On the Bayesian estimation of species richness and related quantities from quadrat sampling

Jérôme Dupuis

On the Bayesian estimation of species richness and related quantities from quadrat sampling

Jérôme A. Dupuis

L.S.P. Université Paul Sabatier, 118 Route de Narbonne, Toulouse, France

SUMMARY. We consider the problem of estimating the number of species of a biological community located in a region R divided into J quadrats. Recently, two parametric approaches have been developed which both take into account in a same modeling framework the detectability and the occurence of species in the quadrats. One developed by Dorazio and Royle (2005, J.A.S.A. **100**, 389-398) ignores the unsampled part of R (thus also J) and models the occurence of species only in the sampled quadrats. We show that this approach can be used only if J is large, which limits its use in practice, since the value of J can strongly vary from one study to another. The other developed by Dupuis and Joachim (2006, Biometrics **62**, 706-712) does not have this limitation but it applies only in the presence of prior information. In this paper, we propose a new approach which extends these two approaches since it can be used in non informative setups, and applies without limitation on J. We develop our approach within a simple model, which assumes that the species population is homogeneous. This constitutes the suitable framework for examining the effect that ignoring J has on the Bayesian estimate of species richness. In particular, a simulation study is undertaken which shows that the approach of Dorazio and Royle generates an error which can be important for small or moderate values of J, when species are spatially rare or hard to detect.

KEY WORDS: Bayesian estimate; Biodiversity, 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 and fundamental 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 (Krebs, 1989). Typically, the region R under investigation is first divided into J quadrats, and then a random sample of T quadrats from J is taken. From quadrat sampling data, the knowledge of the biodiversity of the region R can be increased by estimating quantities closely related to S, such as the number S_a of species present in any subregion R_a of R. For biological motivations, see e.g. Chao (2000).

Nonparametric methods have been proposed for estimating the species richness when quadrat sampling is used. They include the Jacknife and the boostrap estimates; for a complete review see Bunge and Fitzpatrick (1993) and Chao (2005). Recently, two parametric approaches have been developed which both take into account in a same modeling framework the two biological processes which underlie the data: the one related to occurence of species in the quadrats, and the other related to species detection. One approach has been developed by Dorazio and Royle (2005), the other by Dupuis and Joachim (2006). Both are preferable to the popular capture recapture methods, which are actually rarely relevant for estimating the species richness of a biological community, as underlined by Dorazio and Royle (2005). However, both suffer from serious limitations.

We consider that the approach developed in Dorazio and Royle (2005), and called afterwards the DR approach, needs to be discussed on three points.

• First, the number of quadrats in which R has been divided plays no part in the DR approach, while it is expected that the estimation of S depends both on T and J; see e.g. Mingoti and Meeden (1992), Hass *et al.* (2006). Furthermore, we note that no indication

or discussion is provided concerning the use of the DR approach with regard to the sampled fraction of R. One of the objectives of this paper is precisely to clarify this question.

• Second, Dorazio and Royle (2005) model the occurence of species in the sampled quadrats, by T independent Bernoulli outcomes. We show in this paper that this assumption of independence is not consistent with the fact that their approach is conditional, which means that they model the occurence of species effectively present in R.

• Third, their approach is not fully Bayesian but rather *ad hoc*. Indeed, credible posterior intervals are provided for S, but no prior distribution is actually put on S. By way of contrast, in this paper we pay particular attention to the prior adopted for this parameter.

Dupuis and Joachim (2003, 2006) developed a Bayesian parametric model to estimate S. In contrast to Dorazio and Royle (2005), their approach, called afterwards the DJ approach, is unconditional, in that it models the occurence of species liable to be present in the region R. It applies without limitation on J, but it is limited to situations where a list $\mathcal{L}$ of species liable to be present in the region R can be drawn up, and requires prior information on the probabilities of presence of each species of $\mathcal{L}$ not detected, which significantly limits the use of this approach. In particular, the approach of Dupuis and Joachim (2003, 2006) cannot be used in unknown regions. We note that Dorazio *et al.* (2006) have also developed an unconditional approach. It presents similarities with the DJ approach. In particular, these authors introduce a supercommunity of species which is supposed to encompass the species population of interest. Its implementation is relatively easy when the region R is known, but is problematic in the opposite case. Furthermore, this approach can be used only if J is large, simply because Dorazio *et al.* (2006) model the occurence of species present in R as in Dorazio and Royle (2005).

In this paper we propose a new approach which does not require the above list $\mathcal{L}$, and

can be used with totally unknown regions. It also provides answers to the reservations formulated regarding the DR approach. As Dorazio and Royle (2005) we adopt a conditional approach, but the key difference is that we model the occurence of species in the J quadrats (not only in the T quadrats). Therefore J and T are included in our model. We show that the DR approach is valid only asymptotically (with respect to J), which significantly limits its use (since the range of possible values of J is particularly large in practice, as illustrated in Section 2.1). In fact, the DR approach can be viewed as a particular case of ours, more precisely as a limiting case (namely $J \longrightarrow \infty$). Since J is finite (even though it might be very large in some studies), using the DR approach thus induces an error which needs to be evaluated.

In this paper we focus on a simple model, called $\mathcal{M}_0$, which assumes that the species population is homogeneous; it can be viewed as the correct version of the DR model (in its homogeneous version). We emphasize that the methodology developed in this paper also applies to heterogeneous populations. (However, the heteregeneous case requires specific and rather lengthy developments and is beyond the scope of this paper.) The model $\mathcal{M}_0$, which includes a small number of parameters, is actually a suitable framework for pointing out the limitations of the DR approach, as well as for examining the extent to which the error entailed by the use of the DR model (instead of $\mathcal{M}_0$) is liable to affect the Bayesian estimation of S. However, the model $\mathcal{M}_0$ exhibits some statistical difficulties which have to be overcome before undertaking this analysis. They are due to the presence of randomly missing data: a species can be present in R but, for practical reasons, not have been detected. There are two types of statistical difficulty. First, obtaining the Bayesian estimates of S and S_a involves computational difficulties, which are overcome by implementing a Markov Chain Monte Carlo (MCMC) algorithm which uses a suitable partitioning of the missing data. Second, some parameters can be non identifiable (as it

is often the case in missing data models). In our context this is effectively the case if each sampled quadrat is visited only once; but we prove that all the parameters are identifiable if repeated observations are made in sampled quadrats.

The paper is organized as follows. The experimental protocol and the missing data structure are described in Section 2. Section 3 is devoted to the model $\mathcal{M}_0$. The approach of Dorazio and Royle (2005) is discussed in Section 4. The links existing between conditional and unconditional approaches are clarified in Section 5. We conclude in Section 6.

2. The experimental protocol and the missing data structure

2.1 *The experimental protocol and data description*

Before describing the protocol, we stress that the region R under investigation has to be of course bounded, or else the quantity S is not well defined, and estimating S is an ill-posed problem (as pointed out by Dorazio *et al.,* 2006). Quadrat sampling can be done in two distinct variants.

- The first variant is described in details for example in Dorazio *et al.* (2006), cf the Section *Protocol for Sampling Communities*. It proceeds as follows. The region R is divided into J 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 K times each sampled quadrat and records the species detected in each. As in Dorazio and Royle (2005), we assume that $K \geq 2$. Detections are typically based on visual or aural recognitions; we assume that species are correctly identified.

- The second variant slightly differs from the one above described. It proceeds as follows: T quadrats of equal area, say $\mathcal{A}$, are placed (typically at random) in the region R; then, they are explored as above indicated (see e.g. Mingoti and Meeden, 1992). Contrary to

the previous protocol, the region R is not beforehand divided in quadrats. However, we emphasize that the model M_0 also applies to this procotcol, provided the unsampled part of R can be divided (even virtually) in T^* quadrats of area $\mathcal{A}$ (which is generally the case, in practice); so we have $J = T + T^*$.

When $K = 4$ and $T = 6$, a possible record (or history) for a species s present in R 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, its presence has not been detected in quadrat $2, 3, 4, 6$.

The problem is to estimate the number S of species present in R, from the data formed by the records of species whose the presence has been detected at least once. Sometimes, all the J quadrats are explored (thus $T = J$). Although the methodology developed in this paper focuses on the case $J < T$, we emphazise that it also applies to the case $T = J$, with straightforward changes. This remark is of importance, since when J is not too large, sampling the whole region R is not rare (see below).

The size of J is a key element in the discussion concerning the validity of the DR approach (cf Section 4.3). In practice, the range of J is particularly large, since its value can strongly vary from one survey to another. Let us illustrate this variability through a few examples, by limiting ourselves to ornithological surveys made in the south of France. In Joachim *et al.* (1990), ten forests have been divided in $J \leq 15$ quadrats; in this study the values of J are thus small. The size of J can be moderate, as in Dupuis and Joachim (2006) where the forest of Montech has been divided in $J = 40$ quadrats; or as in Decamps *et al.* (1987) where the Bouconne forest has been divided in $J = 98$ quadrats. It can be large, as in Lauga and Joachim (1992), where R includes $J = 5865$ quadrats. The distinction between small, moderate and large J is of course rather arbitrary; it actually rests on some observations made in the framework of our simulation study (cf Section

4.3). The data analysed by Dorazio and Royle (2005), as well as by Dorazio *et al.*, (2006), concern the *North American breeding bird survey*; it includes around 4000 roadsides, and each route contains 50 quadrats. In this survey, J is thus particularly large.

2.2 *A specific missing data structure.*

To specify the missing data structure inherent in quadrat sampling data, we view the record $y_s = (y_{sj}; j = 1, \ldots, T)$ as the result of two processes: one is related to the presence-absence process, and the other is related to the detection process. Such a formalism also allows us to formulate rigorously the biological assumptions made and to introduce, in a natural way, the parameters of biological interest.

- For $s = 1, \ldots, S$ and $j = 1, \ldots, J$, we denote by z_{sj} the indicator of presence of species s in quadrat j. The vector $(z_{sj}; j = 1, \ldots, J)$ is denoted by z_s.

- For a species s present in a sampled quadrat j, we denote by x_{sj} the number of times that species s has been detected in quadrat j during the K visits. Note that x_{sj} is thus defined only conditionally on $z_{sj} = 1$; the vector formed by the x_{sj}'s is denoted by x_s. Note also that $x_{sj} = 0$ and $y_{sj} = 0$ do not have the same meaning (see below).

Missing data can occur in different circumstances. First, when a species s has not been detected in quadrat j, it is clear that z_{sj} is missing; this event covers in fact two exclusive situations: either species s is present in quadrat j but has not been detected, or it is not present in quadrat j (and cannot have been detected). Formally, one has the equivalence $(y_{sj} = 0) \Longleftrightarrow (z_{sj} = 1 \text{ and } x_{sj} = 0)$ or $(z_{sj} = 0)$. Conversely, when $1 \leq k \leq K$, one has $(y_{sj} = k) \Longleftrightarrow (z_{sj} = 1 \text{ and } x_{sj} = k)$. Secondly, z_{sj} is missing, when quadrat j is not a part of the sampled quadrats. Thirdly, a species s present in R and undetected is such that the whole vector z_s is missing. The set of the missing z_{sj}'s is denoted by $\mathbf{z}_m$. The above observations lead to partitioning $\mathbf{z}_m$ into four blocks, according to the status of quadrat j (sampled or not) and the status of species s (detected or not); these blocks are

denoted by $\mathbf{z}_m^{[i]}$; $i = 1, .., 4$ and defined as follows. The blocks $\mathbf{z}_m^{[1]}$ and $\mathbf{z}_m^{[2]}$ are formed by the missing z_{sj}'s where s designates any detected species, and j denotes either a sampled quadrat ($i = 1$) or an unsampled quadrat ($i = 2$). The blocks $\mathbf{z}_m^{[3]}$ and $\mathbf{z}_m^{[4]}$ are formed by the missing z_{sj}'s where s designates any undetected species, and j denotes either a sampled quadrat ($i = 3$) or an unsampled quadrat ($i = 4$).

3. The homogeneous model: $\mathcal{M}_0$

3.1 *Notation, assumptions and parameters.*

Notation is basically the one adopted in Dupuis and Joachim (2003, 2006). $p(.)$ denotes a probability mass function, and $\pi(.)$ the density of any prior distribution. The null vector is denoted by $\vec{0}$, and the vector which has all its components equal to 1, by $\vec{1}$. We denote by $\mathcal{Z}$ the set $\{0,1\}^J$, and by $\mathcal{Z}^*$ the set $\mathcal{Z}\backslash\{\vec{0}\}$. Let v be a vector; we denote by $|v|$ the sum of all its components. $\mathbb{1}_{(C)}$ denotes an indicator function that takes the value 1 when the condition C is true and zero otherwise. The Bayesian estimate of S is denoted by $\widehat{S}_0$. We denote by $\overline{z}_{sj}$ the vector z_s from which the j-th component has been removed, and by z_s^* the vector formed by the sampled z_{sj}; thus $z_s^* = (z_{sj}; j = 1, \ldots, T)$. Finally, we set $\mathbf{x} = (x_s; s = 1, \ldots, S)$ and $\mathbf{z} = (x_s; s = 1, \ldots, S)$.

We assume that the species population is closed, in that the number of species present in R is constant throughout the quadrat sampling experiment. Biological assumptions related to the occurence of species in the J quadrats are supported by the z_s's, and those related to the detections in the T quadrats by the x_s's.

Assumption A1. We assume that:

$$p(\mathbf{z}) = \prod_{s=1}^{S} p(z_s) \quad \text{and} \quad p(\mathbf{x}|\mathbf{z}) = \prod_{s=1}^{S} p(x_s|z_s).$$

Assumption A2. We assume that the probability of detecting species s in quadrat j does not depend on its (possible) detections in the other quadrats.

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

Assumptions A1, A2 and A3 are standard (they are also present in the DR's model). A1 means that the species present in R do not interact relative to their detectability and their presence (in the quadrats). We thus exclude predator-prey relationships between species. A3 means that, s and j being fixed, the detections of species s during the K visits in quadrat j are independent. Note that q represents the probability of detecting species s in quadrat j during any visit, given that it is present in quadrat j.

Assumption A4. For any species s present in R, we assume that:

$$p(z_s|\varphi) = \frac{\varphi^{|z_s|}(1-\varphi)^{J-|z_s|}}{1-(1-\varphi)^J} \tag{3.1}$$

where $\varphi \in]0,1[$, and $|z_s|$ represents the number of quadrats in which species s is present. Assumption A4 thus models the occurence of species in the J quadrats; this step of the modeling is absent from the DR approach which model their presence only in the T sampled quadrats (see Section 4.1 for details). This is a key difference between the two approaches. Note that z_s takes its values in $\mathcal{Z}^*$ since z_s cannot take the value $\vec{0}$ (recall that s designates a species present in R) and that $p(.)$ given by (3.1) effectively defines a probability mass function on $\mathcal{Z}^*$. We now provide two results concerning the distribution of z_s, subsequently useful to establish the distribution of the missing data conditionally on the (observed) data.

Proposition 3.1. Let $z_s(a)$ be a vector extracted from z_s of length J_a. We have:

$$p(z_s(a)|\varphi) = \frac{\varphi^{|z_s(a)|}(1-\varphi)^{J_a-|z_s(a)|}}{1-(1-\varphi)^J}$$

if $z_s(a) \neq \vec{0}$, and $p(z_s(a)) = \left[(1-\varphi_s)^{J_a} - (1-\varphi)^J\right] / \left[1-(1-\varphi)^J\right]$ otherwise. Let $z_s(a)$, $z_s(b)$, $z_s(c)$ denote three vectors extracted from z_s and pairwise disjoint. If $z_s(c) \neq \vec{0}$, thus $z_s(a)$ and $z_s(b)$ are independent, conditionally on $z_s(c)$.

proof. See Appendix A1.

By using the proposition (3.1) above, it is easy to verify that $\text{pr}(z_{sj} = 1|\overline{z}_{sj} \neq \vec{0}) = \varphi$; therefore, φ represents the probability that species s is present in quadrat j, given that it

is present in at least one other quadrat. Note that the z_{sj}'s are not independent, although a certain form of conditional independence between the z_{sj}'s holds. Indeed, from the proposition (3.1), one deduces that if i, j, l denote three distinct quadrats, $z_{s,i}$ and $z_{s,j}$ are independent, conditionally on $z_{s,l} = 1$. In other words, the presence of species s in quadrat i does not affect its presence in quadrat j (on condition that it is present in a third quadrat). This biological assumption is standard; it has been discussed in Dupuis and Joachim (2006, 2003).

Note that we do not make the parameters q and φ depend on s and j, which constitutes the characteristics of the homogeneous model $\mathcal{M}_0$. We stress that the model M_0 deals only with species s present in R; the approach adopted in this paper is thus conditional. To avoid needless discussions, we assume that at least one species has been detected by the quadrat sampling, so that the data set is not empty. Concerning S, this implies that we assume that $S \geq 1$.

3.2 *Likelihood under the model $\mathcal{M}_0$.*

Recall that each record (or history) is described by a vector of length T and having all its components in $\{0, 1, \ldots, K\}$ (cf Section 2.1). Histories are numbered from $h = 0$ to $H = K^T - 1$; the history numbered 0 being associated with the record $\vec{0}$. For convenience, the history numbered h is afterwards refered as the history h. For $h = 0$ to $H = K^T - 1$, we denote by n_h the number of species having the history h. Moreover we denote by d the number of detected species; note that $n_0 = S - d$. Data are denoted by $\mathbf{y}$ and formed by the H counts $\{n_1, \ldots, n_H\}$ which clearly constitutes a sufficient statistic. The probability $\mathrm{pr}(y_s = h)$ is denoted by β_h. Let s be any species having $h \neq 0$ as its history, we denote by v_h the number of quadrats in which species s is detected, and by w_h the total number of visits during which it is detected. Let $h \in \{1, \ldots, H\}$ and $(b_1, \ldots, b_j, \ldots, b_T)$ be the

history h, thus we set

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

Due to the presence of missing data, writing down the likelihood in a closed form is not immediate, but performing this task is greatly facilitated by the formalism introduced in Section 2.2.

Proposition 3.2. The likelihood of the data $\mathbf{y}$ under the model $\mathcal{M}_0$ is:

$$L_0(\theta; \mathbf{y}) = p(\mathbf{y}|S, \varphi, q) = \frac{S!}{(S-d)! \prod_{h=1}^{H} n_h!} \, \beta_0^{S-d} \prod_{h=1}^{H} \beta_h^{n_h} \tag{3.2}$$

where

$$\beta_h = \frac{\rho_h \, q^{w_h} (1-q)^{K v_h - w_h} \, \varphi^{v_h} \left[(1-q)^K \varphi + 1 - \varphi \right]^{T - v_h}}{1 - (1-\varphi)^J} \tag{3.3}$$

if $h \neq 0$, and

$$\beta_0 = \frac{\left[(1-q)^K \varphi + 1 - \varphi \right]^T - (1-\varphi)^J}{1 - (1-\varphi)^J}. \tag{3.4}$$

proof. See Appendix A2.

We set $V = \sum_{h=1}^{H} v_h$ and $W = \sum_{h=1}^{H} v_h$. From the above proposition, we deduce that:

$$L_0(\theta; \mathbf{y}) \propto \frac{S!}{(S-d)!} \, \frac{\varphi^V \, q^W (1-q)^{KV - W} \, F(\varphi, q)^{dT - V} \left[F(\varphi, q)^T - (1-\varphi)^J \right]^{S-d}}{\left[1 - (1-\varphi)^J \right]^d}$$

where $F(\varphi, q) = (1-q)^K \varphi + 1 - \varphi$ and the constant of proportionality depends only on $\mathbf{y}$. It is thus clear that $\mathcal{Y} = (d, V, W)$ is a sufficient statistic. This remark is used later, in Section 4.3. Moreover, we emphasize that the expressions (3.3) and (3.4) also hold when $T = J$ (for brevity, details are omitted).

3.3 *The prior distributions.*

We assume that S, φ, q are a priori independent; thus, $\pi(\theta) = \pi(S)\pi(\varphi)\pi(q)$. Uniform distributions are placed on q and φ. Of course, beta distribution can be used if some prior information is available. The negative binomial distribution is usually the distribution adopted in practice for incorporating some prior information on an integer parameter;

see, eg King and Brooks (2001), in a capture-recapture set-up where the parameter of interest is the size of the animal population. Since the negative binomial distribution can be parametrized by its mean and its variance (cf Appendix 3), adopting such a distribution is thus very flexible; in particular, a large variance for S will reflect a poor prior on S.

When no prior information on an integer parameter is available, the improper Jeffreys prior distribution, is usually adopted: that is $\pi(S) \propto \frac{1}{S}$ in our context. Now, to our knowledge, no 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 put for a real parameter $\alpha > 0$ (namely, $\pi(\alpha) \propto 1/\alpha$) to the case of an integer parameter. Our motivation for the use of the Jeffreys prior as a non informative prior rests on that it coincides with the limiting case of a negative binomial distribution in which the prior variance tends to ∞ (the prior mean being fixed). Details appear in Appendix A3. The main alternative to the Jeffreys prior is to take $\pi(S) \propto 1$ (eg Casteldine, 1981). In the next Section, we explain why these two non informative priors should give, in practice, very similar estimations of S.

3.4 *Identifiability issues.*

As mentioned in the introduction, the parameter θ of the model $\mathcal{M}_0$ is not identifiable if each sampled quadrat is visited only once (that is when $K = 1$); for brevity, the proof is omitted. When $K \geq 2$, we have established an identifiability result which holds under a very slight restriction.

Theorem 3.3. We assume that $S \geq 2$. The parameters S, φ and q are all identifiable.

proof. See Appendix A4.

3.5 *MCMC algorithms for estimating S and related quantities*

This Section is organized as follows. At first, we consider the problem of estimating S alone. To obtain $\mathrm{E}[S|\mathbf{y}]$, two MCMC algorithms are considered: one uses the missing data

structure exhibited in Section 2.2, the other uses only the likelihood of the observed data (3.2). The first algorithm is called Alg01, and the second one Algo2. Both are described in the next Section. Next, we show how to obtain the Bayesian estimates of some quantities related to S; only Alg01 is able to do this (with slight modifications according to which quantity we aim to estimate).

3.5.1 Estimating S alone.

We first describe the MCMC algorithm Algo1. It is implemented on $(\theta, \mathbf{z}_m)$; such a strategy is actually standard in missing data models (see eg Robert and Casella, 2004). Algo1 is a Metropolized Gibbs sampling algorithm. The parameter φ is up-dated via a Hastings-Metropolis step. We up-date q and $\xi = (S, \mathbf{z}_m)$ via Gibbs steps, as follows:

$$q \sim \pi(q|\xi, \varphi, \mathbf{y}) \quad \text{and} \quad \xi \sim p(\xi|\varphi, q, \mathbf{y}),$$

where $\xi = (S, \mathbf{z}_m)$ is simulated as follows:

$$S \sim \pi(S|\varphi, q, \mathbf{y}) \quad \text{and} \quad \mathbf{z}_m \sim p(\mathbf{z}_m|S, \varphi, q, \mathbf{y}).$$

Before indicating how to implement each step, we provide the expression of the complete data likelihood denoted by $L_0(\theta; \mathbf{y}, \mathbf{z}_m)$. By using assumptions A1, A2, A3 and A4, and by observing that $(\mathbf{y}, \mathbf{z}_m)$ and $(\mathbf{x}, \mathbf{z})$ provide the same information on θ, it is easy to check that:

$$L_0(\theta; \mathbf{y}, \mathbf{z}_m) \propto \frac{\varphi^{V'}(1-\varphi)^{JS-V'}}{[1-(1-\varphi)^J]^S}\ q^W(1-q)^{K(V_m^{[1]}+V_m^{[3]})} \tag{3.7}$$

where $V' = V + V_m$, $V_m = V_m^{[1]} + V_m^{[2]} + V_m^{[3]} + V_m^{[4]}$, and $V_m^{[i]}$ denotes the sum of the z_{sj}'s belonging to $\mathbf{z}_m^{[i]}$.

Updating S. S is simulated according to the distribution of $S|q, \varphi, \mathbf{y}$. Let $\pi(\theta|\mathbf{y})$ denote the density of the posterior distribution. We have $\pi(\theta|\mathbf{y}) \propto L_0(\theta; \mathbf{y})\pi(\theta)$. Taking into

account the expression of $L_0(\theta; \mathbf{y})$, we have:

$$\pi(S|q, \varphi, \mathbf{y}) \propto \frac{S!}{(S-d)!}\, \lambda_0^{S-d}\, \pi(S)\; \mathbb{1}_{(S \geq d)} \tag{3.8}$$

where λ_0 is given by (3.4). The indicator $\mathbb{1}_{(S \geq d)}$ expresses the fact that, conditionnally on $\mathbf{y}$, S is necessarily greater than d. From (3.8), it is straightforward to deduce that, $S-d|q, \varphi, \mathbf{y}$ follows a NegBin$(d, 1-\beta_0)$ distribution if $\pi(S) \propto 1/S$, a NegBin$(d+1, 1-\beta_0)$ distribution if $\pi(S) = 1$. Therefore, these two non informative priors should give very close estimates of S, as long as 1 is small compared with d (which is the case in most studies). Finally, $S-d|q, \varphi, \mathbf{y}$ follows a NegBin$(d+r, 1-(1-b)\beta_0)$ if $S \sim$ NegBin(r, b).

Updating $\mathbf{z}_m$. Due to the form of the complete data likelihood we simply have to simulate the $V_m^{[i]}$'s. Simulating $V_m^{[1]}$ and $V_m^{[2]}$ directly uses the proposition 3.5

proposition 3.4. Conditionally on $(\theta, \mathbf{y})$, the two blocks $\mathbf{z}_m^{[1]}$ and $\mathbf{z}_m^{[2]}$ are independent. $V_m^{[1]}|\mathbf{y}, \varphi, q \sim$ Binomial$(dT - V, \gamma)$ where $\gamma = [\varphi(1-q)^K]/[\varphi(1-q)^K + (1-\varphi)]$, and $V_m^{[2]}|\mathbf{y}, \varphi, q \sim$ Binomial$((J-T)d, \varphi)$. For any undetected species s we have:

$$p(z_s|y_s = \vec{0}, \varphi, q) = \frac{(1-q)^{K|z_s^*|}\; \varphi^{|z_s|}\; (1-\varphi)^{J-|z_s|}}{\left[(1-q)^K\varphi + 1 - \varphi\right]^T - (1-\varphi)^J} \tag{3.9}$$

proof. See Appendix A5.

Simulating $V_m^{[3]}$ and $V_m^{[4]}$ needs special attention, because these two blocks are not, conditionally on $(\theta, \mathbf{y})$, independent, due to the constraint $z_s \neq \vec{0}$. Recall also that the species s involved in the simulation of $V_m^{[3]}$ and $V_m^{[4]}$ are those which have not been detected (that is such that $y_s = \vec{0}$), and for which the whole vector z_s is missing and has to be simulated according to (3.9). Simulating $V_m^{[3]}$ and $V_m^{[4]}$ proceeds as follows. Set $V_m^{[3]} = 0$ and $V_m^{[4]} = 0$, and repeat $S-d$ times the two following steps:

1. Simulate z_s according to (3.9).

2. Set $V_m^{[3]} = V_m^{[3]} + |z_s^*|$ and $V_m^{[4]} = V_m^{[4]} + |z_s^\circ|$, where $z_s^\circ = (z_{sj}; j = T+1, J)$.

Updating q**.** Simulate q according to a Beta$\Big(1+W, 1+K(V_m^{[1]}+V_m^{[3]})\Big)$.

Updating φ**.** This is done via a Hastings-Metropolis step, as follows. The proposal φ' is accepted with probability:

$$\min\Big\{\, 1, \frac{\pi(\varphi'|S,q,\mathbf{z}_m,\mathbf{y})}{\pi(\varphi|S,q,\mathbf{z}_m,\mathbf{y})} \times \frac{g(\varphi|\varphi')}{g(\varphi'|\varphi)} \,\Big\}$$

where $\varphi, q, S, \mathbf{z}_m$ represent the current values, and $g(.|.)$ denotes the density of the instrumental distribution (typically an uniform distribution). Note that $\pi(\varphi|S,q,\mathbf{z}_m,\mathbf{y}) \propto L_0(\theta;\mathbf{y},\mathbf{z}_m)$ which is given by (3.7).

The other algorithm Algo2 does not use the missing data structure of quadrat sampling data; it only uses the observed data likelihood given by (3.2). Parameter S is updated exactly as in Algo1. φ and q are updated via Hastings-Metropolis steps. Details are omitted since this is straightforward.

3.5.1 Estimating some quantities related to S.

We denote by M_j the number of species present in the sampled quadrat j, and by M the total number of species present in the sampled part of R. More generally, we consider the quantity S_a, which represents the number of species present in some subregion R_a composed both of sampled and unsampled quadrats. Only the quantities M_j (where j designates a sampled quadrat) and M have been considered by Dorazio and Royle (2005), and we have doubts about the validity of their estimates of M_j and M. Explanations are postponed until the end of the Section 5.

Estimating M, M_j and S_a, uses the fact that each of these quantities is a (simple) function of S and $\mathbf{z}$. For example, $S_a = \sum_{s=1}^{S} \mathbb{1}_{(|z_s(a)|\geq 1)}$ where $z_s(a)$ denotes the vector $(z_{sj}; j \in R_a)$. Now, recall that, once the data $\mathbf{y}$ are available, only a part of $\mathbf{z}$ is known, the other part being missing. Thus, conditionally on $\mathbf{y}$, estimating any of these quantities comes down to estimating its missing part. It is the reason why the missing data approach

is particularly well suited to estimating M_j, M and S_a.

• For obtaining $\widehat{M}$, we use Algo1 without any modification. The Bayesian estimate of M is obtained by applying the ergodic theorem, as follows:

$$d + \frac{1}{L}\sum_{l=1}^{L} h(S^{(l)}, \mathbf{z}_m^{(l)}) \longrightarrow E[M|\mathbf{y}] = \widehat{M} \quad (L \longrightarrow +\infty)$$

where (l) indicates the step of the algorithm, and $h(S, \mathbf{z}_m) = \sum_{s=d+1}^{S} \mathbb{1}_{(|z_s^*| \geq 1)}$.

• For obtaining $\widehat{M}_j$, in addition to simulating $V_m^{[1]}$ as indicated in the previous Section, we simulate separately each missing $z_{sj} \in \mathbf{z}_m^{[1]}$ (j being fixed); it is done according to the distribution of $z_{sj}|y_s$, that is according to a Bernoulli(γ) distribution, where γ is given by proposition 3.5. The simulation of $V^{[2]}$, $V^{[3]}$ and $V^{[4]}$ are without change. The Bayesian estimate of S_j is then obtained by applying the ergodic theorem:

$$d_j + \frac{1}{L}\sum_{l=1}^{L} h(S^{(l)}, \mathbf{z}_m^{(l)}) \longrightarrow E[M_j|\mathbf{y}] = \widehat{M}_j \quad (L \longrightarrow +\infty)$$

where d_j denotes the number of species detected in the sampled quadrat j, and

$$h(S, \mathbf{z}_m) = \sum_{s=1}^{d} \mathbb{1}_{(z_{sj}=1,\, y_{sj}=0)} + \sum_{s=d+1}^{S} \mathbb{1}_{(z_{sj}=1)}.$$

• We denote by J_a the number of quadrats (among the J quadrats) located in R_a, and by T_a the number of sampled quadrats (among the J_a). We denote by $\mathbf{y}_a$ the part of the data collected in these T_a quadrats, and by d_a the number of species detected in R_a. Note that it will not be correct to perform inference on S_a, only on the basis of $\mathbf{y}_a$, because $\mathbf{y}_a$ and $\mathbf{y}_b$ (where $\mathbf{y}_b = \mathbf{y} \backslash \mathbf{y}_a$) are not marginally independent (they are independent only conditionally on θ). For obtaining the species richness $\widehat{S}_a$ of R_a, we complete the simulation of $(V_m^{[1]}, V_m^{[2]})$ and $(V_m^{[3]}, V_m^{[4]})$, as follows.

- Concerning $(V_m^{[1]}, V_m^{[2]})$, set $E_a = 0$ and repeat $d - d_a$ times the two following steps:

1. Simulate $|z_s^*(a)| \sim$ Binomial(T_a, γ) and $|z_s^\circ(a)| \sim$ Binomial$(J_a - T_a, \varphi)$, where $z_s^*(a) = (z_{sj}(a); j = 1, T_a)$ and $z_s^\circ(a) = (z_{sj}(a); j = T_a + 1, J_a)$.

2. Set $E_a = E_a + \mathbb{1}_{(|z_s(a)|\geq 1)}$.

- Concerning $(V_m^{[3]}, V_m^{[4]})$, set $F_a = 0$ and add to the steps 1 and 2 of the paragraph untitled *Updating* $\mathbf{z}_m$ of the Section 3.5.!1, the following step: $F_a = F_a + \mathbb{1}_{(|z_s(a)|\geq 1)}$.

Note that E_a represents the number of species present in R_a, not detected in R_a, but detected in R_b, where $R_b = R \backslash R_a$. F_a represents the number of species present in R_a but undetected. $\widehat{S}_a$ is obtained by applying the ergodic theorem:

$$d_a + \frac{1}{L}\sum_{l=1}^{L}[E_a^{(l)} + F_a^{(l)}] \longrightarrow E[S_a|\mathbf{y}] = \widehat{S}_a \quad (L \longrightarrow +\infty)$$

4. Discussion of the model of Dorazio and Royle.

This Section is organized as follows. In Section 4.1, we explain why the way of which Dorazio and Royle (2005) model the occurence of species in the quadrats is not consistent with the fact that their approach is conditional. In Section 4.2, we first provide the likelihood of the model of Dorazio and Royle (2005) in its homogeneous version (called afterwards the DR model). Then, we show that the DR approach is valid only asymptotically (with respect to J). In Section 4.3, a simulation study is performed to quantify the error resulting from the use of the DR model instead of $\mathcal{M}_0$.

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 (2003, 2006).

4.1 *The model of Dorazio and Royle and the constraint* $z_s \neq \vec{0}$

First, recall that the approach of Dorazio and Royle (2005) is conditional; that is they model the occurence of species present in R. Consequently the constraint $z_s \neq \vec{0}$ applies to their approach. It is of interest to point out that this constraint does not exist when an unconditional approach is adopted since it models the occurence of species liable to be present in R, and it is quite possible that $z_s = \vec{0}$, for some s (see Section 6 for details).

Recall also that Dorazio and Royle (2005) do not model the occurence of species s in the J quadrats, but only its occurence in the T sampled quadrats, by assuming that

$z_{s,1}, \ldots, z_{s,T}$ are independent outcomes of a Bernoulli random variable of parameter ψ (cf line 6 of the right part of the page 391 of their paper). The $J - T$ remaining quadrats are thus ignored in their modeling, as well as the constraint $z_s \neq \vec{0}$. From the independence assumption of Dorazio and Royle (2005), called afterwards the DR assumption, it follows that:

$$p(z_s^*) = \psi^{|z_s^*|}(1-\psi)^{T-|z_s^*|}, \tag{4.1}$$

where $|z_s^*|$ represents the number of sampled quadrats in which species s is present. The dependence of ψ on s, adopted by these authors, is without importance in this discussion, and has been removed for convenience. The dependence of ψ on j, equally adopted by these authors, is considered later. The expression of $p(z_s^*)$ given by (4.1) has to be compared with the one calculated under the model M_0, and deduced from the proposition (3.1):

$$p(z_s^*) = [\varphi^{|z_s^*|}(1-\varphi)^{T-|z_s^*|}] \,/\, [1-(1-\varphi)^J] \quad \text{if} \quad z_s^* \neq \vec{0} \tag{4.2}$$

and $[(1-\varphi)^T - (1-\varphi)^J] \,/\, [1-(1-\varphi)^J]$ if $z_s^* = \vec{0}$. Note that $p(z_s^*)$ in (4.2) depends on J and T (contrary to 4.1). We suggest that the modeling adopted by Dorazio and Royle (2005) is not correct: the problem comes from the fact that the DR assumption is not consistent with the constraint $z_s \neq \vec{0}$. We support this assertion by means of two arguments.

Argument 1. This concerns the case $T = J$ (that is when the whole region R has been sampled). Assuming that $z_{s,1}, \ldots, z_{s,T}$ are independent implies that $\mathrm{pr}(z_s = \vec{0}) = (1-\psi)^J$; hence the contradiction, since it gives a positive probability to the event $z_s \neq \vec{0}$, which is not consistent with the constraint $z_s \neq \vec{0}$. In fact, $\mathrm{pr}(z_s = \vec{0}) = 0$ implies $\psi = 1$, that is $z_s = \vec{1}$ with probability one, which means that species s is present in all the quadrats of R (with probability one). Now, this situation has been discarded by Dorazio and Royle (2005) as being not tenable from a biological point of view; recall that it is precisely this

latter point which motivated their paper (cf the introduction).

Argument 2. The distribution of z_s should (of course) not depend on the number T of sampled quadrats. Now, this is not the case when the DR assumption is adopted, as illustrated by a simple example. We assume that $J = 5$. When all the quadrats have been sampled (that is when $T = J = 5$), we have $\text{pr}(z_s = 00001) = (1-\psi)^4\psi$, by using (4.1). We now asssume that 80% of the quadrats have been sampled (thus $T = 4$), and that the sampled quadrats are the quadrats numbered from 1 to 4. We can write $\text{pr}(z_s = 00001) = \text{pr}(z_{s5} = 1 | z_{s1} = z_{s2} = z_{s3} = z_{s4} = 0) \times \text{pr}(z_{s1} = z_{s2} = z_{s3} = z_{s4} = 0)$. Due to the constraint $z_s \neq \vec{0}$ we have $\text{pr}(z_{s5} = 1 | z_{s1} = z_{s2} = z_{s3} = z_{s4} = 0) = 1$; moreover due to (4.1) we have $\text{pr}(z_{s1} = z_{s2} = z_{s3} = z_{s4} = 0) = (1-\psi)^4$; hence $\text{pr}(z_s = 00001) = (1-\psi)^4$ which differs from the value obtained when $T = 5$. Therefore, when one adopts the DR assumption, the distribution of z_s depends on T, which is not acceptable. Note that the two values of $\text{pr}(z_s = 00001)$ are in fact equal if and only if $\psi = 1$.

If we now make ψ depend on j, that does not modify the conclusions. Details appear in Appendix A6.

4.2 *The likelihood under the DR model*

The model considered by Dorazio and Royle meets the Assumptions A1, A2, and A3. The model $\mathcal{M}_0$ and the DR model (that is the homogeneous version of the model considered in Dorazio and Royle, 2005) actually differ only from the way of modeling the occurence of species in the quadrats. The DR model is parametrized by $\omega = (S, \psi, q)$, where ψ denotes the probability that $z_{sj} = 1$. Establishing the expression of the likelihood under the DR model, denoted afterwards by $L_{DR}(\omega; \mathbf{y})$, is straigthforward, contrary to $L_0(\theta; \mathbf{y})$. (The fact of ignoring the unsampled quadrats and of assuming the independance between the

sampled quadrats greatly facilitates the calculations.) If we let μ_h=pr$(y_s = h|q, \psi)$

$$L_{DR}(\omega; \mathbf{y}) = \frac{S!}{(S-d)! \prod_{h=1}^{H} n_h!} \, \mu_0^{S-d} \prod_{h=1}^{H} \mu_h^{n_h}$$

where

$$\mu_h = \rho_h \, q^{w_h} (1-q)^{K v_h - w_h} \, \psi^{v_h} \, [(1-q)^K \psi + 1 - \psi]^{T - v_h} \quad \text{if} \quad h \neq 0,$$

where ρ_h has been defined in Section 3.2 and $\mu_0 = [(1-q)^K \psi + 1 - \psi]^T$.

If we make $J \longrightarrow +\infty$ (φ, q, T and K being fixed), we observe that the likelihood of the DR model and ours coincide. Indeed, $(1-\psi)^J \longrightarrow 0$ when J tends to $+\infty$, and the two parameters φ and ψ are confounded asymptotically. To check this second point, we introduce the parameter ν equal to the probability that $z_{sj} = 1$ under the model $\mathcal{M}_0$; by using the proposition 3.1, we have $\nu = \varphi/[1-(1-\varphi)^J]$. We note that the parameters ψ and ν do not take their values in the same space; indeed, $\psi \in]0,1[$ while $\nu \in]1/J, 1[$; the verification is immediate. But, when $J \longrightarrow +\infty$, the three parameters φ, ν and ψ are confounded, since $\nu = \varphi/[1-(1-\varphi)^J] \longrightarrow \varphi$ and ν and ψ take now their values in the same set. The approach of Dorazio and Royle is thus valid asymptotically with respect to J, and it is therefore expected that, for large J, the Bayesian estimation of S under $\mathcal{M}_0$ and $\mathcal{M}_{DR}$ should be close. The question is to know from which value of J, it is effectively the case. This issue is examined in the next Section, using a wide sample of data sets. We now propose another reading of the fact that the DR likelihood and ours coincide when $J \longrightarrow +\infty$. Using the DR model, and thus ignoring the number of quadrats - called J^* for this remark - in which R has been divided, comes down actually to using the model $\mathcal{M}_0$ with $J \longrightarrow \infty$, while in fact $J = J^*$; that clearly entails an error which is quantified in the next Section.

4.3 *A simulation study.*

It is expected that, for some given data $\mathbf{y}$, more or less important differences appear between $\mathrm{E}[S|\mathbf{y},\mathcal{M}_0]$ and $\mathrm{E}[S|\mathbf{y},\mathcal{M}_{DR}]$, as well as between $\sigma[S|\mathbf{y},\mathcal{M}_0]$ and $\sigma[S|\mathbf{y},\mathcal{M}_{DR}]$, where $\sigma[S|\mathbf{y},\mathcal{M}] = \sqrt{\mathrm{var}[S|\mathbf{y},\mathcal{M}]}$. These quantities are afterwards respectively denoted by $\widehat{S}_0$, $\widehat{S}_{DR}$, σ_0 and σ_{DR} (dropping $\mathbf{y}$ in this notation, for convenience). The differences $|\widehat{S}_{DR} - \widehat{S}_0|$ and $|\sigma_0 - \sigma_{DR}|$ are respectively denoted by e_1 and e_2. Our objective is to exhibit some examples for which some significant differences exist, and to examine when they occur (with regard to J, T, K, φ and q). The differences e_1 and e_2, both together, may be globally interpreted as the error (in a statistical sense) one makes concerning S, when one uses the DR model, instead of $\mathcal{M}_0$. Of course, other alternatives could be taken to appreciate the proximity between the posterior distributions of S under $\mathcal{M}_0$ and $\mathcal{M}_{DR}$.

We have chosen to examine the values taken by e_1 and e_2 from average data sets. To clarify what we mean by average data set, recall that the statistic $\mathcal{Y} = (d, V, W)$ is sufficient. J, T, K and θ being fixed, we call an average data set, any data set $\mathbf{y}$ having $\mathcal{Y} = (a, b, c)$ as sufficient statistic, where a is the integer closest to $\mathrm{E}[d|\theta]$, b the integer closest to $\mathrm{E}[V|\theta]$, and c the integer closest to $\mathrm{E}[W|\theta]$, the expectation being taken with respect to the distribution of $\mathbf{y}$ given θ, and under $\mathcal{M}_0$. Choosing average data sets is, in a way, neutral, compared with arbitrary (or simulated) data sets. Finally, when one uses average data sets, it makes sense to compare the errors to each other (when q, φ, J, T, K vary).

Throughout this study S is fixed: $S = 100$. For different values of J, T, K, φ and q, we have calculated the corresponding sufficient statistics $\mathcal{Y}$ by classical Monte Carlo methods (by simulating 1000 data sets similar to $\mathbf{y}$). The resulting average data sets have been numbered from $n = 1$ to 18. Then, for each average data set, we have calculated $\widehat{S}_0$ by implementing the MCMC algorithm Algo2 (based on the observed likelihood L_0).

Concerning $\widehat{S}_{DR}$, we have used the same algorithm (except that L_0 has been replaced by L_{DR}). The convergence of the Markov chain has been diagnosed by using standard techniques. Independent replications of the simulation run for five million iterations (with the first 10% discarded as burn-in) and from different starting points $(\varphi^{(0)}, q^{(0)})$ in $]0,1[\times]0,1[$ produced identical results to one unit place.

We mainly focus on small values for φ, namely in $[0.05, 0.1]$, that is on populations composed of spatially rare species. It is, in fact, for this type of population that we expect to observe significant differences between $\widehat{S}_0$ and $\widehat{S}_{DR}$, since the term $(1-\varphi)^J$ is as less negligible as φ is small (J being fixed). For q, we consider values in $[0.1, 0.3]$, which correspond to species that are relatively hard to detect during a visit. Nevertheless, we emphasize that $q^* = 1-(1-q)^K$ (which represents the probability of detecting a species s during the quadrat sampling experiment, given that it is present in a quadrat j) can be high even if q is small. For example, $q^* = 0.19$ if $q = 0.1$ and $K = 2$, but $q^* = 0.65$ if $q = 0.1$ and $K = 10$. These ranges for φ and q are not unusual in some animal populations, such as birds (see eg Dorazio and Royle, 2005; Dupuis and Joachim, 2006).

For each fixed value of J, T, K, φ and q, we provide in Table 1 below q^*, $\mathcal{Y}$, $\widehat{S}_0$, $\widehat{S}_R$, σ_0, σ_{DR}, e_1 and e_2; non informative prior distributions having been adopted (the Jeffreys prior for S, and the uniform distribution for q, φ and ψ).

• We make two general comments concerning the results. First, we observe that $\widehat{S}_0$ is close to 100 (for most values of K, J, T, φ and q); that is explained by the fact that, for these values of K, J, T, φ and q, the bias of $\widehat{S}_0$ is close to 0; this observation comes from an additional simulation study, not presented in this paper, for brevity. Secondly, we note that $\widehat{S}_{DR}$ is systematically greater than $\widehat{S}_0$. That is simply explained by the fact that using the DR model comes down to using the model $\mathcal{M}_0$ with $J \longrightarrow \infty$.

• Let us briefly comment qualitatively on the different values of the posterior standard

Table 1

q^*, (d, V, W), $\widehat{S}_0$, σ_0, $\widehat{S}_{DR}$, σ_{DR}, e_1, e_2 for different values of J, T, K, q, φ ($S = 100$)

n	J	T	K	φ	q	q^*	d	V	W	$\widehat{S}_0$	σ_0	$\widehat{S}_{DR}$	σ_{DR}	e_1	e_2
1	10	10	4	0.1	0.3	0.76	84	117	184	101	6	154	20	53	14
2	10	10	4	0.3	0.1	0.34	69	107	124	101	10	105	12	4	2
3	10	10	4	0.1	0.1	0.34	45	53	61	101	22	145	48	47	26
4	10	10	10	0.1	0.1	0.65	75	100	154	99	7	152	24	53	17
5	20	10	4	0.1	0.3	0.76	62	86	136	100	11	115	18	15	7
6	20	10	4	0.3	0.1	0.34	66	103	120	99	11	99	11	0	0
7	20	10	4	0.1	0.1	0.34	33	39	45	92	26	105	41	13	15
8	20	10	10	0.1	0.1	0.65	56	74	114	101	13	117	22	16	9
9	20	20	4	0.1	0.1	0.34	57	78	91	98	13	113	20	15	7
10	40	20	4	0.1	0.1	0.34	51	70	81	99	17	101	19	2	2
11	40	40	4	0.1	0.1	0.34	76	139	161	99	7	101	8	2	1
12	40	20	2	0.1	0.1	0.19	32	38	40	104	36	110	43	6	7
13	40	40	4	0.05	0.1	0.34	57	78	91	99	14	116	21	17	7
14	40	20	4	0.05	0.1	0.34	33	39	46	104	29	134	61	30	32
15	80	20	4	0.05	0.1	0.34	30	35	41	101	37	107	48	6	11
16	120	20	4	0.05	0.1	0.34	29	34	40	98	39	101	45	3	6
17	160	20	4	0.05	0.1	0.34	29	34	40	100	41	101	45	1	4
18	200	20	4	0.05	0.1	0.34	29	34	40	100	43	100	44	0	1

deviation σ_0 appearing in Table 1. J being fixed, we note that, not surprisingly, σ_0 decreases with T, K, φ and q (the other factors remaining each time fixed). When T, K, φ and q are fixed, it of interest to observe that σ_0 increases with J, which is not surprising (though that is - maybe - less intuitive than the previous observations); indeed, as J increases - while T remains fixed - the fraction of the unsampled part increases, which creates more and more uncertainty. See the cases: 3 and 7; 9 and 12; 11, 15, 16 and 17.

• We now comment on the magnitude of the errors e_1 and e_2, with respect to J, T, K, φ and q by distinguishing two cases, according to $J \leq 20$ or $J \geq 40$.

Case 1: $J \leq 20$. We observe that, when $q = \varphi = 0.1$, the error e_1 is important if $J = 10$ (whatever T and K) and smaller if $J = 20$. We also note that, T, K, q and φ being fixed, σ_{DR} is markedly larger than σ_0. Moreover, we observe that the estimation of S produced by the DR model may exhibit surprising behaviors. So, e_1 is not modified (even increases) when K increases (cf the cases 3 and 4, as well as the cases 7 and 8); e_1 increases when q increases (cf the cases 1 and 3, as well as the cases 5 and 7); and e_1 increases when T increases (cf the cases 7 and 9). Finally, when $\varphi = 0.3$, we note that the errors e_1 and e_2 are relatively small, even null (see the cases 2 and 6).

Case 2: $J \geq 40$. We note that, when $\varphi = 0.1$ and $K = 4$, the errors e_1 and e_2 are now small, even very small: see the cases 10 and 11. But, if we now consider for φ values smaller than 0.1, as 0.05, the errors may be particular high: see the case 14 (which has to be compared with the case 10). Keeping the value 0.05 for φ and increasing J (cf the cases 15,16,17, 18), we observe that $\widehat{S}_0$ and $\widehat{S}_{DR}$, as well as σ_{DR} and σ_0, practically coincide from about $J = 200$ (cf the case 18). If we now consider species both relatively spatially rare ($\varphi = 0.1$) and hard to detect ($q = 0.1$ and $K = 2$), the values $e_1 = 6$ and $e_2 = 7$ (cf the case 12) are not negligible; the case 12 has to be compared with the case 10 where $e_1 = e_2 = 2$. When $J = 40$, we have not observed the undesirable behaviors of S_{DR} above

mentioned (that is when $J = 10, 20$).

• To sum up, two striking facts emerge from this simulation study. Firstly, the errors e_1 and e_2 due to the use of the DR model (instead of $\mathcal{M}_0$) globally decrease when J increases (as forecasted in Section 4.2); but, the value of J, beyond which $\widehat{S}_{DR}$ and $\widehat{S}_0$, as well as σ_0 and σ_{DR}, will practically coincide (T, K, ϕ and q being fixed) may be rather large. Furthermore, this value depends on T, K, φ, and q (as well as on the true value of S which has been fixed in our study) which prevents from indicating a threshold beyond which the DR approach will yield numerically acceptable results (although Table 1 gives some orders of magnitude). Secondly, we observe that, for small or moderate values of J (≤ 20), the estimations yielded by the DR model may exhibit undesirable behaviors, with respect to K, T and q.

• Our study is limited to homogeneous species populations, but it can reasonably be expected that similar observations will be made in practically homogeneous populations (in the sense that the variability within the φ_s's and the q_s's is small); for heterogeneous populations, the situation is difficult to apprehend intuitively, and additional simulation studies - out of the scope of our paper - will be considered in the future.

5. Conditional and unconditional approaches

We first clarify the links existing between the unconditional approach of Dupuis and Joachim (2003, 2006) and the conditional approach developed in this paper. Then, we briefly comment the unconditional approach of Dorazio *et al.* (2006).

The DJ approach first assumes that a list $\mathcal{L}$ of species liable to be present in R is available. The idea is that in a known region we may have a precise idea of which species are liable to be present in R; this prior information is typically based on previous studies made in R (or encompassing R), or else in regions similar as R. For any species s of the list $\mathcal{L}$ we introduce the indicatrice ξ_s which takes the value 1 if species s is present in the

region R and 0 otherwise. Note that: $S = \sum_{s=1}^{L} \xi_s$ where L denotes the size of $\mathcal{L}$, and that $S \leq L$. The probability mass function (*pmf*) of z_s is as follows:

$$p(z_s|\lambda_s, \varphi_s) = \lambda_s \frac{\varphi_s^{|z_s|}(1-\varphi_s)^{J-|z_s|}}{1-(1-\varphi_s)^J} \mathbb{1}_{(z_s \neq \vec{0})} + (1-\lambda_s)\mathbb{1}_{(z_s=\vec{0})},$$

where λ_s represents the probability that $\xi_s = 1$. It is clear that $p(.)$ effectively defines a *pmf* on $\mathcal{Z}$, that $p(.)$ gives a positive probability to the event $z_s = \vec{0}$ (namley $1 - \lambda_s$), and that

$$p(z_s|\xi_s = 1, \varphi_s) = \frac{\varphi_s^{|z_s|}(1-\varphi_s)^{J-|z_s|}}{1-(1-\varphi_s)^J}.$$

since $z_s \neq \vec{0}$ is equivalent to $\xi_s = 1$. Note that $p(z_s|\xi_s = 1)$ defines a *pmf* on $\mathcal{Z}^*$. We emphasize that it coincides with the *pmf* given in (3.1); in other words, the occurence of any species s present in R is modeled as in the conditional approach. An analogous characteristics, is also present in the unconditional approach developed by Dorazio *et al.* (2006); see below.

Dorazio *et al.* (2006) have also developed an unconditional approach. As mentioned in the introduction, it presents similarities with the DJ approach. In particular, a supercommunity of species is introduced, which is supposed to include the species population of interest located in the region R. Dorazio *et al.* (2006) uses the notion of *maximal species list* to make this idea more concrete. The size of this supercommunity (afterwards denoted by L) is assumed to be known; it has to be fixed so that $L \geq S$, while S is actually unknown (a similar difficulty also exits in the DJ approach). In known regions, one may have some information on the size of S, which facilitates this task. In unkown regions, drawing up the above mentioned *maximal species list* is not faisable, and fixing L is clearly problematic (even illusive). Furthermore, the strategy consisting in choosing a very high value for L does not actually solve the difficulty, since high values of L imply high computational costs, as stressed by Dorazio *et al.* (2006). (A way of overcoming these difficulties

would be probably to put a non informative prior distribution on L.) In short, one can consider that this unconditional approach (as it is) suffers from the same limitation as the DJ approach.

Furthermore, we observe that Dorazio *et al.* (2006) have modeled the occurence of species present in R as in Dorazio and Royle (2005); cf the Section untitled *Modeling heterogeneity in occurence and detection of species*. In particular, the independence assumption between the sampled z_{sj}'s appears, page 845, above the equation (1). Consequently, the unconditional approach developed by these authors suffers from the same limitation as the (conditional) DR approach: that is, it can be used only when J is large.

Recall that M_j represents the number of species present in quadrat j. As for the estimate of M_j calculated by Dorazio and Royle (2005), we have doubts about the validity of the estimate of M_j calculated by Dorazio *et al.* (2006). In the first paper j designates a sampled quadrat, and an unsampled quadrat in the second one. The problem comes from the fact that these authors ignore the status of s (detected or not) and of j (sampled or not), when they indicate the distribution of a missing z_{sj} conditional on the data: concerning Dorazio and Royle (2005), see Section 3.4, page 393, line 1; and concerning Dorazio *et al.* (2006), see page 847, Section untitled *Predicting species accumulation as a function of species occurence*, second paragraph, line 14. More precisely, they consider that $z_{sj}|\mathbf{y} \sim$ Bernoulli (φ) (ψ with their notation) whatever the status of s and of j, while, on the contrary, the distribution of $z_{sj}|\mathbf{y}$ depends on these status (cf the Proposition 3.4). For concision, we do not comment more the unconditional approach of Dorazio *et al.* (2006).

6. Conclusion

In this paper we have developed a new conditional approach to estimate the species richness S of a biological community located in a specified region R divided in J quadrats. It takes into account both the occurence and the detectability of species in the quadrats, as does

the DR approach. But, contrary to the latter, it models the occurence of species in the J quadrats (not only in the sampled quadrats).

We have shown that using the DR model, and therefore ignoring the unsampled part of R, comes down to use the model $\mathcal{M}_0$ with $J \longrightarrow \infty$. From a practical point of view, this means that the DR approach can be used only for large values of J. Concerning the data set analysed by Dorazio and Royle (2005), the use of the DR approach poses of course no problem, since J is particularly large. For small or moderate values of J (≤ 100 to give an order of magnitude) our simulation study shows that the error resulting from the use of the DR model (instead of $\mathcal{M}_0$) may be important; it is especially the case if the population is composed of spatially rare species and/or hard to detect. We note that our conclusions agree, in a way, with those of Mingoti and Meeden (1992) who observed that approaches which ignore J, such as the standard Jacknife estimate, can perform poorly. Hass *et al.* (2006) made similar comments. Our study has thus pointed out some limits of the DR approach. However, it is difficult to provide a general threshold beyond which the DR model could be used (though our study provides some orders of magnitude for some values of T, K and S). Finally, we are of the same opinion as Hass *et al.* (2006), Mingoti and Meeden (1992), when they assert that any statistical analysis should *a priori* include J in the model (except if J is particularly large, as in the the data set analysed by Dorazio and Royle, 2005)

Compared now with the DJ approach, the one developed here has the advantage of applying, to both informative as to non-informative settings, while the DJ approach is strictly limited to situations where a list of species liable to be present in R can be drawn up. In other words, on constrat to the DJ approach, ours applies to both known and unknown regions.

Being able to model the occurence of species in the whole region R allows us to estimate

quantities of biological interest which were out of reach before, such as the number of species present in any subregion of R. In fact, the MCMC algorithm implemented in this paper, which takes advantage of the missing data structure of quadrat sampling data, allows us to tackle more complex problems, such as estimating the number of species shared by two (or more) distinct subregions of R. This problem, said to be difficult, has been solved by Chao *et al.* (2000), but only when data consist of a sample of individuals (not of a sample of quadrats) and when the region R has been divided into two subregions.

The homogeneous model $\mathcal{M}_0$ is a basic one, which has turned out to be a suitable framework for showing the importance of introducing J in species richness modeling. Of course, subsequent work will have to focus on more complex models. From this point of view, the model $\mathcal{M}_0$ can be taken as a starting point. For example, the methodology developed in this paper easily extends to models including heterogeneity at a species and/or spatial level. Other extensions of biological interest are equally possible. Current models assume that species occupy the quadrats independently, a challenging problem would thus be to develop a model for estimating the species richness of an animal population within which predator-prey relationships exist.

Acknowledgements

The author is grateful to Professor Emmanuelle Cam and to Jean Joachim , for helpful discussions.

Appendix A1

Let $z_s(a)$ be a vector extracted from z_s of length J_a. We partition z_s in $z_s(a)$ and $z_s(b)$; hence $p(z_s(a)) = \sum_{z_s(b)} p(z_s(a), z_s(b))$. By observing that this sum is over all the possible values of $z_s(b)$ when $z_s(a) \neq \vec{0}$, and over all the possible values of $z_s(b)$, apart from $\vec{0}$, when

$z_s(a) = \vec{0}$ (due to the constraint $z_s \neq \vec{0}$), it is easy, by using (3.1), to establish that:

$$p(z_s(a)|\varphi) = \frac{\varphi^{|z_s(a)|}(1-\varphi)^{J_a-|z_s(a)|}}{1-(1-\varphi)^J}$$

if $z_s(a) \neq \vec{0}$, and $[(1-\varphi_s)^{J_a}-(1-\varphi)^J]\,/\,[1-(1-\varphi)^J]$ otherwise. Afterwards, the conditioning on φ is omitted, for convenience.

Let $z_s(c)$ be a vector extracted from z_s of length J_c and different from $\vec{0}$. Starting from

$$p[z_s(a)|z_s(c)] = \frac{p(z_s(a), z_s(c))}{p(z_s(c))},$$

and by applying the above result to the vectors $(z_s(a), z_s(c))$ and $z_s(c)$, we have:

$$p[z_s(a), z_s(c)] = \frac{\varphi^{|z_s(a)|+|z_s(c)|}(1-\varphi)^{J_a-|z_s(a)|+J_c-|z_s(c)|}}{1-(1-\varphi)^J}$$

and $p[z_s(c)] = [\varphi^{|z_s(c)|}(1-\varphi)^{J_c-|z_s(c)|}]\,/\,[1-(1-\varphi)^J]$ from which we deduce that:

$$p[z_s(a)|z_s(c) \neq \vec{0}] = \varphi^{|z_s(a)|}(1-\varphi)^{J_a-|z_s(a)|}.$$

Similarly, one has $p[z_s(b)|z_s(c) \neq \vec{0}] = \varphi^{|z_s(b)|}(1-\varphi)^{J_b-|z_s(b)|}$ and

$$p[z_s(a), z_s(b)|z_s(c) \neq \vec{0}] = \varphi^{|z_s(a)|+|z_s(b)|}(1-\varphi)^{J_a-|z_s(a)|+J_b-|z_s(b)|},$$

from which we deduce the result.

Appendix A2

• Due to the assumption A1, the random variables $y_1, \ldots, y_S$ are independent (conditionally on θ). They are also identically distributed, since q and φ do not depend on s. Consequently,

$$(n_1, \ldots, n_h, \ldots, n_H)|S, \beta \sim \text{Multinomial}(S, \beta)$$

where $\beta = (\beta_1, \ldots, \beta_h, \ldots, \beta_H)$ and $\beta_h = \text{pr}(y_s = h|\varphi, q)$; hence (3.2). Afterwards, the conditioning on (φ, q) is omitted, for convenience.

- We first calculate $p(y_s)$, as a function of q and φ, for any $y_s \neq \vec{0}$. 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 $z_s^{mis} = \{z_{sj} | y_{sj} = 0\}$; so, we can write:

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

z_s^{mis} can itself be partitioned into $z_1^{mis}(s)$ and $z_2^{mis}(s)$; $z_1^{mis}(s)$ including the missing z_{sj}'s where j is a sampled quadrat, and $z_2^{mis}(s)$ including those where j is an unsampled quadrat. Hence, dropping afterwards the index s (for simplicity of notation), we have:

$$p(y) = \sum_{z_1^{mis},\ z_2^{mis}} p(x | z^{obs}, z_1^{mis}) \ \ p(z^{obs}, z_1^{mis}, z_2^{mis}) \tag{1}$$

where $z_2^{mis} \in E^{J-T}$ and $z_1^{mis} \in E^{T-|z^{obs}|}$. The normalizing constant which appears in the expression of $p(z)$, namely $[1-(1-\varphi)J]^{-1}$, is afterwards, denoted by c. Taking into account that:

$$p(z^{obs}, z_1^{mis}, z_2^{mis}) = c\, \varphi^{|z^{obs}|+|z_1^{mis}|}(1-\varphi)^{T-|z^{obs}|-|z_1^{mis}|} \times \varphi^{|z_2^{mis}|}(1-\varphi)^{J-T-|z_2^{mis}|}$$

and that

$$p(x | z^{obs}, z_1^{mis}) = \rho\, q^{|x|}(1-q)^{K\,|z^{obs}|-|x|} \times (1-q)^{K\,|z_1^{mis}|}$$

$|x|$ represents the total number of times that the presence of species s has been detected during the experiment, we deduce that:

$$p(y) = c\,\rho \sum_{z_1^{mis}} \varphi^{|z^{obs}|+|z_1^{mis}|}(1-\varphi)^{T-|z^{obs}|-|z_1^{mis}|} q^{|x|}(1-q)^{K\,[|z^{obs}|+|z_1^{mis}|]-|x|}$$

by noting that the double sum (1) over $(z_1^{mis},\ z_2^{mis})$ can be written as the product of two simple sums, one over z_1^{mis} and the other over z_2^{mis}, and that:

$$\sum_{z_2^{mis}} \varphi^{|z_2^{mis}|}(1-\varphi)^{J-T-|z_2^{mis}|} = 1.$$

Hence,

$$p(y) = c\,\rho\,\varphi^{|z^{obs}|}q^{|x|}(1-q)^{K|z^{obs}|-|x|}\sum_{z_1^{mis}}[\varphi(1-q)^K]^{|z_1^{mis}|}\ (1-\varphi)^{T-|z^{obs}|-|z_1^{mis}|}$$

from which we deduce the expression (3.3), by observing that

$$\sum_{z_1^{mis}}[\varphi(1-q)^K]^{|z_1^{mis}|}\ (1-\varphi)^{T-|z^{obs}|-|z_1^{mis}|} = [\varphi(1-q)^K + (1-\varphi)]^{T-|z^{obs}|}$$

since $|z_1^{mis}|$ varies from 0 to $T - |z^{obs}|$.

• We now calculate $p(y_s)$ for a not detected species s, that is β_0. We again start from $p(y_s) = \sum_{z_s^{mis}} p(y_s, z_s^{mis})$. Note that now $z_s^{obs} = \emptyset$ and that $z_s^{mis} = z_s$. We partition the set z_s^{mis} into two parts defined, as before. It is easy to check that:

$$p(y) = c\sum_{z_1^{mis},\ z_2^{mis}}(1-q)^{K|z_1^{mis}|}\varphi^{|z_1^{mis}|}(1-\varphi)^{T-|z_1^{mis}|}\varphi^{|z_2^{mis}|}(1-\varphi)^{J-T-|z_2^{mis}|} \qquad (2)$$

where index s has been dropped (for simplicity of notation). The previous technique used to calculate $p(y)$ leads here to rather tedious calculations because the spaces in which z_1^{mis} and z_2^{mis} takes their values are no longer independent (since z_1^{mis} and z_2^{mis} cannot take together the value $\vec{0}$). To get around that computational difficulty, we calculate the double sum appearing in (2) over $E^T \times E^{J-T}$, from which we remove the value of the term

$$(1-q)^{K|z_1^{mis}|}\varphi^{|z_1^{mis}|}(1-\varphi)^{T-|z_1^{mis}|}\varphi^{|z_2^{mis}|}(1-\varphi)^{J-T-|z_2^{mis}|}$$

evaluated at $z = \vec{0}$. Since $z = \vec{0} \Longleftrightarrow |z| = 0 \Longleftrightarrow |z_1^{mis}| = |z_2^{mis}| = 0$, this value is equal to $(1-\varphi)^J$. The sum over $E^T \times E^{J-T}$ now decomposes as the product of two independent sums:

$$\sum_{z_2^{mis}\in E^{J-T}}\varphi^{|z_2^{mis}|}(1-\varphi)^{J-T-|z_2^{mis}|}$$

which is equal to 1, and

$$\sum_{z_1^{mis}\in E^{T}}(1-q)^{K|z_1^{mis}|}\varphi^{|z_1^{mis}|}(1-\varphi)^{T-|z_1^{mis}|}$$

which is equal to $[\varphi(1-q)^K + (1-\varphi)]^T$. Finally, we obtain:

$$\beta_0 = \frac{[\varphi(1-q)^K + (1-\varphi)]^T - (1-\varphi)^J}{1-(1-\varphi)^J}.$$

Appendix A3

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

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

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

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

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

Appendix A4

Let $\mathbb{N}_2 = \mathbb{N} \setminus \{0, 1\}$. Let $\theta_1 = (S_1, \varphi_1, q_1)$ and $\theta_2 = (S_2, \varphi_2, q_2)$ denote any two values of $\theta \in \Theta = \mathbb{N}_2 \times]0, 1[\times]0, 1[$. We have to prove the following implication:

$$p(.|\theta_1) = p(.|\theta_2) \Longrightarrow \theta_1 = \theta_2,$$

where the hyptohesis $p(.|\theta_1) = p(.|\theta_2)$ means that: $p(\mathbf{y}|\theta_1) = p(\mathbf{y}|\theta_2)$ whatever the data set $\mathbf{y}$. We can thus choose any data set $\mathbf{y}$ (see nevertheless the restriction below); the difficulty being to find suitable and, as far as possible, simple data sets which us allow to prove that

$\theta_1 = \theta_2$. Let us notice, that conditionally on θ_1 and θ_2, a data set $\mathbf{y}$, in which d species have been detected, has to satisfied $d \leq \inf(S_1, S_2)$. This remark concerns only the fourth data set denoted by $\mathbf{y}''$ (see later).

• We first prove that the parameter q is identifiable. For that, we consider two particular data sets. In the first one, called $\mathbf{y}^*$, only one species has been detected, and it has been detected at each visit (therefore $|\mathbf{y}^*| = KT$). In the second one, called $\mathbf{y}^{**}$, again only one species has been detected, and it has been detected at each visit apart from once (therefore $|\mathbf{y}^{**}| = KT - 1$). Using the proposition 3.2, it is easy to check that:

$$p(\mathbf{y}^*|\theta_1) = S_1 \frac{(\varphi_1 q_1^K)^T \ [(\varphi_1(1-q_1)^K + (1-\varphi_1))^T - (1-\varphi_1)^J]^{S_1-1}}{[1-(1-\varphi_1)^J]^{S_1}}$$

and that:

$$p(\mathbf{y}^{**}|\theta_1) = S_1 \frac{K\,\varphi_1^T(1-q_1)q_1^{KT-1}\ \Big[[\varphi_1(1-q_1)^K + 1 - \varphi_1]^T - [1-\varphi_1]^J\Big]^{S_1-1}}{[1-(1-\varphi_1)^J]^{S_1}}.$$

Now, by hypothesis, $p(\mathbf{y}^*|\theta_1) = p(\mathbf{y}^*|\theta_2)$ and $p(\mathbf{y}^{**}|\theta_1) = p(\mathbf{y}^{**}|\theta_2)$; hence

$$\frac{p(\mathbf{y}^*|\theta_1)}{p(\mathbf{y}^{**}|\theta_1)} = \frac{p(\mathbf{y}^*|\theta_2)}{p(\mathbf{y}^{**}|\theta_2)}.$$

After simplification, we obtain:

$$\frac{q_1}{1-q_1} = \frac{q_2}{1-q_2}$$

which implies $q_1 = q_2$.

• We now prove that φ is identifiable. We introduce a third data set, called $\mathbf{y}'$, in which only one species s has been detected and $y_{sj} = K$ for all the quadrats j, except for one quadrat in which species s has not been detected; otherwise $|\mathbf{y}'| = K(T-1)$. Now, using the proposition 3.2, we have:

$$p(\mathbf{y}'|\theta_1) = S_1\ \frac{[\varphi_1 q_1^K]^{T-1}[\varphi_1(1-q_1)^K + 1 - \varphi_1]\Big[[\varphi_1(1-q_1)^K + 1 - \varphi_1]^T - [1-\varphi_1]^J\Big]^{S_1-1}}{[1-(1-\varphi_1)^J]^{S_1}}.$$

After simplification, we obtain:

$$\frac{p(\mathbf{y}^*|\theta_1)}{p(\mathbf{y}'|\theta_1)} = \frac{\varphi_1}{\varphi_1(1-q_1)^K + 1 - \varphi_1}.$$

Using now the hypothesis that $p(\mathbf{y}^*|\theta_1) = p(\mathbf{y}^*|\theta_2)$ and $p(\mathbf{y}'|\theta_1) = p(\mathbf{y}'|\theta_2)$ yields

$$\frac{\varphi_1}{\varphi_1(1-q_1)^K + 1 - \varphi_1} = \frac{\varphi_1}{\varphi_2(1-q_2)^K + 1 - \varphi_2},$$

from which we deduce that $\varphi_1 = \varphi_2$ (since $q_1 = q_2$).

• To prove that S is identifiable, we use a proof by contradiction. We assume that $S_2 \neq S_1$; for example, $S_2 > S_1$. We introduce a new data set, called $\mathbf{y}''$, in which $d = S_1$ species have been detected, and all these species have been detected during each visit. We have:

$$p(\mathbf{y}''|\theta_1) = \frac{[\varphi_1\, q_1^K]^{TS_1}}{[1-(1-\varphi_1)^J]^{S_1}}.$$

and

$$p(\mathbf{y}''|\theta_2) = \left(\begin{array}{c} S_2 \\ S_1 \end{array} \right) \frac{[\varphi_2 q_2^K]^{TS_1} \Big[[\varphi_2(1-q_2)^K + 1 - \varphi_2]^T - [1-\varphi_2]^J \Big]^{S_2-S_1}}{[1-(1-\varphi_2)^J]^{S_2}}$$

where the presence of the terms $[[\varphi_2(1-q_2)^K + 1 - \varphi_2]^T - [1-\varphi_2]^J]^{S_2-S_1}$ and $\left(\begin{array}{c} S_2 \\ S_1 \end{array} \right)$ in $p(\mathbf{y}''|\theta_2)$ is an immediate consequence of our assumption $S_2 > S_1$.

Considering that $q_1 = q_2$ and $\varphi_1 = \varphi_2$, it is easy to verify that $p(\mathbf{y}''|\theta_1) = p(\mathbf{y}''|\theta_2)$ implies that

$$\left(\begin{array}{c} S_2 \\ S_1 \end{array} \right) = \left[\frac{1}{\beta_0} \right]^{S_2-S_1} \tag{1}$$

where

$$\beta_0 = \frac{[\varphi(1-q)^K + 1 - \varphi)]^T - (1-\varphi)^J}{1-(1-\varphi)^J}$$

in which $q = q_1 = q_2$ and $\varphi = \varphi_1 = \varphi_2$.

Moreover, it is straightforward to check that $p(\mathbf{y}^*|\theta_1) = p(\mathbf{y}^*|\theta_2)$ implies that

$$\frac{S_2}{S_1} = \left[\frac{1}{\beta_0} \right]^{S_2-S_1}. \tag{2}$$

From (1) and (2), we deduce that:

$$\left(\begin{array}{c} S_2 \\ S_1 \end{array} \right) = \frac{S_2}{S_1} \tag{3}$$

We set $n_1 = S_1 - 1$ and $n_2 = S_2 - 1$; note that n_1 and n_2 are such that $n_2 > n_1 \geq 1$ since $S_2 > S_1 \geq 2$. It is clear that (3) is equivalent to $\left(\begin{array}{c} S_2 \\ S_1 \end{array} \right) = 1$. Now, $\left(\begin{array}{c} S_2 \\ S_1 \end{array} \right)$ is always > 1 whatever $n_2 > n_1 \geq 1$ (the verification is immediate); hence the contradiction. (Note that $\left(\begin{array}{c} S_2 \\ S_1 \end{array} \right)$ is not > 1 when $n_1 = 0$, hence the condition $S \geq 2$ in the terms of our proposition.) It is clear that starting with $S_2 < S_1$ leads to the same contradiction. Therefore, we conclude that our starting assumption (namely $S_2 \neq S_1$) is false, and that $S_1 = S_2$ is thus true.

Appendix A5

• Let s be a detected species; first note that z_s^{obs} is not empty and is necessarily different from $\vec{0}$. Moreover, given y_s, we can partition z_s into three parts: z_s^{obs}, $z_s^{mis}(1) = \{z_{sj}^{mis} | j \in \mathcal{T}\}$ and $z_s^{mis}(2) = \{z_{sj}^{mis} | j \notin \mathcal{T}\}$ where $\mathcal{T} = \{1, \ldots, T\}$. Due to proposition 3.1, $z_s^{mis}(1)$ and $z_s^{mis}(2)$ are independent conditionally on (y_s, θ) (since $z_s^{obs} \neq \vec{0}$). By using now the first part of Assumption A1, we deduce that the two blocks $\mathbf{z}_m^{[1]}$ and $\mathbf{z}_m^{[2]}$ are independent conditionally on $(\mathbf{y}, \theta)$. Using the same arguments, we deduce that the z_{sj}'s of $\mathbf{z}_m^{[2]}$ are independent conditionally on $(\mathbf{y}, \theta)$. Thus $V_m^{[2]} | \mathbf{y}, \varphi, q \sim \text{Bin}((J - T)d, \varphi)$ as the sum of $(J - T)d$ independent Bernoulli (φ) r.v.s.

• Using again the same arguments, we deduce that the missing z_{sj}'s of $\mathbf{z}_m^{[1]}$ are independent conditionally on $(\mathbf{y}, \theta)$. We now calculate the probability $\gamma = \text{pr}(z_{sj} = 1 | y_s, \varphi, q)$, where y_s is such that $y_s \neq \vec{0}$ and $y_{sj} = 0$ since s represents a species which has not been detected in quadrat j. By appling the Bayes formula, we have:

$$\text{pr}(z_{sj} = 1 | y_s) = \frac{p(y_s | z_{sj} = 1) \ \text{pr}(z_{sj} = 1)}{p(y_s | z_{sj} = 1) \ \text{pr}(z_{sj} = 1) + p(y_s | z_{sj} = 0) \ \text{pr}(z_{sj} = 0)}$$

where φ and q have been omitted in the conditionings (for convenience). Now, as mentioned at the end of Section 3.3, we have: $\text{pr}(z_{sj}=1)=\varphi/[1-(1-\varphi]^J]$. Moreover, it is easy to verify that: $p(y_s|z_{sj}=1)=(1-q)^T p(y_s|z_{sj}=0)$ (recall that $y_s\neq\vec{0}$). Hence, the expression of γ given in proposition 3.5. Moreover, $V_m^{[1]}|\mathbf{y},\varphi,q\sim\text{Bin}(dT-V,\gamma)$ as the sum of $dT-V$ independent Bernoulli (γ) r.v.s.

- Let s be a species that is not detected. We start from:

$$p(z_s|y_s=\vec{0})=\frac{\text{pr}(y_s=\vec{0}|z_s)p(z_s)}{\text{pr}(y_s=\vec{0})}.$$

where φ and q have been omitted in the conditionings (for convenience). If $z_s=\vec{0}$ then $\text{pr}(y_s=\vec{0}|z_s=\vec{0})=1$ else $\text{pr}(y_s=\vec{0}|z_s)=(1-q)^{K|z_s^*|}$. Note that this formula holds when $z_s=\vec{0}$. By replacing $p(z_s)$ by its expression (cf 3.1), and $\text{pr}(y_s=\vec{0})=\beta_0$ by its own (cf 3.4), we obtain the result. estimates of the parameters of the model (in particular S) and those of any related quantities to S (as N_j).

Appendix A6

The Argument 1 still holds when ψ depends on j, due to the following remark which is parameter-free. Whatever j, one has $\text{pr}(z_{sj}=1|\overline{z}_{s,j}=\vec{0})=1$ (due to the constraint $\overline{z}_{s,j}\neq\vec{0}$), and $\text{pr}(z_{sj}=1|\overline{z}_{s,j}\neq\vec{0})=\varphi$. Consequently, the distribution of $z_{sj}|\overline{z}_{s,j}$ depends on $\overline{z}_{s,j}$, which implies that the J r.v. z_{sj}'s cannot be independent.

We now examine the Argument 2, and we set $\text{pr}(z_{sj}=1)=\psi_j$. For any $i\in\{1,\ldots,J\}$ we introduce the vector $z_s^{[i]}=(z_{sj}^{[i]};j=1,\ldots,J)$ where $z_{sj}^{[i]}=1$ if $j=i$ and zero otherwise. i being fixed, we assume that one wishes to infer on S from the sample $\mathcal{E}_i$ which includes all the quadrats, except the quadrat i; therefore, $\mathcal{E}_i$ includes $T=J-1$ quadrats. Omitting, for convenience, the conditionings on the ψ_j's, we can write:

$$p(z_s^{[i]})=\text{pr}(z_{si}=1|A_i)\times\text{pr}(A_i)$$

where A_i denotes the event: $z_{sj} = 0$ for all $j \neq i$. Moreover, due to the constraint $z_s \neq \vec{0}$, we have $\text{pr}(z_{si} = 1|A_i) = 1$; due to the DR assumption of independence, we have $\text{pr}(A_i) = \prod_{j \neq i}(1 - \psi_j)$, from which we deduce that: $p(z_s^{[i]}) = \prod_{j \neq i}(1 - \psi_j)$. We now assume that $T = J$; due to DR assumption of independence, we have $p(z_s^{[i]}) = \psi_i \prod_{j \neq i}(1 - \psi_j)$ which differs from the value obtained when $T = J - 1$.

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., Hwang,W-H, Chen, Y-C, and Kuo C-Y (2000) Estimating the number of shared species in two communities. *Statistica Sinica* **10**, 227-246.

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 rive 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 occurence of species. *Journal of the American Statistical Association* **100**, 389-398.

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

Dupuis, J. A. and Joachim, J. (2003) Bayesian estimation of species richness. Publication No LSP-2003-10.

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.

George, I.E. and Robert, C.P. (1992) Capture-recapture estimation via Gibbs sampling. *Biometrika* **79**, 677-683.

Hass, P.J., Liu, Y., and Stokes, L. (2006) An estimator of number of species richness 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.

Joachim, J., Clouet, M., Bousquet, J.F., and Faure, C. (1990). Peuplements d'oiseaux nicheurs des forêts pyrénéennes centrales; comparaison du peuplement de différentes essences. *Acta Biologica Montana* **10**, 135-157.

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. (2001) On the Bayesian estimation of population size. *Biometrika* **88**, 841-851.

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.

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

Robert, C. and Casella, G. (2004) *Monte-Carlo Statistical Methods.* 2nd edition. Springer-Verlag, New York.