2 are the mean and variance of the log-likelihood along the chain.

Please cite this article in press as: Flores, O., et al., Autocorrelation offsets zero-inflation in models of tropical saplings density. Ecol. Model. (2009), doi:10.1016/j.ecolmodel.2009.01.030

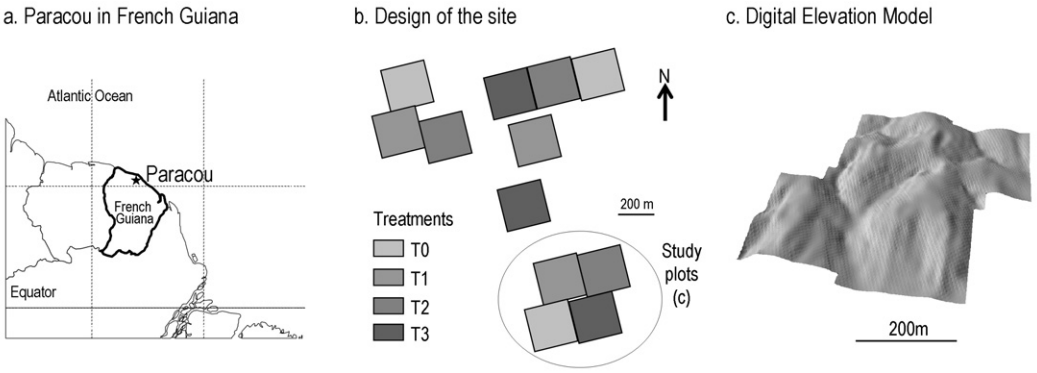


Fig. 5. Location and map of the study site.

An alternative approach is the posterior predictive loss described by Gelfand and Ghosh (1998) which addresses model predictive power. It uses the distribution of replicate data conditional on the posterior distribution of observations (the posterior predictive distribution). We note $\mathbf{z}^{\text{rep}}$ a replicate dataset simulated with sampled values of parameters θ along the Markov chain. These data could have been observed under the studied model with those values of θ (Gelman et al., 1996). Conditional on θ, $\mathbf{z}^{\text{rep}}$ and $\mathbf{z}$ are assumed to be independent. The posterior predictive distribution of replicates is then

$$p(\mathbf{z}^{\text{rep}}|\mathbf{z}) = \int p(\mathbf{z}^{\text{rep}}|\theta, \mathbf{z})p(\theta|\mathbf{z})\,d\theta$$

The best model then minimizes the posterior predictive loss defined as

$$D_k = \frac{k}{k+1}G + P$$

where

$$G = \sum_{i=1}^{n}(\hat{\mu}_i - z_i)^2 \quad \text{and} \quad P = \sum_{i=1}^{n}\hat{\sigma}_i^2$$

where $\hat{\mu}_i$ and $\hat{\sigma}_i^2$ are the mean and variance of the posterior predictive distribution. The loss function D_k reflects the classical compromise between bias and variance, depending on the choice of k: G and P are, respectively, the bias (the goodness of fit) and the variance of the prediction. In the paper, we use the variance-orientated loss function $D_1 = (G/2) + P$ ($k = 1$) as a second Bayesian criterion. In order to estimate D_1, we simulate 100 replicated data for each of 1000 values of θ sampled along the chain ($\theta^{(k)}$, $k = 1, \ldots, 1000$).

Finally, we derived a last criterion of model validation using the posterior predictive check approach (Gelman et al., 1996). This approach also requires simulated replicates of the data. A discrepancy measure based on the residual sum of squares, $T(\mathbf{z}, \theta)$, quantified model fit:

$$T(\mathbf{z}, \theta^{(k)}) = \sum_i (z_i - E[z_i|\theta^{(k)}])^2$$

where (k) indicates values sampled along the chain.

The goodness-of-fit of a model is then evaluated by comparing the posterior distribution of $T(\mathbf{z}, \theta^{(k)})$ with the posterior predictive reference distribution $T(\mathbf{z}^{\text{rep}}, \theta^{(k)})$ (Stern and Cressie, 2000). We quantified the closeness of two discrepancy measures based on parameters estimates and either the observations or a simulated replicate dataset. Graphically, scattering away from the 1:1 line in the plot of $T(\mathbf{z}, \theta^{(k)})$ and $T(\mathbf{z}^{\text{rep}}, \theta^{(k)})$ indicates that data generated by the model greatly differ from the observed data, with respect to T. Numerically, this information can be summarized by a posterior predictive p-value:

$$p_{\text{ppc}} = \mathbb{P}[T(\mathbf{z}^{\text{rep}}, \theta) \geq T(\mathbf{z}, \theta)]$$

Values close to 0 or 1 tend to indicate model rejection (Gelman et al., 1996). In order to estimate p_{ppc}, we use the approximation:

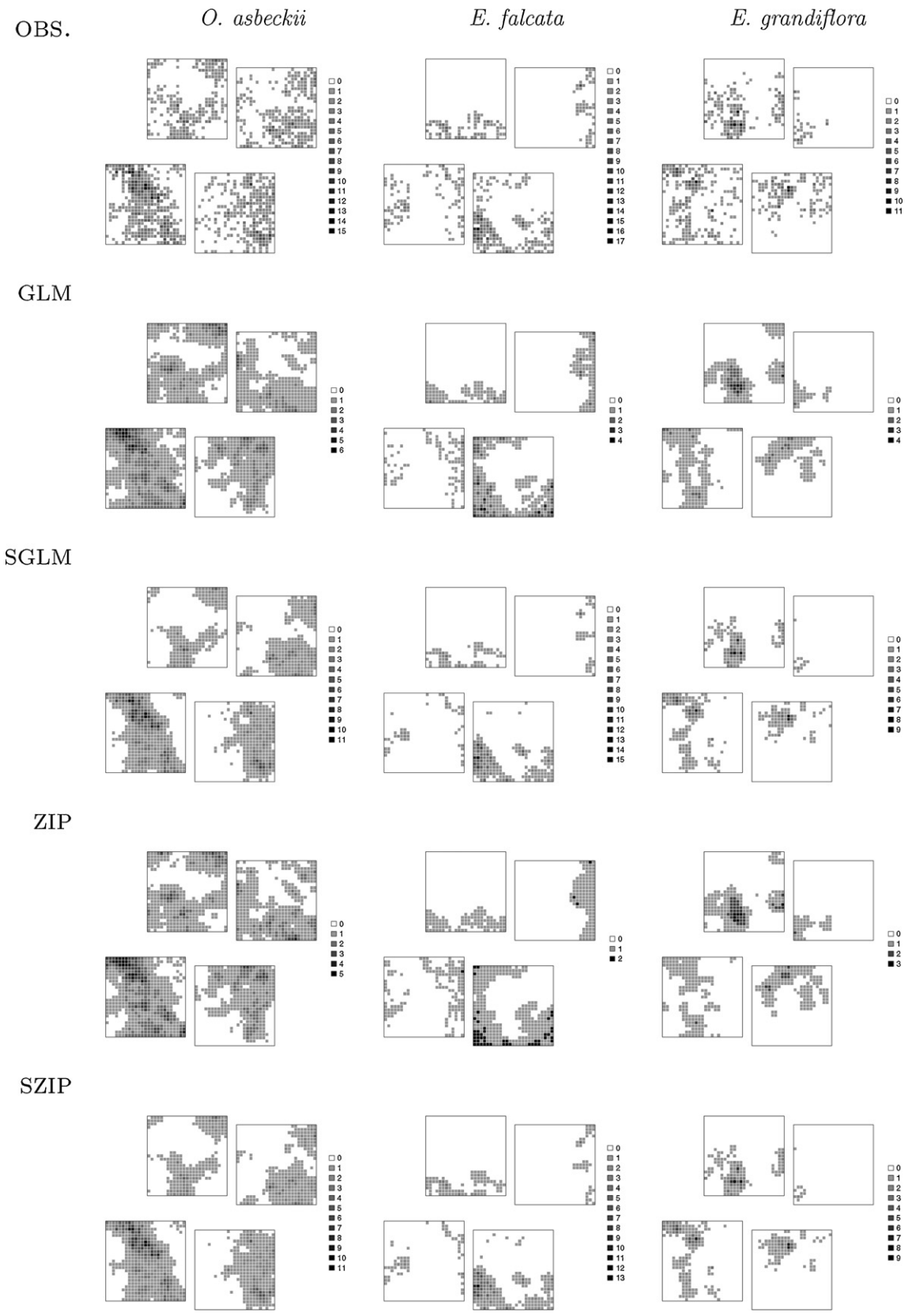
$$p_{\text{ppc}} = \sum_k \sum_j \mathbb{I}_{T(\mathbf{z}_j^{\text{rep}(k)}, \theta^{(k)}) \geq T(\mathbf{z}, \theta^{(k)})}$$

where j indicates a simulated dataset using parameters $\theta^{(k)}$ ($j = 1, \ldots, 100$).

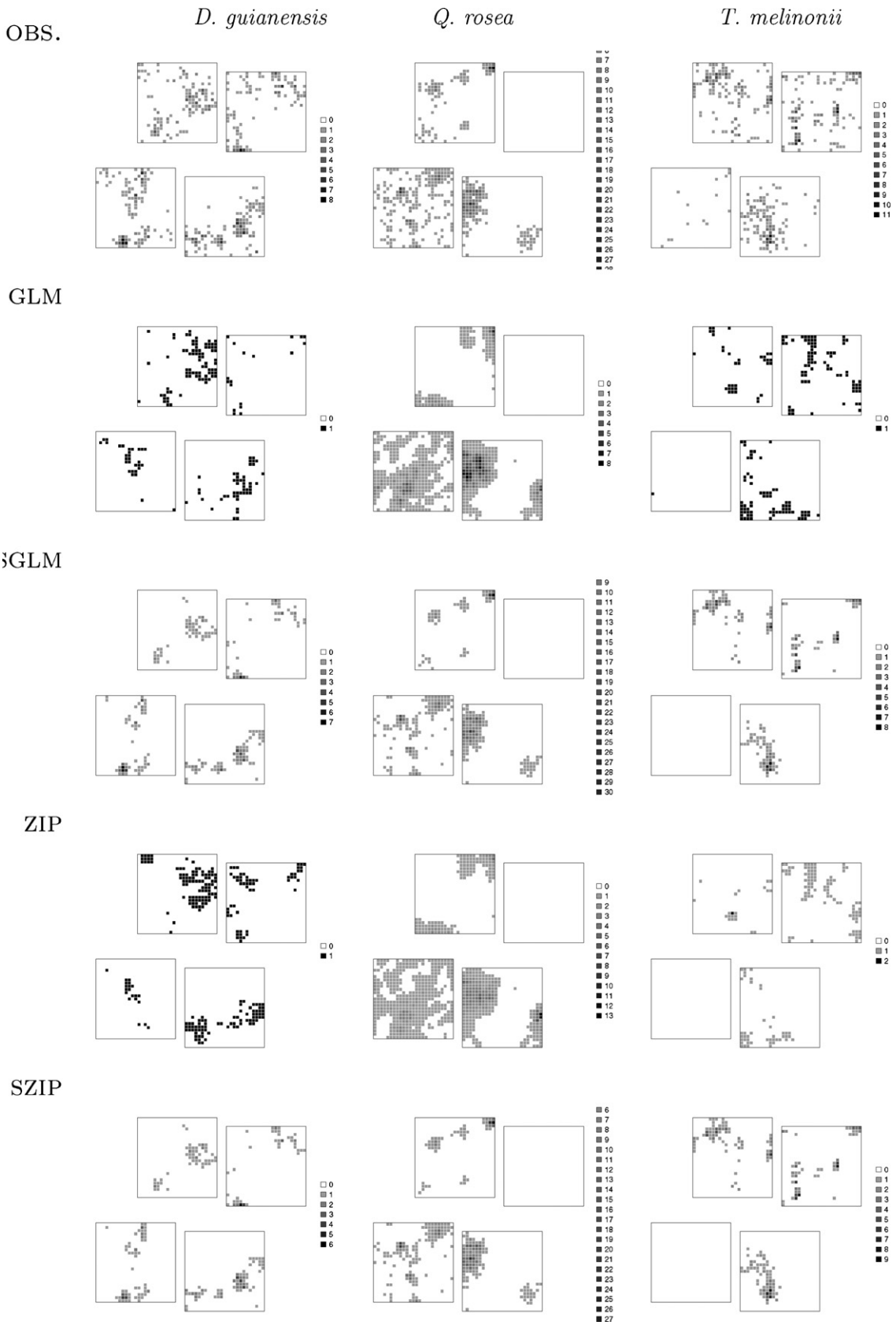
Another common criterion in the comparison of Bayesian models is the Deviance Information Criterion (DIC, Spiegelhalter et al., 2002). However, when hidden structures and random effects are addressed, the definition of the posterior estimates of parameters, $\bar{\theta}$ is not fixed so that DIC depends on model parametrization Spiegelhalter et al. (2002), Celeux et al. (2006), and Raftery et al. (2007), and on a certain focus on the hierarchy (Plummer, 2006). Although Celeux et al. (2006) proposed several versions of the DIC, none seems to be well suited to such cases (see discussion in Celeux et al., 2006 paper).

Please cite this article in press as: Flores, O., et al., Autocorrelation offsets zero-inflation in models of tropical saplings density. Ecol. Model. (2009), doi:10.1016/j.ecolmodel.2009.01.030

Appendix C. Density maps



Please cite this article in press as: Flores, O., et al., Autocorrelation offsets zero-inflation in models of tropical saplings density. Ecol. Model. (2009), doi:10.1016/j.ecolmodel.2009.01.030



Please cite this article in press as: Flores, O., et al., Autocorrelation offsets zero-inflation in models of tropical saplings density. Ecol. Model. (2009), doi:10.1016/j.ecolmodel.2009.01.030

References

Agarwal, D.K., Gelfand, A.E., Citron-Pousty, S., 2002. Zero-inflated models with application to spatial count data. Environ. Ecol. Stat. 9 (4), 341–355.
Angers, J.F., Biswas, A., 2003. A Bayesian analysis of zero-inflated generalized Poisson model. Comput. Stat. Data Anal. 42 (1–2), 37–46.
Austin, M., 2002. Spatial prediction of species distribution: an interface between ecological theory and statistical modelling. Ecol. Model. 157 (2–3), 101–118.
Austin, M., 2007. Species distribution models and ecological theory: a critical assessment and some possible new approaches. Ecol. Model. 200 (1–2), 1–19.
Banerjee, S., Carlin, B., Gelfand, A., 2003. Hierarchical modeling and analysis for spatial data. In: Monographs on Statistics and Applied Probability, vol. 101. Chapman & Hall/CRC.
Barry, S., Welsh, A., 2002. Generalized additive modelling and zero inflated count data. Ecol. Model. 157 (2–3), 179–188.
Besag, J., 1974. Spatial interaction and the statistical analysis of lattice systems. J. Roy. Stat. Soc. Ser. B 36, 192–236.
Celeux, G., Forbes, F., Robert, C., Titterington, M., 2006. Deviance information criteria for missing data models. Bay. Anal. 1, 651–674.
Clark, J., 2005. Why environmental scientists are becoming Bayesians? Ecol. Lett. 8, 2–14.
Clark, J., Beckage, B., Camill, P., Cleveland, B., Hillerislambers, J., Lichter, J., McLachlan, J., Mohan, J., Wyckoff, P., 1999. Interpreting recruitment limitation in forests. Am. J. Bot. 86 (1), 1–16.
Condit, R., Ashton, P., Baker, P., Bunyavejchewin, S., Gunatilleke, S., Gunatilleke, N., Hubbell, S., Foster, R., Itoh, A., LaFrankie, J., Lee, H., Losos, E., Manokaran, N., Sukumar, R., Yamakura, T., 2000. Spatial patterns in the distribution of tropical tree species. Science 288, 1414–1418.
Dalling, J., Hubbell, S., Silvera, K., 1998. Seed dispersal, seedling establishment and gap partitioning among pioneer tropical trees. J. Ecol. 86, 674–689.
Dellaportas, P., Forster, J.J., Ntzoufras, I., 2000. Bayesian variable selection using the Gibbs sampler. In: Dey, D.K., Ghosh, S.K., Mallick, B.K. (Eds.), Generalized Linear Models: A Bayesian Perspective. Chemical Rubber Company Press, New York, USA, pp. 273–286.
Dellaportas, P., Forster, J., Ntzoufras, I., 2002. On Bayesian model and variable selection using MCMC. Stat. Comput. 12 (1), 27–36.
Flores, O., Gourlet-Fleury, S., Picard, N., 2006. Local disturbance, forest structure and dispersal effects on sapling distribution of light-demanding and shade-tolerant species in a French Guianan forest. Act. Oec. 29 (2), 141–154.
Forget, P.-M., 1992. Regeneration ecology of *Eperua grandiflora* (Caesalpiniaceae), a large-seeded tree in French Guiana. Biotropica 24 (2a), 146–156.
Gelfand, A.E., Ghosh, S.K., 1998. Model choice: a minimum posterior predictive loss approach. Biometrika 85 (1), 1–11.
Gelman, A., Carlin, B., Stern, H., Rubin, D.B., 2004. Bayesian Data Analysis. CRC Press.
Gelman, A., Meng, X.-L., Stern, S., 1996. Posterior predictive assessment of model fitness via realized discrepancies. Stat. Sin. 6, 733–807.
Guisan, A., Edwards, T.J., Hastie, T., 2002. Generalized linear and generalized additive models in studies of species distributions: setting the scene. Ecol. Model. 157 (2–3), 89–100.
Guisan, A., Zimmermann, N.E., 2000. Predictive habitat distribution models in ecology. Ecol. Model. 135 (2/3), 147–186.
Howe, H., 1989. Scatter- and clump-dispersal and seedling demography: hypothesis and implications. Oecologia 79, 417–426.
Keitt, T., Bornstad, O., Dixon, P., Citron-Pousty, S., 2002. Accounting for spatial pattern when modeling organism–environment interactions. Ecography 25, 616–625.
Kneib, T., Hothorn, T., Tutz, G., 2008. Variable selection and model choice in geoadditive regression models. Biometrics on line.
Kuhnert, P.M., Martin, T.G., Mengersen, K., Possingham, H.P., 2005. Assessing the impacts of grazing levels on bird density in woodland habitat: a Bayesian approach using expert opinion. Environmetrics 16 (7), 717–747.
Lambert, D., 1992. Zero-inflated Poisson regression, with an application to defects in manufacturing. Technometrics 34, 1–14.
Legendre, P., 1993. Spatial autocorrelation: trouble or new paradigm? Ecologia 74 (6), 1659–1673.
Lichstein, J., Simons, T., Shriner, S., Franzreb, K., 2002. Spatial autocorrelation and autoregressive models in ecology. Ecol. Monogr. 72 (3), 445–463.
Martin, T.G., Wintle, B.A., Rhodes, J.R., Kuhnert, P.M., Field, S.A., Low-Choy, S.J., Tyre, A.J., Possingham, H.P., 2005. Zero tolerance ecology: improving ecological inference by modelling the source of zero observations. Ecol. Lett. 8 (11), 1235–1246.
McCullagh, P., Nelder, J., 1989. Generalized linear models. In: Monographs on Statistics and Applied Probability, vol. 37, Chapman & Hall edition. Chapman & Hall/CRC, London.
Miller, J., Franklin, J., 2002. Modeling the distribution of four vegetation alliances using generalized linear models and classification trees with spatial dependence. Ecol. Model. 157 (2–3), 227–247.
Miller, J., Franklin, J., Aspinall, R., 2007. Incorporating spatial dependence in predictive vegetation models. Ecol. Model. 202 (3–4), 225–242.
Moisen, G., Frescino, T., 2002. Comparing five modelling techniques for predicting forest characteristics. Ecol. Model. 157 (2–3), 209–225.
Nicotra, A.B., Chazdon, R.L., Iriarte, S.V.B., 1999. Spatial heterogeneity of light and woody seedling regeneration in tropical forests. Ecologia 80 (6), 1908–1926.
Ntzoufras, I., Forster, J.J., Dellaportas, P., 2000. Stochastic search variable selection for log-linear models. J. Stat. Comput. Simul. 68 (1), 23–37.
Oksanen, J., Minchin, P., 2002. Continuum theory revisited: what shape are species responses along ecological gradients? Ecol. Model. 157 (2–3), 119–129.
Plummer, M., 2006. Comment on article by Celeux et al. Bay. Anal. 1 (4), 681–686.
R Development Core Team (2008). R: A Language and Environment for Statistical Computing. ISBN 3-900051-07-0.
Raftery, A., Newton, M., Satagopan, J., Krivitsky, P., 2007. Estimating the integrated likelihood via posterior simulation using the harmonic mean identity. Bay. Stat. 8, 1–45.
Richardson, S., Green, P.J., 1997. On Bayesian analysis of mixtures with an unknown number of components. J. Roy. Stat. Soc. Ser. B 59 (4), 731–792.
Ridout, M., Demetrio, C., Hinde, J., 1998. Models for count data with many zeros. In: Proceedings of the International Biometric Conference, Cape Town.
Russo, S., Augspurger, C., 2004. Aggregated seed dispersal by spider monkeys limits recruitment to clumped patterns in *Virola calophylla*. Ecol. Lett. 7 (11), 1058–1067.
Sabatier, D., 1983. Fructification et dissmination en fort guyanaise - L'exemple de quelques espces ligneuses. Doctorat de 3?me cycle, Universit des Sciences et Techniques du Languedoc.
Spiegelhalter, D., Best, N., Carlin, B., Van der Linde, A., 2002. Bayesian measures of model complexity and fit (with discussion). J. Roy. Stat. Soc. Ser. B 6, 583–639.
Stephenson, C., MacKenzie, M., Edwards, C., Travis, J., 2006. Modelling establishment probabilities of an exotic plant, *Rhododendron ponticum*, invading a heterogeneous, woodland landscape using logistic regression with spatial autocorrelation. Ecol. Model. 193 (3–4), 747–758.
Stern, H., Cressie, N., 2000. Posterior predictive model checks for disease mapping models. Stat. Med. 19, 2377–2397.
Svenning, J.-C., 2001. Environmental heterogeneity, recruitment limitation and the mesoscale distribution in a tropical Montane rain forest (maquipucuna, ecuador). J. Trop. Ecol. 17, 97–113.
Svenning, J.-C., Engelbrecht, B.M.J., Kinner, D.A., Kursar, T.A., Stallard, T.A., Wright, S.J., 2006. The relative roles of environment, history and local dispersal in controlling the distributions of common tree and shrub species in a tropical forest landscape, panama. J. Trop. Ecol. 22, 575–586.
Ulft, L.v., 2004. Regeneration in natural and logged tropical rain forest. Modelling seed dispersal and regeneration of tropical trees in Guyana. In: Volume 12 of Tropenbos-Guyana Series. Tropenbos International, Georgetown.
Wall, M., 2004. A close look at the spatial structure implied by the CAR and the SAR models. J. Stat. Plan. Inf. 121, 311–324.
Wang, B., Smith, T., 2002. Closing the seed dispersal loop. Trend Ecol. Evol. 17 (8), 379–385.
Welsh, A., Cunningham, R., Donnelly, C., Lindenmayer, D., 1996. Modelling the abundance of rare species: statistical models for counts with extra zeros. Ecol. Model. 88 (1–3), 297–308.
Wikle, C., 2003. Hierarchical Bayesian models for predicting the spread of ecological processes. Ecologia 84, 1382–1394.

Please cite this article in press as: Flores, O., et al., Autocorrelation offsets zero-inflation in models of tropical saplings density. Ecol. Model. (2009), doi:10.1016/j.ecolmodel.2009.01.030