Bayesian estimation of abundance based on removal sampling under weak assumption of closed population with catchability depending on environmental conditions. Application to tick abundance

S. Bord [a],[*], P. Druilhet [b], P. Gasqui [a], D. Abrial [a], G. Vourc'h [a]

[a] INRA, UR346 Unité Epidémiologie Animale, Centre INRA de Theix, F-63122 Saint Genès Champanelle, France
[b] Laboratoire de mathématiques, UMR CNRS 6620 Campus des Cézeaux, B.P. 80026, F-63171 Aubière cedex, France

ARTICLE INFO

ABSTRACT

The estimation of animal abundance is essential to understand population dynamics, species interactions and disease patterns in populations. Estimations of relative abundance classically are based on a single observation of several sites. In this case, the mapping of abundance assumes that the probability of detecting an individual, hence the sampling rate, remains constant across the observed sites. In practice, however, this assumption is often not satisfied as the sampling rate may fluctuate between sites due to random fluctuations and/or fluctuations associated with the sampling process, notably associated with the characteristics of the site. It is therefore important to account for variations in detection probability. Using a removal sampling design, we studied the performance of a Bayesian approach to estimate both sampling rates and abundance under the assumption of a closed population. The assumption of a closed population often is weakened when the number of successive samplings is large. The number of samplings has to be limited and optimal. We therefore examined the minimal number of successive samplings needed to achieve sufficient statistical accuracy while respecting underlying model assumptions. Using the same simulations, we also compared the performance of the Bayesian approach to the performance of the frequentist Hayne method based on linear regression. We show that the Bayesian approach proposed gives generally better estimations of population size than the Hayne method. The two methods give approximately the same results for the estimation of sampling rate. We then studied the variability of detection probability of *Ixodes ricinus* ticks sampled under several environmental conditions by using a hierarchical Bayesian model with a random effect. The estimated sampling rate $\hat{\tau}_c$ varied between 33.9% and 47.4% for shrubs and 53.6% and 66.7% for dead leaves. The variability of the sampling rate due to the site decreased when the number of successive samplings considered in the model increased. The variability was lower in dead leaves than shrubs. This approach could be used routinely for ecological or epidemiological studies of ticks and species with comparable life histories.

© 2013 Elsevier B.V. All rights reserved.

1. Introduction

The estimation of animal abundance is essential in ecology to understand fundamental processes, such as population dynamics and species interactions, as well as in epidemiology to understand and generate disease patterns in populations (Anderson, 1991). In the majority of biological systems, relevant indicators of abundance are based on count point surveys (Alldredge et al., 2007) obtained using convenient and calibrated sampling methods (Anderson, 2001; Pollock et al., 2002). As a part of the population is often not observable, the probability of detecting an individual and the sampling rate are both less than 1 (Kery, 2004; Kery and

Royle, 2010; Pellet and Schmidt, 2005). Consequently, indicators calculated in this way only give an index of relative abundance. These indicators of abundance are implicitly based on the assumption that the sampling rate is constant from site to site (Williams et al., 2002; Pollock et al., 2002; Royle and Dorazio, 2006). However, the sampling rate may depend on environmental conditions, such as weather, season, sampler and habitats. If this is the case, considering the sampling rate to be constant leads to confusion between the variability of the rate and the variability of abundance (Thompson et al., 1998; MacKenzie and Kendall, 2002). Therefore, an effort needs to be made to estimate both the abundance and the sampling rate. The two main sampling methods used to estimate both abundance and sampling rate are Capture Mark Recapture (CMR) and removal sampling (RS) (see e.g. Williams et al., 2002; Thompson et al., 1998). CMR involves capturing a random sample of individuals in a population of interest, marking them, and then

releasing them back into the population. Another sample of individual from the same population is subsequently recaptured (Williams et al., 2002). The ratio of unmarked to marked individuals is used to estimate capture probability and abundance. CMR is generally labour intensive (Pollock et al., 2002). Moreover, the capture, marking and release steps could have an effect on the behaviour of the marked population (Dodd and Dorazio, 2004), especially in some species. RS consists of capturing individuals on several occasions over a short period of time. The captured individuals are removed from the population (Williams et al., 2002). The total number of individuals captured over the successive samplings is used to estimate capture probability and abundance. Like CMR, RS is labour intensive (Dodd and Dorazio, 2004) and may modify the observed site when the number of successive samplings is high. The choice of a given protocol (CMR or RS) and its ease of implementation depend on the species studied. Although available, CMR and RS methods are rarely used for certain species. One such species are ticks, which are the most important vectors of human and animal diseases after mosquitoes (Parola and Raoult, 2001). The classical index of tick abundance used is estimated by the number of tick nymphs by dragging a piece of cloth once over the vegetation of a delimited area, generally $10\,m^2$ (Vassallo et al., 2000) in a selection of sites. Host-seeking nymphs, *i.e.* those waiting for a host on the top of the vegetation, are collected by the drag. The numbers of nymphs collected on the different sites are then compared. The drag method is distinguished from RS methods in that the cloth is dragged only once over each site. To our knowledge, the sampling rate of the drag sampling method has been studied little. Only one study (Talleklint-Eisen and Lane, 2000) has used a RS design to estimate the abundance and the sampling rate of the drag method. In this study, 17 successive samplings were conducted over 23 days. This protocol could not guarantee that the population remained closed over the sampling period. The authors estimated the sampling rate to be 5.9% using the Hayne method (Hayne, 1949).

To estimate parameters, the Hayne method makes a linear regression of the number of successive captures on the cumulative number of captures. The sampling rate and the population size are estimated respectively as the slope of the regression line and as the intersection point between the horizontal axis and regression line. However, the Hayne method is known to produce poor estimations of population size (White et al., 1982), especially when the sampling rate is potentially low (less than 10%) and variable. Moreover, this method does not allow covariate effects to be taken into account, nor does it provide confidence intervals for estimates.

In this paper, a hierarchical Bayesian approach was used to estimate both the sampling rate and the population size. This approach allows the inclusion of prior knowledge and provides posterior distributions of parameter estimates. Moreover, because it is hierarchical, one can take into account parameters which are either observable or not observable, and which are located at different scales such as the area sampled and the site of sampling (Gelman and Hill, 2006; Cressie et al., 2009). First, we studied the performance of a Bayesian approach using simulations to determine the minimum number of successive samplings needed to achieve sufficient statistical accuracy for estimates while respecting an underlying model assumption of a closed population. We then compared the performance of the Bayesian approach to the Hayne method. Finally, our Bayesian approach was applied to RS tick data collected using the drag method on several sites in September 2011. In a first hierarchical Bayesian model (HBM), the sampling rate was assumed to be specific to each site. In a second HBM, the sampling rate was assumed to be the same for similar conditions with small uncontrolled spatial variations due to the characteristics of the site observed.

2. Materials and methods

2.1. Removal sampling data structure

The statistical unit considered in this study was the sampling site. A sampling site was defined as a delimited area on which successive samplings were performed. For a given sampling site, we considered a closed population of an unknown size N_0, *i.e.* with no immigration, emigration, birth or death during the successive samplings. We denoted by $X = (X_1, \ldots, X_K)$ the sequence of captures observed on a given site where K was the total number of successive samplings carried out. Let X_k be the number of captures at the kth sampling and N_k the remaining population after the kth sampling where $N_k = N_{k-1} - X_k$ for $k \in 1, \ldots, K$. Furthermore, each individual of the closed population was assumed to be captured independently with the same probability of capture τ (Moran, 1951; Zippin, 1956). Hence, we assumed that X_k followed a binomial distribution with population size N_{k-1} and probability of capture τ (Eq. (1)):

$$(X_k|N_{k-1}, \tau) \sim \mathcal{B}(N_{k-1}, \tau), \quad \text{where} \quad N_k = N_{k-1} - X_k. \tag{1}$$

The capture probability was considered as independent and the same for all individuals, so we considered that it was equal to the sampling rate, *i.e.* the percent of captured individuals in the population.

2.2. Hierarchical Bayesian models

A first HBM (HBM1) assumed that the population size N_{0s} and the sampling rate τ_s was specific to each site s. A second HBM (HBM2) assumed that the sampling rate of a given site τ_s was associated to both the effect of sampling conditions c (considered to be a fixed effect) and a small variation due to the site sampled (considered to be a random effect). The logit transformation of sampling rate τ_s denoted logit(τ_s) was used. The logit(τ_s) was decomposed as the sum of the logit transformation of the sampling rate τ_c under the sampling conditions c and a random effect ϵ_s (Eq. (2)):

$$\text{logit}(\tau_s) = \text{logit}(\tau_c) + \epsilon_s. \tag{2}$$

The random effect ϵ_s was introduced to add a fluctuation of the sampling rate τ_s due to the observed site s. The range of variation of the random effect ϵ_s, denoted by σ_c^2, was considered to be specific to each sampling condition c where ϵ_s was assumed to follow a normal distribution with zero-mean and variance σ_c^2 which depended on the sampling conditions c. The HBM1 and HBM2 models described above are summarised in Figs. 1 and 2 by Directed Acyclic Graphs (DAGs) (Thulasiraman and Swamy, 1992; Clark and Gelfand, 2006). These DAGs represent relationships (lines) between the observed data and the unknown parameters or hypotheses of the model (nodes). The lines represent the relations and the hierarchy between nodes. The nodes symbolise the data observed X_k, parameters to estimate N_{0s}, τ_s, τ_c, prior distribution of parameters to estimate for N_{0s}, τ_s, τ_c and ϵ_s and the distribution of hyperparameter σ_c^2. If $\epsilon_s = 0$ and τ_c was specific to each site s, then the DAG corresponded to the HBM1 model. HBM1 and HBM2 were implemented in OpenBUGS (version 3.2.1.). Source codes are available in the supplementary material.

The prior distribution is a key part of Bayesian inference and represents information about an uncertain parameter. In the absence of accurate information, a uniform prior distribution over the interval [0 ; 1000] was assigned to N_{0s}. For HBM1, a uniform prior distribution over the interval [0 ; 1] was assigned to τ_s. For HBM2, a uniform prior distribution over the interval [0 ; 1] was assigned to τ_c and a uniform prior distribution over the interval [0 ; 3] was assigned to σ_c^2. The prior distribution was then combined with the probability distribution of data to obtain the posterior distribution of parameters. Parameters were estimated by the mean and