Estimating species richness from quadrat sampling data: a general approach

Jérôme Dupuis, Goulard Michel

▶ **To cite this version:**

Jérôme Dupuis, Goulard Michel. Estimating species richness from quadrat sampling data: a general approach. 2010. hal-00503260

HAL Id: hal-00503260
https://hal.science/hal-00503260

Preprint submitted on 18 Jul 2010

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Estimating species richness from quadrat sampling data: a general approach

Jérôme Dupuis*

IMT, Laboratoire de Statistique et Probabilités, Université Toulouse III, France

and

Michel Goulard

INRA, Dynafor, UMR 1201, Castanet Tolosan, Toulouse, France

SUMMARY. We consider the problem of estimating the number of species (denoted by S) of a biological community located in a region composed of n quadrats. To address this question, different parametric approaches have been recently developed. However, they all have some limitations which reduce their use in practice: indeed, either they presuppose that the sampled quadrats are taken from a large population of quadrats (theoretically infinite), or they require an upper bound on S. Our approach is more general in that it applies without limitation on n and it can be used in the presence of prior information on S, as well as in totally unknown regions. We pay attention to the prior adopted for S; in particular, different non informative priors are considered and motivated. We first consider a simple model which assumes that occurrence and detectability parameters do not depend on quadrats. It constitutes a suitable framework to clarify the links existing between current approaches and ours. We then extend this model by assuming that the region of study is spatially heterogeneous. We illustrate our approach by estimating the number of species of a birds community located in a forest.

KEY WORDS: Bayesian estimate; Biodiversity, Jeffreys prior; Missing data; Quadrat sampling; Species richness.

---

* *email:* dupuis@math.ups-tlse.fr

1. Introduction

The species richness of a community of animals or plants - that is the number of species present within this community - is a basic measure of its bio-diversity (Huston, 1994). Estimating the species richness (denoted by S) of a biological community located in some specified region, called afterwards R, often relies on quadrat sampling data (Krebs, 1989). Assume that that the region R is composed of n quadrats, inference on S is thus based on a sample of quadrats (of size $T < n$). Of course, the region R has to be bounded, or else estimating S is an ill-posed problem (eg Dorazio *et al.,* 2006). Nonparametric methods have been used for estimating S when quadrat sampling is used. They include the Jacknife and the boostrap estimates; see Mingoti and Meden (1992); Bunge and Fitzpatrick (1993), Chao (2005), Hass *et al.* (2006).

Recently, parametric approaches have been developed for estimating species richness from quadrat sampling. They are all characterized by a hierarchical modeling of the data which separates assumptions related to occurrences of species in the quadrats and those related to their detectability (this constitutes an interesting characteristic of these approaches, compared to the nonparametric approaches). Three approaches can be distinguished according to whether an upper bound on S is required or not, and according to whether the number n of quadrats which composes R is assumed to be large (in theory infinite) or not. From here on, it is important to point out that, in practice, n is not necessarily large; in fact, its value strongly varies from one survey to another (see Section 2). We recall these three approaches below, and make clear the limitation(s) of each.

Dorazio and Royle (2005) have developed an approach (afterwards called the DR approach) which does not require an upper bound on S, but assumes that the number of quadrats which composes R is theoretically infinite (we will write $n = \infty$, for convenience). Consistently, n is not a part of the model, and Dorazio & Royle (2005) define the

species richness S as the limit of S_T when $T \longrightarrow \infty$, where S_T represents the number of species present in T sampled quadrats (or as the asymptote of a cumulative species-area curve). This approach is thus not appropriate to deal with the problem examined in this paper (recall that we assume nothing regarding the size of n).

Dupuis and Joachim (2006) have developed an approach which has no limitation on n, but requires to have some information on S; in particular, an upper bound. More precisely, their approach assumes that it is possible to draw up a list of species liable to be present in R. Consequently, it cannot be implemented in little known regions. Furthermore, it requires to have some prior information on the probabilities of presence (in R) of species not detected by the quadrat sampling, which constitutes a limitation of this approach.

Royle *et al.* (2006) have developed a Bayesian approach which presents some similarities with the previous one, in that they introduce a supercommunity of species which is supposed to include the species population of interest (see also Dorazio, Royle and Link, 2007). As in Dupuis and Jochim (2006), the size M of this supercommunity is assumed to be known and has to satisfy the constraint $M \geq S$, while S is actually unknown. Not surprisingly the estimate of S depends on M, and the choice of M is thus important in this method (as in Dupuis and Jochim, 2006). As a rule, the approach of Royle *et al.* (2006) requires to have some idea on the size of S, to meet - as far as possible - the constraint $M \geq S$. However, it can be used when nothing is known on the size of S, by assigning large values to M; indeed, such values lead effectively to flat priors on S because Royle *et al.* (2006) assume that $S|M$ follows a uniform distribution on $\{1, \ldots, M\}$. Nevertheless, such a strategy is not very satisfactory in practice, because assigning large values to M induces high computational costs, as stressed by Royle *et al.* (2006) and by Dorazio, Royle and Link (2007). In other respects, we note that the approach of Royle *et al.* (2006) is able to estimate the number of species present in any finite subset of quadrats, but it

is unclear if the region R has to be composed of an infinite number of quadrats or not. However, we note that n plays no part in their modeling and that its value is actually not mentioned; this element and others (cf Section 3.3) lead us to believe that the authors implicitly assume that n is theoretically infinite (as in Dorazio and Royle, 2005). Lastly, we also note that Royle *et al.* (2006) assume that the probability of presence (in R) of any species s belonging to the supercommunity does not depend on s, which constitutes a strong biological assumption which is not always reasonable (cf Section 3.3.2); by contrast, Dupuis and Joachim (2006) do not make such an assumption. For all these reasons, we believe that the approach of Royle *et al.* (2006) is not suitable to deal with the situation considered in this paper.

The model developed in this paper is closer to that of Dorazio and Royle (2005) than the two other models (Royle *et al.*, 2006; Dupuis and Jochim, 2006). Contrary to Dorazio and Royle (2005), we do not assume that n is theoretically infinite, which allows us to introduce it in the model. When n is finite, it is indeed preferable that it is a part of the model. Ignoring n can actually lead to unreasonable estimates of S, simply because the resulting model will produce the same estimate of S whatever the sampling fraction T/n (eg Hass *et al.*, 2006). Our approach extends the one of Dorazio and Royle (2005) to the situation where no assumption is made regarding the size of n; we show in particular that the DR likelihood and ours coincide asymptotically (that is when $n \longrightarrow \infty$). Moreover, obtaining the Bayesian estimate of S involves serious computational difficulties (cf Section 3.7), and the DR approach is rather *ad hoc*, in that it is not fully Bayesian (in particular, no prior distribution is placed on S). In this paper, we show how to overcome these difficulties and provide a full Bayesian analysis of the problem.

We stress that existing approaches cannot incorporate satisfactorily some prior information on S, except (of course) for a prior consisting of an upper bound on S; for example,

it is not possible to incorporate information on S consisting of an estimate S^* of S and an interval $[a^*, b^*]$ containing S^*, such that $\Pr(S \in [a^*, b^*]) = 0.95$, while prior information is typically available under this form (when it exists). As far as the approaches of Dupuis & Joachim (2006), and Royle *et al.* (2006) are concerned, the key reason is that S is not a parameter of the model (see Sections 3.3.2 and 3.6). In Dorazio and Royle (2005), S is a parameter of the model, but no prior is placed on S (as already mentioned above). The approach developed in this paper has no such a limitation. We also pay particular attention to the situation where no prior information is available; so, different non informative priors on S are considered and the sensitivity to such priors is examined. In particular, we provide arguments for the standard non-informative prior of Jeffreys. To our knowledge, no theoretical justification has yet been provided for this choice of prior; see, for example, Kass and Wasserman (1996).

2. The experimental protocol and underlying processes

2.1 *The experimental protocol and data description*

The experimental protocol to collect the data is standard (Krebs, 1984; or Mingoti and Meeden, 1992; Dupuis and Joachim, 2006). The region R is divided into n spatial units, called quadrats for convenience, though they may have different shapes. In this paper we assume that these quadrats are of equal area. A sample of T quadrats is then taken, and the sampled quadrats are numbered from 1 to T. The draw is usually performed at random so as to have a sample representative of the whole region R. Finally, an experimenter visits each sampled quadrat $K \geq 2$ times, and records the species detected. Detections are typically based on visual or oral recognitions; we assume that species are correctly identified. The protocol used for collecting the data set analyzed in Dorazio and Royle (2005) enters in this framework. It is in fact described in details in Royle *et al.* (2006); see the Section *Protocol for Sampling Communities*. We note that the number n of spatial

units (quadrats) which compose the study region does not appear in this description; this is not actually surprising since this quantity is not a part of the model. In fact, in most papers which ignores n during the modelling, its value is not provided, as pointed out by Mingoti and Meeden (1992), and by Hass *et al.*, (2006).

Sometimes, all the n quadrats are explored, thus $T = n$; such a situation typically occurs when the size n is small or moderate. Interestingly, the methodology developed in this paper also applies to this particular situation.

We denote by $y_s = (y_{sj}; j = 1, \ldots, T)$ the record (or history) related to species s. When $K = 4$ and $T = 6$, a possible record is: $y_s = (3\ 0\ 0\ \ 0\ 4\ 0)$. Such a record means that species s has been detected in quadrat 1 during three visits, and detected in quadrat 5 during each visit. Moreover, it has not been detected in quadrats $2, 3, 4, 6$. We stress that when $y_{sj} = 0$, it does not necessarily mean that species s is absent from quadrat j. Indeed, it is possible that a species is present in a quadrat, but is not detected during the K visits. From a statistical point of view, the problem is to estimate the number S of species present in R, from the records of species whose the presence (in R) has been detected at least once.

The size of n is a key element of the paper. In practice, its range is particularly large, since its value strongly varies from one survey to another. Let us illustrate this variability through a few examples, by limiting ourselves to ornithological surveys. The size of n can be moderate as in Dupuis and Joachim (2006) where $n = 40$ quadrats, or relatively large as in Decamps *et al.* (1987) where $n = 98$ quadrats. In this study the values of n are relatively small 18 and 22 (or moderate 40). In Dorazio and Royle (2005), Royle *et al.* (2006), Dorazio, Royle and Link (2007), the value of n is is not provided.

2.2 *Underlying processes and missing data.*

For any species s, we introduce two underlying processes: the first one is related to its occurrences in the quadrats, and the other one is related to its detections.

- For $j = 1, \ldots, n$, we denote by z_{sj} the indicator of presence of species s in quadrat j; thus $z_{sj} = 1$ if species s is present in quadrat j, and zero otherwise.

- Let j be a sampled quadrat. For a species s present in quadrat j (thus such that $z_{sj} = 1$), $x_{sj} \in \{0, 1, \ldots, K\}$ denotes the number of times that species s has been detected in quadrat j during the K visits. If $z_{sj} = 0$ thus $x_{sj} = 0$ (with probability one).

Introducing these two latent processes allows us to formulate rigorously the biological assumptions made, and to introduce, in a natural way, the parameters of biological interest; such a strategy is standard when one process is partially observed; see eg Dupuis (1995) in a capture-recapture context. Dupuis, Bled, and Joachim (2010) have exhibited the missing data structure of quadrat sampling data; this structure will be particularly useful here to obtain the expression of the likelihood in closed form. It is recalled below. When species s has not been detected in a sampled quadrat j (that is when $y_{sj} = 0$), it is clear that z_{sj} is missing. In fact the event $(y_{sj} = 0)$ covers two exclusive situations: either species s is present in the sampled quadrat j but has not been detected, or it is not present in it (that is $z_{sj} = 0$). Formally, we have the equivalence $(y_{sj} = 0) \Longleftrightarrow (z_{sj} = 1$ and $x_{sj} = 0)$ or $(z_{sj} = 0)$. When $(y_{sj} = k)$, where $1 \leq k \leq K$, it is clear that z_{sj} is not missing $(z_{sj} = 1)$, and formally one has $(y_{sj} = k) \Longleftrightarrow (z_{sj} = 1$ and $x_{sj} = k)$. Moreover, z_{sj} is of course missing, when quadrat j is not a part of the sampled quadrats. Lastly, note that the whole vector $(z_{s1}, \ldots, z_{sn})$ is missing when species s has not been detected by the quadrat sampling, that is when $y_s = (0, \ldots, 0)$.

3. The homogeneous model M_0

3.1 *Some additional notation.*

Our notation is essentially the same as the one adopted in Dupuis and Joachim (2006). Throughout the paper, $p(.)$ denotes a probability mass function. We denote by $\mathbb{1}_{(C)}$ the indicator function that takes the value 1 when the condition C is true and zero otherwise. The null vector is denoted by $\vec{0}$. Let v be a vector; we denote by $|v|$ the sum of all its components. The number of species detected by the quadrat sampling is denoted by d. The vectors $(z_{sj}; j = 1, \ldots, T)$ and $(x_{sj}; j = 1, \ldots, T)$ are denoted by z_s and x_s, respectively. Lastly, we denote by $\mathbf{z}_s$ the vector $(z_{sj}; j = 1, \ldots, n)$. We stress that, in Dorazio and Royle (2005), J denotes the number of sampled quadrats, while this quantity is denoted by T in this paper (as in Dupuis and Joachim, 2006).

3.2 *Modeling detectability*

Biological assumptions related to detections are supported by the random vectors x_s's.

Assumption A1. We assume that $x_1, \ldots, x_s, \ldots, x_S$ are independent.

Assumption A2. We assume that: $p(x_s|z_s) = \prod_{j=1}^{T} p(x_{sj}|z_{sj})$

Assumption A3. We assume that $x_{sj}|z_{sj} = 1 \sim \text{Binomial}(K, q_s)$.

Assumptions A1, A2 and A3 are standard; they are also present in Dorazio & Royle (2005) and in Royle *et al.* (2006). A2 means that the probability of detecting species s in quadrat j does not depend on its (possible) detections in the other quadrats. A3 means that, s and j being fixed, the detections of species s during the K visits in quadrat j are independent. Moreover, q_s represents the probability of detecting species s in any quadrat j during any visit, given that it is present in quadrat j.

3.3 *Modeling occurrence of species in the quadrats.*

Assumption A4. We assume that $\mathbf{z}_1, \ldots, \mathbf{z}_s, \ldots, \mathbf{z}_S$ are independent.

A4 is standard and means that the species present in R do not interact relative to their

presence (in the quadrats). We thus exclude predator-prey relationships between species.

Concerning the probabilistic assumptions made on the z_{sj}'s (s being fixed), one can distinguish two types of approach according to whether one models occurrences of species liable to be present in R (unconditional approach), or whether modeling involves species present in R (conditional approach). This terminology has been introduced by Dupuis, Bled and Joachim (2010). This distinction between conditional and unconditional approaches is crucial when estimating species richness is of concern; it is the reason why we indicate how each approach deals with this stage of the modeling. The way of modeling occurrence is unconditional in Dorazio, Royle & Link (2007) and in Dupuis & Joachim (2006); it is conditional in Dorazio & Royle (2005), as well as in this paper. Moreover, it is useful to introduce the indicator ξ_s equal to 1 if species s is present in R and zero otherwise. We denote by λ_s the probability that species s is present in R; thus $\lambda_s = \Pr(\xi_s = 1)$.

3.3.1. Conditional approaches

• The way we model occurrences in this paper is the one adopted by Dupuis, Bled and Joachim (2010). We briefly recall this approach. The following assumption is made.

Assumption A5. Let s be a species present in R; then it exists at least one quadrat $j \in \{1, \ldots, n\}$ such that $z_{sj} = 1$ (or equivalently $\mathbf{z}_s \neq \vec{0}$). We assume that:

$$p(\mathbf{z}_s|\varphi_s) = \frac{\varphi_s^{|\mathbf{z}_s|}(1-\varphi_s)^{n-|\mathbf{z}_s|}}{1-(1-\varphi_s)^n} \tag{3.1}$$

where $|\mathbf{z}_s|$ represents the number of quadrats in which species s is present.

It is easy to check that φ_s represents the probability that species s is present in quadrat j, given that it is present in at least one other quadrat (Dupuis et *al.*, 2010). Parameter φ_s is connected to the rarety (in a spatial sense) of species s; a small occurrence probability φ_s being associated with a species s which occupies (on average) a small number of quadrats. Indeed, it is easy to check that the expectation of $|\mathbf{z}_s|$ is equal to $n\varphi_s[1-(1-\varphi_s)^n]^{-1}$

which is an increasing one-to-one function of φ_s (it varies from 1 to n). Moreover, due to the constraint $\mathbf{z}_s \neq \vec{0}$, the z_{sj}'s are clearly not independent. However, a certain form of conditional independence between the z_{sj}'s holds: namely z_{si} and z_{sk} are independent conditionally on the presence of species s in any other quadrat j, distinct from i and k (Dupuis et *al.*, 2010). The result below makes it possible to obtain the likelihood in closed form and also links the way we model occurrences in the sampled quadrats and the way Dorazio and Royle (2005) proceed.

Proposition 3.1. The probability mass function of z_s is:

$$p(z_s|\varphi_s) = [\varphi_s^{|z_s|}(1-\varphi_s)^{T-|z_s|}] \,/\, [1-(1-\varphi_s)^n] \quad \text{if} \quad z_s \neq \vec{0} \tag{3.2}$$

and $p(z_s|\varphi_s)) = [(1-\varphi_s)^T - (1-\varphi_s)^n] \,/\, [1-(1-\varphi_s)^n]$ otherwise.

Proof. See Appendix A1.

Note that the above Proposition also holds in the particular case $T = n$.

• Dorazio and Royle (2005) model occurrences of species s only in the sampled quadrats, by assuming that $z_{s1}, \ldots, z_{sT}$ are independent outcomes of a Bernoulli random variable. We note that the above constraint $\mathbf{z}_s \neq \vec{0}$ (which characterizes the fact that species s is present in R when n is finite) has no equivalent in the DR modeling. Dorazio and Royle (2005) assume that:

$$p(z_s|\psi_s) = \psi_s^{|z_s|}(1-\psi_s)^{T-|z_s|} \tag{3.3}$$

where $|z_s|$ represents the number of sampled quadrats in which species s is present, and $\psi_s = \text{pr}(z_{sj} = 1)$. Note that ψ_s and φ_s do not have the same meaning, hence two distinct notations; in fact, φ_s is a conditional occurrence parameter (contrary to ψ_s); as stressed in Dupuis et *al.*, (2010). Here, it is of interest to compare the probability mass function of z_s we use in this paper (cf Propostion 3.1) and the one adopted by Dorazio and Royle (2005). First, note that both T and n appear in (3.2), contrary to (3.3). More importantly, if in

(3.2) we make $n \longrightarrow \infty$ (T being fixed), we observe that both p.m.f. coincide (φ_s and ψ_s being confounded when $n \longrightarrow \infty$, as explained in Section 3.5).

3.3.2. Unconditional approaches

Unconditional approaches - which model occurrences (in the quadrats) of species liable to be present in R - are essentially characterized by three elements. First, a supercommunity, denoted by $\mathcal{M}$, is introduced; it includes the species community of interest and its size (denoted by M) is assumed to be known. Note that $S \leq M$; in other words, M is an upper bound of S. Second, the species richness S is not a parameter of the model, one has $S = \sum_{s=1}^{M} \xi_s$. Third, unconditional approaches give a positive probability to the event $\xi_s = 0$ (or equivalently to the event $\mathbf{z}_s = \vec{0}$ when n is finite), contrary to conditional approaches. We can distinguish two unconditional approaches according to whether n plays a part in the way occurrences are modeled (Dupuis and Joachim, 2006), or not (Royle et *et al.*, 2006); see below.

• Let s be a species belonging to the supercommunity $\mathcal{M}$, Dupuis and Joachim (2006) consider the hierarchical model below:

$$\xi_s|\lambda_s \sim \text{Bernouilli}(\lambda_s) \qquad p(\mathbf{z}_s|\xi_s = 1, \varphi_s) = \frac{\varphi_s^{|\mathbf{z}_s|}(1-\varphi_s)^{n-|\mathbf{z}_s|}}{1-(1-\varphi_s)^n} \tag{3.4}$$

where the meaning of φ_s is the one given above. The way a species s present in R occupies the n quadrats is thus modeled as in (3.1).

• Similarly, for any species $s \in \mathcal{M}$, Royle *et al.* (2006) introduce the indicator ξ_s and the parameter λ. Note that λ (denoted by Ω by these authors) is therefore assumed to be the same for all species, which constitutes a strong biological assumption which is not always reasonable. For example, in the paper of Dupuis and Joachim (2006), estimating S requires to estimate λ_s for each species s not detected; now, a strong variability among these different estimates can be observed, since they vary from 0.13 to 0.95. Consequently, for this species population (which is also the one considered in this paper), the assumption

of Royle *et al.* (2006) concerning λ would not be tenable from a biological point of view. Moreover, Royle *et al.* (2006) assume independence between the z_{sj}'s, conditionally on $\xi_s = 1$. This independence assumption (called afterwards A6) is more general than in Dorazio and Royle (2005), in that it applies to any finite subset of quadrats taken from the population of quadrats which composes R. Assumption A6 proves to be necessary when estimating the number of species present in such a subset is of interest (cf the introduction). Now, it is not clear if the model assumes that the region of study has to be composed of an infinite number of quadrats or not. However, different elements lead us to believe that their model presupposes that n is theoretically infinite (as Dorazio and Royle, 2005). First, no notation is introduced to designate the number of quadrats which compose R (consequently, this quantity is not a part of the model); second, and more importantly, assuming that n is finite will be not consistent with Assumption A6. Indeed, let us assume that n is finite and that the size of the subset (above mentioned) is equal to n; thus assumption A6 will imply that $z_{s1}, \ldots, z_{sn}$ are independent (conditionally on $\xi_s = 1$), which is not possible since $\xi_s = 1$ implies exactly that $z_{s1}, \ldots, z_{sn}$ are not independent, as stressed in Section 3.3.1 (first paragraph). Therefore the Royle *et al.* (2006) approach does not apply to the situation considered in our paper, where n is assumed to be finite and not necessarily large.

3.4 *Modeling species heterogeneity*

As in Dorazio and Royle (2005) we model heterogeneity between species via random effects, as follows:

$$\text{logit}(q_{sj}) = \alpha + \mu_s \quad \text{and} \quad \text{logit}(\varphi_{sj}) = \beta + \nu_s \tag{3.6}$$

where the μ_s's are $i.i.d$ according to a normal distribution $\mathcal{N}(0, \sigma_\mu^2)$, and the ν_s's are $i.i.d$ according to a $\mathcal{N}(0, \sigma_\nu^2)$. Contrary to the DR paper, and for brevity, we do not introduce a prior correlation between μ_s and ν_s.

3.5 *Likelihood.*

It is convenient to number histories from 0 to $H = (K+1)^T - 1$ (note that one has $(K+1)^T$ distinct histories) where the history 0 is associated with the record $\vec{0}$; only histories distinct from 0 are observable. Moreover, we denote by N_h the number of species having the history h, and by d the number of detected species; therefore, we have $d = \sum_{h=1}^{H} N_h$. Note that the count $N_0 = S - d$ which represents the number of undetected species is not a part of the data. Due to Assumptions A1 and A4, it is clear that

$$(N_1, \ldots, N_H)|S, \alpha, \beta, \sigma_\mu^2, \sigma_\nu^2 \sim \text{Multinomial}(S; \omega_1, \ldots, \omega_H)$$

where $\omega_h = \Pr(y_s = h|\alpha, \beta, \sigma_\mu^2, \sigma_\nu^2)$. Note that ω_h cannot be analytically calculated; indeed:

$$\omega_h = \int\int \Pr(y_s = h|\alpha, \beta, \mu_s, \nu_s)\pi(\mu_s|\sigma_\mu^2)\pi(\nu_s|\sigma_\nu^2) d\,\mu_s d\,\nu_s\,;$$

and this integral is untractable; see indeed the expression of $\Pr(y_s = h|\alpha, \beta, \mu_s, \nu_s)$ below. Nevertheless this integral can be easily evaluated via classical Monte Carlo simulation methods. The expression of $\Pr(y_s = h|\alpha, \beta, \mu_s, \nu_s)$, which is also equal to $\Pr(y_s = h|\varphi_s, q_s)$, is given by Proposition 3.2.

Proposition 3.2. Let s be a species detected by the quadrat sampling, one has:

$$p(y_s|\varphi_s, q_s) = \frac{\rho_s\, q_s^{W_s}(1-q_s)^{KV_s - W_s}\, \varphi_s^{V_s}\left[(1-q_s)^K\varphi_s + 1 - \varphi_s\right]^{T-V_s}}{1-(1-\varphi_s)^n} \tag{3.7}$$

where V_s denotes the number of quadrats in which species s is detected, W_s the total number of visits during which it is detected, and

$$\rho_s = \prod_{j=1}^{T} \left(\begin{array}{c} K \\ y_{sj} \end{array} \right).$$

Moreover,

$$\Pr(y_s = \vec{0}|\varphi_s, q_s) = \frac{\left[(1-q_s)^K\varphi_s + 1 - \varphi_s\right]^T - (1-\varphi_s)^n}{1-(1-\varphi_s)^n}. \tag{3.8}$$

Proof. See Appendix A2.

The expressions of $p(y_s)$ which appear in Proposition 3.2 have been already mentioned in Dupuis, Bled and Joachim (2010), but the proof has never been published.

Considering that $(N_1, \ldots, N_H)|S, \alpha, \beta, \sigma_\mu^2, \sigma_\nu^2 \sim \text{Multinomial}(S; \omega_1, \ldots, \omega_H)$, the likelihood of the data $\mathbf{y} = \{N_1, \ldots, N_H\}$ is

$$L(\boldsymbol{\theta}_0; \mathbf{y}) = \frac{S!}{(S-d)! \prod_{h=1}^{H} N_h!} \, \omega_0^{S-d} \prod_{h=1}^{H} \omega_h^{N_h}$$

where $\boldsymbol{\theta}_0 = (S, \alpha, \beta, \sigma_\mu^2, \sigma_\nu^2)$ denotes the parameter of the model M_0. It is easy to see that $\{N_{vw}; v = 1, \ldots, n; w = v, \ldots, Kv\}$, where N_{vw} represents the number of detected species s such that $V_s = v$ and $W_s = w$, constitutes a sufficient statistics for θ_0.

Due to Proposition 3.1, it is expected that inference on S based on the DR likelihood should coincide with ours asymptotically (that is when $n \longrightarrow \infty$). It is effectively the case. Under the DR approach, one has

$$p(y_s|\psi_s, q_s) = \rho_s \, q_s^{W_s} (1-q_s)^{KV_s - W_s} \, \psi_s^{V_s} \left[(1-q_s)^K \psi_s + 1 - \psi_s\right]^{T-V_s} \tag{3.9}$$

if $y_s \neq \vec{0}$, and

$$\Pr(y_s = \vec{0}|\psi_s, q_s) = \left[(1-q_s)^K \psi_s + 1 - \psi_s\right]^T. \tag{3.10}$$

If we now make $n \longrightarrow +\infty$ in (3.7) and (3.8), T and K and φ_s being fixed, we observe that the likelihood of the DR model and ours coincide. Indeed, $(1-\varphi_s)^n \longrightarrow 0$ when n tends to $+\infty$, and the two parameters φ_s and ψ_s are confounded asymptotically. To check this second point, it is convenient to introduce the parameter ζ_s equal to the probability that $z_{sj} = 1$ (under the model M_0); one has $\zeta_s = \varphi_s/[1-(1-\varphi_s)^n]$ (see Dupuis *et al.*, 2010). We note that the parameters ψ_s and ζ_s do not take their values in the same set; indeed, $\psi_s \in]0, 1[$ while $\zeta_s \in]1/n, 1[$; the verification is immediate. But, when $n \longrightarrow +\infty$, the three parameters φ_s, ζ_s and ψ_s are confounded, since $\zeta_s = \varphi_s/[1-(1-\varphi_s)^n] \longrightarrow \varphi_s$ and ζ_s and ψ_s now take their values in the same set.

3.6 *Prior distributions*

We assume that the parameters S, α, β, σ_μ^2 and σ_ν^2 are a priori independent. In other words, we assume that: $\pi(\theta_0) = \pi(\alpha)\pi(\beta)\pi(\sigma_\mu^2)\pi(\sigma_\nu^2)\pi(S)$.

• As Dorazio and Royle (2005), we assume that:

$$\pi(\alpha) = \frac{\exp(\alpha)}{(1+\exp(\alpha))^2} \quad \text{and} \quad \pi(\beta) = \frac{\exp(\beta)}{(1+\exp(\beta))^2}\,,$$

so that $\text{logit}^{-1}(\alpha)$ and $\text{logit}^{-1}(\beta)$ follow a uniform distribution.

• As far as σ_μ^2 and σ_ν^2 are concerned, we adopt the following flat priors:

$$\sigma_\mu^2 \sim \Gamma^{-1}(2+\epsilon, 1+\epsilon) \quad \text{and} \quad \sigma_\nu^2 \sim \Gamma^{-1}(2+\epsilon, 1+\epsilon)$$

with small ϵ, arguing that $\text{Var}(\sigma_\mu^2) = 1/\epsilon$ is thus large, while $E(\sigma_\mu^2)$ is constant (equal to 1); and the same for σ_ν^2. In the above Γ^{-1} distribution, the used parametrisation is such that if $Z \sim \Gamma^{-1}(a,b)$, where $a > 2$ and $b > 0$, then the p.d.f. f of Z is $f(z|a,b) \propto z^{-(a+1)}\exp(-b/z)$, $E(Z) = b/(a-1)$ and $\text{Var}(Z) = [E(Z))^2]/(a-2)$. The priors $\Gamma^{-1}(2.1, 1.1)$ and $\Gamma^{-1}(4,3)$ used by King and Brooks (2008) enter in this general framework.

• A negative binomial distribution is usually placed on an integer parameter (eg King and Brooks, 2001). When no prior information is available, the improper Jeffreys prior distribution, is usually proposed: namely $\pi(S) \propto \frac{1}{S}$ in our context. Now, to our knowledge, no theoretical motivation exists in the literature concerning the Jeffreys prior, as already pointed out by Kass and Wasserman (1996); these authors simply note that its extends the standard non informative prior for a real parameter $\beta > 0$ (namely, $\pi(\beta) \propto 1/\beta$) to the case of an integer parameter. Our motivation for the use of the Jeffreys prior as a non informative prior is provided by the following proposition which establishes a natural link between the Negative Binomial distribution and the Jeffreys prior.

Proposition 3.3. The Jeffreys prior coincides with the limiting case of a negative binomial distribution in which the prior variance tends to ∞ (the prior mean being fixed).

Proof. See Appendix A3.

The main alternative to the Jeffreys prior is $\pi(S) = 1$ which is also improper (eg Casteldine, 1981). In the next Section, we indicate in what extent choosing this prior, instead of the Jeffreys prior, could affect the Bayesian estimate of S. Note that in the absence of prior information on S, we can also use a Negative Binomial distribution NegBin(r, p) such that p is small and $r = p$; arguing that Var(S) is thus large, while $E(S)$ is close to 1; Recall that $E(S) = (1-p)\frac{r}{p}$ and Var$(S) = E(S)/p$. Moreover, it is clear that the Bayesian estimate of S based on such a flat proper prior (namely a NegBin(p, p) with small p) will be close to the one yielded by the Jeffreys prior (due to the Proposition 3.3); see also the way S is generated during the MCMC algorithm (cf Section 3.7).

So far, we have focused on non-informative priors on S, but an advantage of our approach, compared to the unconditional approaches of Dupuis and Joachim (2006) and of Royle *et al.* (2006), is that it easily allows prior information on S to be incorporated (when it exists), simply because S is a parameter of the model. For example, it is possible to incorporate - via a negative binomial distribution - prior information consisting of an estimate S^* of S and of its standard error e (on condition that $e^2 > S^*$), as follows: set $E(S) = S^*$ and Var$(S) = e^2$, then use the fact the negative binomial distribution can be parametrized by its mean and variance (cf Appendix A3). More interestingly, it is also possible (in general) to incorporate - via a negative binomial distribution - a prior consisting of $S^* = E(S)$ and an interval $[a^*, b^*]$ (containing S^*) such that $\Pr(S \in [a^*, b^*]) = 0.95$; obtaining the coefficients of the corresponding negative binomial distribution requires a program (available from the authors on request) which uses the above parametrization and proceeds by dichotomy (details are omitted). This is of course an interesting characteristic

of the negative binomial distribution since, in practice, prior information will typically be available under this form. (Note that analogous observations have been made by Dupuis (1995) while incorporating some similar prior information on a parameter belonging to $[0, 1]$.) When an unconditional approach is used, incorporating such an informative prior (consisting of S^* and $[a^*, b^*]$) is not possible, simply because S is not a parameter of the model. So, in Royle *et al.* (2006), the distribution of $S|M$ is in fact completely determined by the prior distribution placed on λ; now, the latter is practically constrained (owing to biological considerations). More precisely, $S|M$ follows a uniform distribution on distribution on $\{1, \ldots, M\}$ because a uniform distribution on $[0, 1]$ has been placed on λ (other distributions on λ will be difficult to justify, in practice).

3.7 *Estimating S and computational issues.*

Obtaining the posterior distribution of S (and in particular the posterior mean) requires the implementation of MCMC methods. A possible algorithm is as follows. The parameters $(\alpha, \beta, \sigma_\mu^2, \sigma_\nu^2)$ are updated via Metropolis-Hastings steps. For example, updating α proceeds as follows. We propose a new value $\alpha' \sim \text{Normal}(\alpha, \delta)$ where δ is fixed via tuning pilot. The proposal α' is accepted with probability $\min(1, r)$ where

$$r = \frac{\pi(\boldsymbol{\theta}_0'|\mathbf{y})}{\pi(\boldsymbol{\theta}_0|\mathbf{y})} = \frac{L(\boldsymbol{\theta}_0'; \mathbf{y})}{L(\boldsymbol{\theta}_0; \mathbf{y})} \times \frac{\pi(\alpha')}{\pi(\alpha)},$$

where $\boldsymbol{\theta}_0 = (S, \alpha, \beta, \sigma_\mu^2, \sigma_\nu^2)$ represents the current state and $\boldsymbol{\theta}_0' = (S, \alpha', \beta, \sigma_\mu^2, \sigma_\nu^2)$ the proposal. The parameter S is updated via a Gibbs step; it is easy to check that if we adopt the Jeffreys prior, then

$$S - d|\alpha, \beta, \sigma_\mu^2, \sigma_\nu^2, \mathbf{y} \sim \text{Negative Binomial}(d, 1 - \omega_0) \tag{3.11}$$

where ω_0=Pr$(y_s = \vec{0}|\alpha, \beta, \sigma_\mu^2, \sigma_\nu^2)$. If $\pi(S) = 1$ thus $S - d|\alpha, \beta, \sigma_\mu^2, \sigma_\nu^2, \mathbf{y}$ follows a Negative Binomial $(d + 1, 1 - \omega_0)$. Therefore, these two non informative priors should give close estimates of S, as long as 1 is small compared with d. When $S \sim$ NegBin (r, p), it is

easy to check that $S - d|\alpha, \beta, \sigma_\mu^2, \sigma_\nu^2, \mathbf{y}$ follows a Negative Binomial $(r + d, 1 - (1 - p)\omega_0)$. Consequently, if one uses a Negative Binomial distribution (r, p), with $r = p$ and small p, as a non informative prior distribution on S (as suggested in Section 3.6), it is clear that the estimate of S based on this proper prior distribution, and the one based on the Jeffreys prior, will be very close.

Implementing this MCMC algorithm requires that ω_0 and the ω_h's such that $N_h \neq 0$ be computed at each iteration; that is integrals which are not analytically tractable (see Section 3.5). Although they can be easily approximated by classical Monte Carlo methods, obtaining the Bayesian estimate of S by this strategy is computationally intensive; simply because the number of ω_h we have to compute (namely those such that $N_h \neq 0$) rapidly increases with T and K (recall that one has $(K + 1)^T$ possible histories). This strategy is actually no more conceivable when one assumes that occurrence and detectability parameters are quadrat dependent (cf Section 4). Hence it is necessary to propose an alternative to this MCMC algorithm, to overcome these computational difficulties (already mentioned by Dorazio and Royle, 2005; as well as by Royle *et al.*, 2006).

Our idea is to implement a MCMC algorithm in which only one integral (namely ω_0) will have to be calculated at each MCMC step. We achieve this objective by treating (in the MCMC algorithm) the random effects μ_s and ν_s of detected species at the same level as the parameters $S, \alpha, \beta, \sigma_\mu^2, \sigma_\nu^2$. In other words the MCMC algorithm is implemented on $(\mu, \nu, S, \alpha, \beta, \sigma_\mu^2, \sigma_\nu^2)$ where $\mu = \{\mu_s; s = 1, \ldots, d\}$ and $\nu = \{\nu_s; s = 1, \ldots, d\}$, instead of $(S, \alpha, \beta, \sigma_\mu^2, \sigma_\nu^2)$. This algorithm relies on the joint distribution of $(\mathbf{y}, d, \mu, \nu, \theta_0)$ where $\mathbf{y} = \{y_s; s = 1, \ldots, d\}$. This density decomposes as follows:

$$p(\mathbf{y}|\mu, \nu, d, \alpha, \beta)\ p(\mu, \nu|d, \sigma_\mu^2, \sigma_\nu^2)\ p(d|S, \alpha, \beta, \sigma_\mu^2, \sigma_\nu^2)\ \pi(\theta_0)$$

where

$$p(\mathbf{y}|\mu,\nu,d,\alpha,\beta)=\prod_{s=1}^{d}p(y_s|y_s\neq\vec{0},\mu_s,\nu_s,\alpha,\beta)=\prod_{s=1}^{d}p(y_s|y_s\neq\vec{0},\varphi_s,q_s),$$

$$p(\mu,\nu|d,\sigma_\mu^2,\sigma_\nu^2)=\prod_{s=1}^{d}p(\mu_s|\sigma_\mu^2)p(\nu_s|\sigma_\nu^2)\quad\text{and}\quad p(d|S,\alpha,\beta,\sigma_\mu^2,\sigma_\nu^2)=\left(\begin{array}{c}S\\d\end{array}\right)\omega_0^{S-d}(1-\omega_0)^d.$$

The expression of $p(y_s|y_s\neq\vec{0},\varphi_s,q_s)$ (where s represents a detected species) is derived from Proposition 3.2; so, one has:

$$p(y_s|y_s\neq\vec{0},\varphi_s,q_s)=\frac{\rho_s\,q_s^{W_s}(1-q_s)^{KV_s-W_s}\,\varphi_s^{V_s}\left[(1-q_s)^K\varphi_s+1-\varphi_s\right]^{T-V_s}}{1-\left[(1-q_s)^K\varphi_s+1-\varphi_s\right]^T}.$$

Updating $\alpha,\beta,\sigma_\mu^2,\sigma_\nu^2$, as well as μ and ν, is done via Metropolis-Hastings steps (details are omitted). Updating S is done via a Gibbs step as in (3.11).

4. The heterogeneous model $\mathbf{M}_1$

In this Section we no longer assume that the region under investigation, R, is spatially homogeneous, but assume that it is composed of $A\geq 2$ homogeneous subregions of biological interest; for example a forest can be divided into the edge and the inner forest (as in our illustration). The experimental protocol is without change, except that the region of study is first divided into A subregions, called $R_1,\ldots,R_a,\ldots,R_A$; then a quadrat sampling is performed in each subregion R_a (as indicated in Section 2). The number of quadrats in which R_a is divided is denoted by n_a, and the number of quadrats sampled in R_a is denoted by T_a.

• The assumptions A1, A2 and A4 of the model M_0 are not modified. The detectability parameters are now indexed by s and a; more precisely we denote by q_{sa} the probability to detect (during a visit) species s in any quadrat j located in the subregion R_a. The distribution of species s in the n quadrats is characterized by the probability mass function:

$$p(\mathbf{z}_s)=\frac{\prod_{a=1}^{A}\varphi_{sa}^{|z_{sa}|}(1-\varphi_{sa})^{n_a-|z_{sa}|}}{1-\prod_{a=1}^{A}(1-\varphi_{sa})^{n_a}}\qquad(4.1)$$

where $|z_{sa}|$ represents the number of quadrats of R_a in which species s is present. It is easy to check that φ_{sa} represents the probability that species s is present in any quadrat j located in region R_a, given that it is present in at least one other quadrat. The p.m.f. (4.1) is in fact a natural extension of the one adopted for the model M_0, as well as a particular case of the more general p.m.f.:

$$p(\mathbf{z}_s) = \frac{\prod_{j=1}^{n} \varphi_{sj}^{z_{sj}} (1-\varphi_{sj})^{1-z_{sj}}}{1 - \prod_{j=1}^{n}(1-\varphi_{sj})}$$

where a distinct occurrence parameter φ_{sj} is introduced for each quadrat j.

• The species effects are modeled as random effects (as in M_0), and the region effects as fixed effects, as follows:

$$\text{logit}(q_{sa}) = \alpha_a + \mu_s(a) \quad \text{and} \quad \text{logit}(\varphi_{sa}) = \beta_a + \nu_s(a)\,, \tag{4.2}$$

where the $\mu_s(a)$ are $i.i.d$ according to a normal distribution $\mathcal{N}(0, \sigma_\mu^2(a))$, and the $\nu_s(a)$ are $i.i.d$ according to a $\mathcal{N}(0, \sigma_\nu^2(a))$.

• The priors are similar to those used for model M_0.

• The likelihood under model M_1, is derived from the Proposition 4.1 below. Let s be a detected species; we denote by V_{sa} the number of quadrats of R_a in which species s has been detected, and by W_{sa} the total number of visits (made in region R_a) during which species s has been detected. Moreover we denote by $\boldsymbol{q}_s$ the vector $(q_{sa}; a = 1, \ldots, A)$ and by $\boldsymbol{\varphi}_s$ the vector $(\varphi_{sa}; a = 1, \ldots, A)$.

Proposition 4.1. Let s be any detected species. We have:

$$p(y_s|\boldsymbol{\varphi}_s, \boldsymbol{q}_s) = \rho_s \frac{\prod_{a=1}^{A} \varphi_{sa}^{V_{sa}} q_{sa}^{W_{sa}} \big[1-q_{sa}\big]^{U_{sa}} \big[(1-q_{sa})^K \varphi_{sa} + 1 - \varphi_{sa}\big]^{T_a - V_{sa}}}{1 - \prod_{a=1}^{A}(1-\varphi_{sa})^{n_a}}$$

where $U_{sa} = KV_{sa} - W_{sa}$ and ρ_s is defined as for the model M_0, except that now $T = \sum_a T_a$. Moreover, we have:

$$\Pr(y_s = \vec{0}|\boldsymbol{\varphi}_s, \boldsymbol{q}_s) = \frac{\prod_{a=1}^{A} \big((1-q_{sa})^K \varphi_{sa} + 1 - q_{sa}\big)^{T_a} - \prod_{a=1}^{A} \big(1-\varphi_{sa}\big)^{n_a}}{1 - \prod_{a=1}^{A}(1-\varphi_{sa})^{n_a}}.$$

The proof is very similar to the one established in the homogeneous case; consequently, details are omitted.

- Estimating S relies on the MCMC algorithm described in Section 3.7 and which requires only one integral to be computed at each iteration.

5. An illustration

5.1 *The Montech forest*

An illustration is given by data collected in May 1987 in order to estimate the number of bird species present in the forest of Montech (France), except for the birds of prey (recall that we exclude predator-prey relationships between species, cf Assumption A4). This forest, with a surface area of 1000 hectares, is relatively spatially homogeneous, mainly composed of oaks and hornbeams (Decamps *et al.*, 1987; Dupuis and Joachim, 2006). However, it is expected that the detectability and occurrence parameters (that is q_{sj} and φ_{sj}) could depend on quadrat j, according to whether j is located at the edge or in the inner forest (species s being fixed). It is why we have first estimated the species richness of the Montech forest (taken as a whole) under model M_0, then under model M_1. Moreover, we have also estimated separately the number of species present at the edge, and the number of species present in the inner forest (by using each time the model M_0).

5.2 *The experimental protocol*

The protocol and field description are detailed in Decamps *et al.* (1987); consequently only the main details are given here. The Montech forest has been divided into $n = 40$ quadrats of equal size; only half of the quadrats have been sampled (thus $T = 20$). The inner forest (called R_1) includes $n_1 = 22$ quadrats; the edge (called R_2) includes $n_2 = 18$; moreover, $T_1 = 14$ quadrats have been sampled in R_1, and $T_2 = 6$ in R_2. We consider that species s occupies quadrat j (that is $z_{sj} = 1$), if at least one pair belonging to species s has nested in quadrat j, during May. Information about the presence of nesting

species was provided by acoustic recognition of singing males according to the following procedure. In spring, songs of males (only) indicate the presence of a nesting pair close by. The researcher spent a prescribed time (twenty minutes in our study) at each station (in the center of quadrat), listening for birds. More precisely, data have been collected according to the following point count protocol: each 20-minute session has been sliced into four subsessions of 5 minutes each, during which the experimenter records whether the presence of the species of interest has been detected or not. Each slice is the equivalent of a visit, therefore $K = 4$. As information 29 species have been detected in the inner forest, 23 in the edge, and 33 in the whole forest. The complete data set is available from the first author on request.

5.3 *Results*

Under model M_0, the non-informative (Jeffreys) Bayesian estimate of the number of species present in the whole forest of Montech is 38.8 (posterior mean); $[34, 47]$ being a 95% posterior credible interval. Under model M_1, the posterior mean is 35.1, and $[33, 39]$ is a 95% posterior credible interval. These results show that taking into the fact that detectability and occurence parameters could differ between the inner forest and the edge has a significant impact on the estimation of S. Dupuis and Joachim (2006) have also estimated the species richness of the Montech forest in 1987, assuming spatial homogeneity of the region of study. But authors limited themselves to passerines species, estimation of S was based on a sub-sample of size $T = 8$, and additional data have been used to built the priors on the parameters λ_s of species s not detected by the quadrat sampling; consequently, comparing estimates of S is not relevant.

Table 1 below provides the posterior means and 95% credible intervals of S, α, β, σ_μ and σ_ν in each part of the Montech forest: the edge and the inner forest.

[Table 1 about here.]

These results are based on the MCMC algorithm described in Section 6.3 (with the Jeffreys prior used for S). Recall that, at each step of the MCMC algorithm, ω_0 (defined in Section 6.1) is calculated by a classical Monte Carlo method. A preliminary simulation study has shown that a run of 500 iterations was sufficient to satisfactorily approximate ω_0. Results appearing in Table 1 are based on a long MCMC sequence which was run for 10^6 iterations (with the first 10% discarded as burn-in). Independent replications (run from overdispersed starting points) produced essentially identical results (as those appearing in Table 1), so that convergence was assumed.

Let us comment on these results.

• An interesting element which emerges from these results is that the estimate of the species richness is significantly greater in the inner forest than at the edge.

• The species richness S constitutes the parameter of main interest in this paper, but the model also allows us to investigate other issues of biological interest, via the parameters α, β, σ_μ and σ_μ. Indeed, parameters β and σ_ν provide some global information on the rarety (in a spatial sense) of the species present in the study region (cf Section 3.3.1), while α and σ_μ provide some global information on their detectability. We observe, on a logit scale, that detectability is (on average) smaller at the edge (-0.9) than in the inner forest (-0.5); conversely, we note that species are more rare (always on a logit scale and on average) in the inner forest (-0.4) than at the edge (0.7). Moreover, we note that, for the inner forest, $\widehat{\alpha}$ (-0.5) and $\widehat{\beta}$ (-0.4) are very close, whereas, for the edge, it is $\widehat{\sigma}_\mu$ (1.2) and $\widehat{\sigma}_\nu$ (1.1) which are practically equal (for convenience, the estimate of α is denoted by $\widehat{\alpha}$; and a similar notation is adopted for β, σ_μ and σ_μ). These two observations lead us to the following comments. As far as the edge is concerned, the fact that $\widehat{\alpha}$ (-0.9) is significantly smaller than $\widehat{\beta}$ (0.7) suggests that species missed by the quadrat sampling in this part of the forest are mainly due to the presence of species not easily detectable. As far the inner

forest is concerned, the fact that $\widehat{\sigma}_\nu$ (2.1) is significantly greater than $\widehat{\sigma}_\mu$ (0.8) suggests, by contrast, that species missed by the quadrat sampling in this other part of the forest are mainly due to the presence of species spatially rare.

• Lastly, it is of interest to compare our estimates of α, β, σ_μ and σ_μ with those obtained by Dorazio and Royle (2005). For the BBS data set; $\widehat{\alpha} = -1.5$, $\widehat{\beta} = -1.9$, $\widehat{\sigma}_\mu = 1.1$, $\widehat{\sigma}_\nu = 2.2$. They are thus of the same order of magnitude; however, we observe non negligible distances between their estimates of α and β, and ours; that suggests that species involved with the BBS data set were, on average, not only more difficult to detect, but also rarer (in a spatial sense) than in our illustration. This could explain why, in Dorazio and Royle (2005), the ratio between the number of detected species and the estimate of S, namely $75/93.3 \approx 80\%$, is smaller than ours: indeed, for the whole forest, this ratio is equal to 85% under model M_0, and to 94% under model M_1; moreover, it is equal to 89% for the inner forest ; and to 86% for the edge.

6. Conclusion

This paper proposes a new approach for estimating the species richness from quadrat sampling when the study region is composed of a finite number n of quadrats. The fact that nothing is assumed regarding the size of n constitutes a key element of our approach which differentiates it from the ones developed by Dorazio & Royle (2005) and by Royle *et al.* (2006) which suppose that the study region is theoretically composed of an infinite number of quadrats. As a consequence, n is not a part of their models, contrary to ours.

Compared with the conditional approach of Dorazio & Royle (2005), our approach can be viewed as an extension, in that the DR likelihood coincides with ours asymptotically (that is when $n \longrightarrow \infty$). In other respects, recall that the estimate of S yielded by Dorazio & Royle (2005) is not fully Bayesian, but rather *ad hoc*. In this paper, we implement an efficient MCMC algorithm which allows us to obtain the Bayesian estimate of S (in spite of

serious computational difficulties). Interestingly, our algorithm could be easily modified to perform a fully bayesian analysis of the problem under the Dorazio and Royle assumption (namely $n = \infty$).

Compared with the unconditional approaches of Dorazio & Royle (2006) and of Dupuis & Joachim (2006), our approach can be used in the absence of information on S (thus in totally unknown regions). By contrast, implementing unconditional approaches in such a context is somewhat delicate as mentioned in the introduction. On the other hand, our approach allows us to incorporate easily (via a negative binomial distribution) some information on S, when its exists; for example, when it consists of an estimate S^* of S and of its standard error. By contrast, unconditional approaches do not offer such a possibility, because S is not a parameter of the model.

Lastly, it is of interest to mention the unconditional approach of MacKenzie *et al.* (2006), which assumes that the region is divided in n quadrats (no assumption is made on the size of n, as in our paper) and models occurrences of species s in the n quadrats as follows:

$$p(\mathbf{z}_s) = \psi_s^{|\mathbf{z}_s|}(1-\psi_s)^{n-|\mathbf{z}_s|} \tag{5.1}$$

To our knowledge, this model has never be used when estimating species richness is of interest, and we believe that it could constitute an interesting alternative to the conditional approach proposed in this paper (cf 3.1). As in Dupuis & Joachim (2006) and Royle *et al.* (2006), this model will require that a supercommunity of known size is introduced. But, contrary to Royle *et al.* (2006) which assume that the probability of presence (in R) of any species s belonging to this supercommunity does not depend on s, MacKenzie *et al.* (2006) do not make such a biological assumption, since $\lambda_s = 1-(1-\psi_s)^n$.

The model developed in this paper has to be considered as starting points for more elaborate models. We assume that the presence of a given species in a given quadrat is not

affected by whether or not the species is present in the neighbouring quadrats, but such an assumption is not always reasonable in animal populations, and a possible extension would thus be to introduce some spatial correlation between z_{sj} and its neighbours (via an autologistic model, for example). We also assume that species occupy the quadrats independently, a challenging problem would be to develop a model for estimating the species richness of an animal population within which predator-prey relationships exist.

Appendices

For convenience, the conditionings on the parameter(s) are omitted.

Appendix A1

We denote by E the set $\{0,1\}^n \backslash \{\vec{0}\}$. We have $p(z_s) = \sum_{z'_s} p(z_s, z'_s)$ where $(z_s, z'_s) \in E$. This sum is over all the possible values of z'_s when $z_s \neq \vec{0}$, and over all the possible values of z'_s, apart from $\vec{0}$, when $z_s = \vec{0}$ (to meet the constraint $\mathbf{z}_s = (z_s, z'_s) \neq \vec{0}$). We first consider the case $z_s \neq \vec{0}$. We have:

$$p(z_s, z'_s) = \frac{\varphi_s^{|z_s|+|z'_s|}(1-\varphi_s)^{n-|z_s|-|z'_s|}}{1-(1-\varphi_s)^n},$$

from which we deduce that

$$p(z_s) = \frac{\varphi_s^{|z_s|}(1-\varphi_s)^{T-|z_s|}}{1-(1-\varphi_s)^n} \sum_{z'_s} \varphi_s^{|z'_s|}(1-\varphi_s)^{n-T-|z'_s|}.$$

Hence:

$$p(z_s) = \frac{\varphi_s^{|z_s|}(1-\varphi_s)^{T-|z_s|}}{1-(1-\varphi_s)^n}$$

by observing that $\sum_{z'_s} \varphi_s^{|z'_s|}(1-\varphi_s)^{n-T-|z'_s|} = 1$. When $z_s = \vec{0}$, we have:

$$p(z_s) = \frac{(1-\varphi_s)^T}{1-(1-\varphi_s)^n} \sum_{z'_s \neq \vec{0}} \varphi_s^{|z'_s|}(1-\varphi_s)^{n-T-|z'_s|}.$$

We deduce the result by observing that

$$\sum_{z'_s \neq \vec{0}} \varphi_s^{|z'_s|}(1-\varphi_s)^{n-T-|z'_s|} = 1-(1-\varphi_s)^{n-T}.$$

Appendix A2

We begin by calculating $p(y_s)$ in function of q_s and φ_s, when $y_s \neq \vec{0}$. We set:

$$\rho_s = \prod_{j=1}^{T} \begin{pmatrix} K \\ y_{sj} \end{pmatrix}.$$

Given y_s, we partition the vector z_s in $z_s^{obs} = \{z_{sj} | y_{sj} \neq 0\}$ and $z_s^{mis} = \{z_{sj} | y_{sj} = 0\}$. Note that z_s^{obs} is known as soon as y_s is available, and that $y_s \neq \vec{0}$ implies that z_s^{obs} is not empty (and therefore z_s^{mis} and $\mathbf{z}_s$ are distinct vectors). We have:

$$p(y_s) = \sum_{z_s^{mis}} p(x_s, z_s^{obs}, z_s^{mis}) = \sum_{z_s^{mis}} p(x_s | z_s^{obs}, z_s^{mis}) p(z_s^{obs}, z_s^{mis}).$$

Due to Assumption A2, we have:

$$p(x_s | z_s^{obs}, z_s^{mis}) = \rho_s \, q_s^{W_s} (1 - q_s)^{KV_s - W_s} \, (1 - q_s)^{K|z_s^{mis}|}$$

where V_s and W_s are defined in Proposition 3.2. Due to Proposition 3.1 we have:

$$p(z_s^{obs}, z_s^{mis}) = \frac{\varphi_s^{V_s} \, \varphi_s^{|z_s^{mis}|} (1 - \varphi_s)^{T - V_s - |z_s^{mis}|}}{1 - (1 - \varphi_s)^n},$$

from which we deduce that:

$$p(y_s) = \rho_s \, \frac{\varphi_s^{V_s} q_s^{W_s} (1 - q_s)^{KV_s - W_s}}{1 - (1 - \varphi_s)^n} \sum_{z_s^{mis}} [\varphi_s (1 - q_s)^K]^{\, |z_s^{mis}|} (1 - \varphi_s)^{T - V_s - |z_s^{mis}|}.$$

Now, it is easy to check that

$$\sum_{z_s^{mis}} [\varphi_s (1 - q_s)^K]^{|z_s^{mis}|} (1 - \varphi_s)^{T - V_s - |z_s^{mis}|} = [\varphi_s (1 - q_s)^K + (1 - \varphi_s)]^{T - V_s};$$

hence the result.

We now calculate $p(y_s)$ for a not detected species s, that is such that $y_s = \vec{0}$. Note that now $z_s^{obs} = \emptyset$ and that $z_s^{mis} = z_s$. We start from:

$$p(y_s) = \sum_{z_s} p(x_s | z_s) p(z_s).$$

Due to Assumption A2, we have: $p(x_s | z_s) = (1 - q_s)^{K|z_s|}$. Two cases have to be distinguished: $T < n$ and $T = n$.

- First case: $T < n$.

The sum $\sum_{z_s} p(x_s, z_s)$ is over all the possible values of z_s (including $\vec{0}$). Using the Proposition 3.1, we have:

$$p(y_s) = \frac{(1-\varphi_s)^T - (1-\varphi_s)^n}{1-(1-\varphi_s)^n} + \sum_{z_s \neq \vec{0}} (1-q_s)^{K|z_s|} \quad \frac{\varphi_s^{|z_s|}(1-\varphi_s)^{T-V_s-|z_s|}}{1-(1-\varphi_s)^n}.$$

Now, it is easy to check that

$$\sum_{z_s \neq \vec{0}} [\varphi_s(1-q_s)^K]^{|z_s|}(1-\varphi_s)^{T-V_s-|z_s|} = [\varphi_s(1-q_s)^K + (1-\varphi_s)]^T - (1-\varphi_s)^T;$$

hence the result.

• Second case: $T = n$. Thus z_s and $\mathbf{z}_s$ are now confounded, and the sum $\sum_{z_s} p(x_s, z_s)$ is over all the possible values of z_s (excluding $\vec{0}$), due to the constraint $\mathbf{z}_s \neq \vec{0}$. We have:

$$p(y_s) = \sum_{z_s \neq \vec{0}} (1-q_s)^{K|z_s|} \quad \frac{\varphi_s^{|z_s|}(1-\varphi_s)^{T-V_s-|z_s|}}{1-(1-\varphi_s)^T}.$$

from which we easily deduce that:

$$p(y_s) = \frac{\left[(1-q_s)^K\varphi_s + 1 - \varphi_s\right]^T - (1-\varphi_s)^T}{1-(1-\varphi_s)^T}.$$

Appendix A3

If $S \sim \text{NegBin}\,(r, p)$, where $r \in]0, +\infty[$ and $p \in]0, 1[$, let us first recall that its probability mass function is such that:

$$\pi(S) \propto \frac{\Gamma(r+S)}{S!}(1-p)^S. \tag{1}$$

We now express r and p in terms of $E(S)$ and $\text{Var}(S)$. This is easily done by using the well known formulae $E(S) = r\frac{1-p}{p}$ and $\text{Var}(S) = r\frac{1-p}{p^2}$, from which we deduce that:

$$r = \frac{[E(S)]^2}{\text{Var}(S) - E(S)} \quad \text{and} \quad p = \frac{E(S)}{\text{Var}(S)}. \tag{2}$$

If we let $\mathrm{Var}(S) \longrightarrow +\infty$ in (2) it is clear that, for any fixed $E(S)$, $p \longrightarrow 0$ and $r \longrightarrow 0$. If we now let p and $r \longrightarrow 0$ in the right-hand side of (1), it is easy to verify that it tends to $1/S$, since $\Gamma(r+S) \longrightarrow \Gamma(S) = (S-1)!$ and $(1-p)^S \longrightarrow 1$.

References

Bunge, J. and Fitzpatrick, M. (1993) Estimating the number of species: A review. *Journal of the American Statistical Association* **8**, 364-373.

Castledine, B. J. (1981) A Bayesian analysis of multiple-recapture sampling for a closed population. *Biometrika* **67**, 197-210.

Chao, A. (2005) Species richness estimation and applications. *Encyclopedia of Statistical Sciences*, 2nd Edition, Vol. 12, 7907-7916, Wiley, New York.

Decamps, H., Joachim, J. and Lauga, J. (1987). The importance for birds of the riparian woodlands within the alluvial corridor of the river Garonne, s.w. France.*Regulated Rivers: Research and Management* **1**, 301-316.

Dorazio, R. M. and Royle, J. A. (2005) Estimating size and composition of biological communities by modeling occurrence of species. *Journal of the American Statistical Association* **100**, 389-398.

Dorazio, R. M., Royle, J. A., Soderstrom, B., and Glimskar, A. (2006) Estimating species richness and accumulation by modeling species occurrence and detectability. *Ecology* **87**, 842-854.

Dupuis, J. A. (1995) Bayesian estimation of movement and survival probabilities from capture-recapture data. *Biometrika* **82**, 761-772.

Dupuis, J. A. and Joachim, J. (2006) Bayesian estimation of species richness from quadrat sampling data in the presence of prior information. *Biometrics* **62**, 706-712.

Dupuis, J. A., Bled, F., and Joachim, J. (2010) Estimating occupancy rate of spatially rare or hard to detect species: a conditional approach. *Biometrics* (in press).

Hass, P.J., Liu, Y., and Stokes, L. (2006) An estimator of number of species from quadrat sampling. *Biometrics* **62**, 135-141.

Huston, M.A. (1994) *Biological diversity.* Cambridge University press, UK.

Jeffreys, H. (1961) *Theory of probability (3rd edition).* Oxford University press, Oxford.

Kass and Wasserman (1996) The selection of prior distributions by formal rules. *Journal of the American Statistical Association* **91**, 1343-1371

King, R. and Brooks, S.P. (2001) On the Bayesian estimation of population size. *Biometrika* **88**, 841-851.

King, R. and Brooks, S.P. (2008) On the Bayesian estimation of a closed population size in the presence of heterogeneity and model uncertainty. *Biometrics* **64**, 816-824.

Krebs, C.J. (1989). *Ecological Methodology.* Harper and Row, NY, USA.

Lauga, J. and Joachim, J. (1992). Modeling the effects of forest fragmentation on certain species of forest-breeding birds. *Landscape Ecology*, **6**, 183-193.

MacKenzie, D.I., Nichols, J.D., Royle J.A., Pollock, K.H., Bailey, L.L., Hines, J.E. (2006) *Occupancy Estimation and Modeling.* Elsevier, Amsterdam.

Mingoti, S.A. and Meeden, G. (1992) Estimating the total number of distinct species using presence and absence data. *Biometrics* **48**, 863-875.

Royle, n.A., Dorazio, R.M. and Link W.A. (2007) Analysis of multinomial models with unknown index using data augmentation. *Journal of Computational and Graphical Statistics* **16**, 67-85.

Smith, E. P. and Van Belle, G.V. (1984). Non parametric estimation of species richness. *Biometrics* **40**, 119-129.

Table 1

Posterior means and 95% *credible intervals of* S, α, β, σ_μ *and* σ_ν

	S	α	β	σ_μ	σ_ν
inner forest	32.5	-0.5	-0.4	0.8	2.1
	$[29, 38]$	$[-1.0, 0.1]$	$[-1.3, 0.5]$	$[0.4, 1.2]$	$[1.3, 3.4]$
edge	26.8	-0.9	0.7	1.2	1.1
	$[23, 32]$	$[-1.6, -0.2]$	$[-0.2, 1.7]$	$[0.7, 1.9]$	$[0.5, 2.4]$