Focused information criterion for capture-recapture models for closed populations

Francesco Bartolucci*† and Monia Lupparelli*‡

November 16, 2007

Abstract

We propose a criterion for selecting a capture-recapture model for closed populations which follows the basic idea of the Focused Information Criterion (FIC) of Claeskens and Hjort (2003). The proposed criterion aims at selecting the model which, among the available models, leads to the smallest Mean Squared Error (MSE) of the resulting estimator of the population size and is based on an index which, up to a constant term, is equal to the asymptotic MSE of the estimator. Two alternative approaches to estimate this FIC index are proposed. We also deal with multimodel inference; in this case, the population size is estimated by using a weighted average of the estimates coming from different models, with weights chosen as to minimize the MSE of the resulting estimator. The proposed model selection approach is compared with more common approaches through a series of simulations performed along the same lines as Stanley and Burnham (1998). It is also illustrated by an application based on a dataset coming from a live-trapping experiment.

KEY WORDS: AIC, CAIC, conditional maximum likelihood estimation, model selection, multimodel inference.

*Dipartimento di Economia, Finanza e Statistica, Università di Perugia, 06123 Perugia, Italy.

†*email:* bart@stat.unipg.it

‡*email:* lupparelli@ds.unifi.it

1 Introduction

In many contexts, the estimation of the size of a certain population, such as that of animals living in a certain region or that of people suffering from a certain disease, is carried out on the basis of capture-recapture data; for a review see Yip *et al.* (1995a,b), Schwarz and Seber (1999), Borchers *et al.* (2004) and Amstrup *et al.* (2005), among others. These data are typically collected by means of a series of trapping experiments or by matching the information contained in two or more administrative lists and consist of sequences of binary outcomes that, for each subject observed at least once, indicate if the subject has been or not "captured" at a given occasion. A wide variety of models is now available for the analysis of data like these. It ranges from very simple models, such as that assuming that capture probabilities are equal across subjects and occasions, to very sophisticated models which take into account all the factors that may affect these probabilities, i.e. time, behavior and heterogeneity (Otis *et al.*, 1978). Since the estimate of the population size crucially depends on the adopted model, model selection is a fundamental issue in the capture-recapture context.

As in many other fields, the Akaike Information Criterion (AIC, Akaike, 1973) is one of the most used criteria for model selection also in this field. This criterion is usually preferred for its simple use and nice interpretation in terms of distance between the *true model*, i.e. the data generating model, and the assumed model. The quality of the inference on the population size, when the model is selected with AIC, was studied by simulation by Burnham *et al.* (1995) and Stanley and Burnham (1998). It results that this criterion usually leads to selecting a model that, in the class of available models, is a good compromise between the largest model, which guarantees a small bias of the resulting estimator of the population size, and the smallest model, which guarantees a small variance of this estimator. These authors also considered other selection criteria, such as AIC_c and CAIC, which are versions of AIC based on a penalization term taking the sample size also into account. These selection criteria perform better than AIC in certain circumstances. Finally, they considered a form of multimodel inference in which the population size is estimated by a weighted average of

the estimates coming from different models, with weights which are simply computed on the basis of the AIC, AIC_c or CAIC index; for a general description of this type of inference see Buckland *et al.* (1997) and Burnham and Anderson (2002, Ch. 4). It turns out that the average estimator is usually more efficient than the corresponding single-model estimator.

An important point to underline is that the AIC, as well as the other selection criteria mentioned above, are aimed at finding the model which, among the available models, provides the best approximation of the true model. However, it is not ensured that such a model also guarantees the smallest estimation error of the population size. In fact, for the same data, we can usually find two or more models which have very close values of the AIC index, but lead to very different estimates of the population size. To this regard, the examples provided by Agresti (1994) and Fienberg *et al.* (1999) are illuminating. For instance, among the models considered by Fienberg *et al.* (1999, Sec. 5.2) for the analysis of a capture-recapture dataset concerning 2,069 cases of diabetes in an Italian region, there are two models for which the difference in terms of AIC index is around 0.6, but one leads to an estimate of the population size equal to 2,381, whereas the other model leads to an estimate of 7,796.

Recently, Claeskens and Hjort (2003) proposed a model selection strategy to be used when the main interest is in estimating a certain parameter, rather than on representing adequately the data generating mechanism. This criterion, known as Focused Information Criterion (FIC), is based on an index that, up to a constant term, is equal to the asymptotic Mean Squared Error (MSE) of the estimator of the parameter of interest under the model to which it is referred. This index is developed by exploiting certain asymptotic results of Hjort and Claeskens (2003) concerning the properties of maximum likelihood estimators in the presence of misspecification. These results, however, are not directly applicable to the capture-recapture context, in particular when the conditional maximum likelihood (CML) approach of Sanathanan (1972) is used to estimate a *closed-population* model, i.e. a model assuming that the population size is constant during the period of observation (Borchers *et al.*, 2004). Note that the inferential framework dealt with in this paper is different from the traditional one, since, in our context, the sample size is a random quantity, whereas the (unknown) population size is a fixed quantity. The asymptotic theory is then developed for

the latter tending to infinity. This is the main reason why the results of Hjort and Claeskens (2003) are not directly applicable to our context.

In this paper, we extend the main results of Hjort and Claeskens (2003) to the case in which a closed-population model is estimated with the CML approach and derive a FIC index for the resulting estimator. We also introduce two estimators of this index, the first of which recalls that proposed by Claeskens and Hjort (2003), whereas the second follows a new idea. We also deal with the estimation of the population size on the basis of a weighted average of the estimates coming from different models, with weights chosen by minimizing a function which provides a measure of error of the resulting estimator of the population size. This criterion for choosing the weights follows a different conception with respect to that used in connection with AIC and similar selection criteria and recalls that exploited by Hansen (2007) in a different context. The approach proposed here is illustrated by a series of simulations performed along the same lines as Stanley and Burnham (1998).

The paper is organized as follows. In Section 2 we review the CML approach for the estimation of a closed-population model and we recall the main asymptotic properties of this approach when the assumed model holds. In Section 3 we study the same properties when the model is misspecified and then we derive an expression for the asymptotic MSE of the estimator of the population size and a FIC index for this estimator and its weighted average version. On the basis of these results, in Section 4 we deal with model selection and multimodel inference on the population size. The approach is illustrated by a simulation study and an application involving a dataset coming from a live-trapping experiment, the results of which are reported in Sections 5 and 6. Finally, in Section 7 we draw the main conclusions and we outline some possible extensions of the proposed approach.

2 Conditional inference for closed-population models

Let N denote the unknown population size, let J denote the number of capture occasions and let $p(\boldsymbol{r})$ denote the probability of the capture configuration $\boldsymbol{r} = (\, r_1 \quad \cdots \quad r_J \,)$, where r_j is a binary variable equal to 1 if a subject is captured at the j-th occasion and to 0

otherwise. Also let $m = 1 - p(\mathbf{0})$ be the probability of being captured at least once and let $q(\boldsymbol{r}) = p(\boldsymbol{r})/m$ be the conditional probability of the capture configuration $\boldsymbol{r}$ for a subject captured at least once. Finally, for every $\boldsymbol{r} \neq \mathbf{0}$, $y(\boldsymbol{r})$ denotes the frequency of the subjects with capture configuration $\boldsymbol{r}$ so that $n = \sum_{\boldsymbol{r}\neq\mathbf{0}} y(\boldsymbol{r})$ is the sample size. We have to stress that n is a random variable whose distribution depends on the capture probabilities. On the other hand, N is a fixed quantity and the asymptotic theory that follows is developed for this quantity tending to infinity, not for n tending to infinity as in standard inferential contexts.

To estimate N on the basis of the frequencies $y(\boldsymbol{r})$, we must assume a model on the probabilities $p(\boldsymbol{r})$. In this paper, we suppose that a collection $\mathcal{M}$ of models is available. The largest model in this class, referred to as *full model*, assumes that $p(\boldsymbol{r}) = f(\boldsymbol{r}; \boldsymbol{\beta}, \boldsymbol{\gamma})$ for a suitable function $f(\cdot)$ and where $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are two vectors of parameters to be estimated. The smallest model in class $\mathcal{M}$, referred to as *null model*, is a particular case of the full model in which $\boldsymbol{\gamma} = \boldsymbol{\gamma}_0$, where $\boldsymbol{\gamma}_0$ is a fixed parameter vector. Between the full and the null models, there exist models in which only certain elements of $\boldsymbol{\gamma}$ are fixed at the corresponding elements of $\boldsymbol{\gamma}_0$ and the other ones have to be estimated, i.e. $\boldsymbol{\gamma} = (\boldsymbol{\gamma}'_{\mathcal{S}}, \boldsymbol{\gamma}'_{\bar{\mathcal{S}},0})'$ for a subset $\mathcal{S}$ of $\mathcal{F} = \{1, \ldots, c\}$, with c denoting the size of $\boldsymbol{\gamma}$ and $\bar{\mathcal{S}}$ its complement. Each model in $\mathcal{M}$ may consequently be denoted by $M_{\mathcal{S}}$ and its parameter vector by $\boldsymbol{\theta}_{\mathcal{S}} = (\boldsymbol{\beta}', \boldsymbol{\gamma}'_{\mathcal{S}})'$, whereas $p_{\mathcal{S}}(\boldsymbol{r})$, $q_{\mathcal{S}}(\boldsymbol{r})$ and $m_{\mathcal{S}}$ denote the corresponding probability functions defined as above. It has to be clear that these are functions of $\boldsymbol{\theta}_{\mathcal{S}}$. Finally, note that $M_{\mathcal{F}}$ obviously corresponds to the full model, whereas $M_{\emptyset}$ corresponds to the null model.

In the following, we show how the parameters of each model $M_{\mathcal{S}}$, and consequently N, may be estimated by the CML approach. We also discuss model selection.

2.1 Estimation of the model parameters and population size

In the CML approach, the parameter vector $\boldsymbol{\theta}_{\mathcal{S}}$ of model $M_{\mathcal{S}}$ is estimated by maximizing the *conditional log-likelihood* of the observed capture configurations given the sample size n. This function may be expressed as

$$\ell(\boldsymbol{\theta}_{\mathcal{S}}) = \boldsymbol{y}' \log(\boldsymbol{q}_{\mathcal{S}}),$$

where $\boldsymbol{q}_{\mathcal{S}}$ denotes the column vector with elements $q_{\mathcal{S}}(\boldsymbol{r})$ for every $\boldsymbol{r} \neq \boldsymbol{0}$ and $\boldsymbol{y}$ denotes the corresponding vector with elements $y(\boldsymbol{r})$. Both vectors have $2^J - 1$ elements.

The Expectation-Maximization (EM) algorithm (Dempster *et al.*, 1977) is normally used to maximize $\ell(\boldsymbol{\theta}_{\mathcal{S}})$. It consists of alternating the following steps until convergence:

- *E-step*: given the current estimate of $\boldsymbol{\theta}_{\mathcal{S}}$, denoted by $\tilde{\boldsymbol{\theta}}_{\mathcal{S}}$, compute the expected value of the number of subjects which have not been captured as

$$\tilde{y}(\boldsymbol{0}) = n\frac{1-\tilde{m}_{\mathcal{S}}}{\tilde{m}_{\mathcal{S}}};$$

- *M-step*: update the estimate of $\boldsymbol{\theta}_{\mathcal{S}}$ by maximizing the conditional expected value of the *complete log-likelihood* given $\boldsymbol{y}$ and $\tilde{\boldsymbol{\theta}}_{\mathcal{S}}$. This log-likelihood may be expressed as

$$\tilde{\ell}^*(\boldsymbol{\theta}_{\mathcal{S}}; \tilde{\boldsymbol{\theta}}_{\mathcal{S}}) = \tilde{y}(\boldsymbol{0}) \log(1 - m_{\mathcal{S}}) + \boldsymbol{y}' \log(\boldsymbol{p}_{\mathcal{S}}),$$

  where $\boldsymbol{p}_{\mathcal{S}}$ denotes the vectors with elements $p_{\mathcal{S}}(\boldsymbol{r})$ for every $\boldsymbol{r} \neq \boldsymbol{0}$. Typically, $\tilde{\ell}^*(\boldsymbol{\theta}_{\mathcal{S}}; \tilde{\boldsymbol{\theta}}_{\mathcal{S}})$ is maximized by means of a Newton-Raphson algorithm based on the following first derivative vector and second derivative matrix

$$\begin{aligned}
\frac{\partial \tilde{\ell}^*(\boldsymbol{\theta}_{\mathcal{S}}; \tilde{\boldsymbol{\theta}}_{\mathcal{S}})}{\partial \boldsymbol{\theta}_{\mathcal{S}}} &= -\frac{\tilde{y}(\boldsymbol{0})}{1 - m_{\mathcal{S}}} \frac{\partial m_{\mathcal{S}}}{\partial \boldsymbol{\theta}_{\mathcal{S}}} + \frac{\partial \boldsymbol{p}'_{\mathcal{S}}}{\partial \boldsymbol{\theta}_{\mathcal{S}}} \mathrm{diag}(\boldsymbol{p}_{\mathcal{S}})^{-1} \boldsymbol{y}, \\
\frac{\partial^2 \tilde{\ell}^*(\boldsymbol{\theta}_{\mathcal{S}}; \tilde{\boldsymbol{\theta}}_{\mathcal{S}})}{\partial \boldsymbol{\theta}_{\mathcal{S}} \partial \boldsymbol{\theta}'_{\mathcal{S}}} &= -\frac{\tilde{y}(\boldsymbol{0})}{(1 - m_{\mathcal{S}})^2} \frac{\partial m_{\mathcal{S}}}{\partial \boldsymbol{\theta}_{\mathcal{S}}} \frac{\partial m_{\mathcal{S}}}{\partial \boldsymbol{\theta}'_{\mathcal{S}}} - \frac{\tilde{y}(\boldsymbol{0})}{1 - m_{\mathcal{S}}} \frac{\partial^2 m_{\mathcal{S}}}{\partial \boldsymbol{\theta}_{\mathcal{S}} \partial \boldsymbol{\theta}'_{\mathcal{S}}} + \\
&\quad -\frac{\partial \boldsymbol{p}'_{\mathcal{S}}}{\partial \boldsymbol{\theta}_{\mathcal{S}}} \mathrm{diag}(\boldsymbol{y}) \mathrm{diag}(\boldsymbol{p}_{\mathcal{S}})^{-2} \frac{\partial \boldsymbol{p}_{\mathcal{S}}}{\partial \boldsymbol{\theta}'_{\mathcal{S}}} + \sum_{\boldsymbol{r} \neq \boldsymbol{0}} \frac{y(\boldsymbol{r})}{p_{\mathcal{S}}(\boldsymbol{r})} \frac{\partial^2 p_{\mathcal{S}}(\boldsymbol{r})}{\partial \boldsymbol{\theta}_{\mathcal{S}} \partial \boldsymbol{\theta}'_{\mathcal{S}}}.
\end{aligned}$$

The value of $\boldsymbol{\theta}_{\mathcal{S}}$ at convergence of the EM algorithm, which we take as the CML estimate of $\boldsymbol{\theta}_{\mathcal{S}}$, is denoted by $\hat{\boldsymbol{\theta}}_{\mathcal{S}}$. We then estimate the population size under model $M_{\mathcal{S}}$ as $\hat{N}_{\mathcal{S}} = n/\hat{m}_{\mathcal{S}}$.

2.2 Asymptotic properties under the assumed model

First of all consider that the score vector for model $M_{\mathcal{S}}$ may be expressed as

$$\boldsymbol{s}(\boldsymbol{\theta}_{\mathcal{S}}) = \frac{\partial \ell(\boldsymbol{\theta}_{\mathcal{S}})}{\partial \boldsymbol{\theta}_{\mathcal{S}}} = \frac{\partial \boldsymbol{q}'_{\mathcal{S}}}{\partial \boldsymbol{\theta}_{\mathcal{S}}} \mathrm{diag}(\boldsymbol{q}_{\mathcal{S}})^{-1} \boldsymbol{y}.$$

Since $\boldsymbol{y}$ has variance equal to $N\boldsymbol{\Omega}_{\mathcal{S}}$, with $\boldsymbol{\Omega}_{\mathcal{S}} = \text{diag}(\boldsymbol{p}_{\mathcal{S}}) - \boldsymbol{p}_{\mathcal{S}}\boldsymbol{p}'_{\mathcal{S}}$, the *Fisher information matrix*, i.e. the variance-covariance matrix of $\boldsymbol{s}(\boldsymbol{\theta}_{\mathcal{S}})$, may be expressed as

$$\boldsymbol{F}(\boldsymbol{\theta}_{\mathcal{S}}) = N\frac{\partial \boldsymbol{q}'_{\mathcal{S}}}{\partial \boldsymbol{\theta}_{\mathcal{S}}}\text{diag}(\boldsymbol{q}_{\mathcal{S}})^{-1}\boldsymbol{\Omega}_{\mathcal{S}}\text{diag}(\boldsymbol{q}_{\mathcal{S}})^{-1}\frac{\partial \boldsymbol{q}_{\mathcal{S}}}{\partial \boldsymbol{\theta}'_{\mathcal{S}}} = m_{\mathcal{S}}N\frac{\partial \boldsymbol{q}'_{\mathcal{S}}}{\partial \boldsymbol{\theta}_{\mathcal{S}}}\text{diag}(\boldsymbol{q}_{\mathcal{S}})^{-1}\frac{\partial \boldsymbol{q}_{\mathcal{S}}}{\partial \boldsymbol{\theta}'_{\mathcal{S}}}.$$

This matrix, evaluated at $\boldsymbol{\theta}_{\mathcal{S},0} = (\boldsymbol{\beta}'_0, \boldsymbol{\gamma}'_{\mathcal{S},0})'$, the true value of the parameter vector $\boldsymbol{\theta}_{\mathcal{S}}$, is denoted by $\boldsymbol{F}_{\mathcal{S},0}$ and $\boldsymbol{J}_{\mathcal{S},0} = \boldsymbol{F}_{\mathcal{S},0}/N$ is the corresponding average information matrix.

Now consider the following assumptions that in our approach need to hold for each model $M_{\mathcal{S}}$ in class $\mathcal{M}$:

A1. for every capture configuration $\boldsymbol{r}$, $p_{\mathcal{S}}(\boldsymbol{r})$ is strictly positive and admits continuous first-order derivatives at every admissible value of $\boldsymbol{\theta}_{\mathcal{S}}$;

A2. for any $\omega > 0$, it is possible to find an $\varepsilon > 0$ such that

$$\inf_{|\boldsymbol{\theta}_{\mathcal{S}} - \boldsymbol{\theta}_{\mathcal{S},0}| > \omega} \sum_{\boldsymbol{r} \neq 0} q_{\mathcal{S},0}(\boldsymbol{r}) \log \frac{q_{\mathcal{S},0}(\boldsymbol{r})}{q_{\mathcal{S}}(\boldsymbol{r})} > \varepsilon,$$

with $q_{\mathcal{S},0}(\boldsymbol{r})$ denoting the probability $q_{\mathcal{S}}(\boldsymbol{r})$ evaluated at $\boldsymbol{\theta}_{\mathcal{S},0}$;

A3. the Jacobian of $\begin{pmatrix} p_{\mathcal{S}}(\boldsymbol{0}) & \boldsymbol{p}'_{\mathcal{S}} \end{pmatrix}$ with respect to $\boldsymbol{\theta}_{\mathcal{S}}$ is of full rank.

Note that A1 is a standard regularity condition on the function $p_{\mathcal{S}}(\boldsymbol{r})$, A2 is a *strong identifiability* condition (see also Rao, 1965, Sec. 5e.2) expressed directly on the conditional probabilities $q_{\mathcal{S}}(\boldsymbol{r})$ and A3 implies that the Fisher information matrix at $\boldsymbol{\theta}_{\mathcal{S},0}$, $\boldsymbol{F}_{\mathcal{S},0}$, is of full rank. It is worth noting that, if the conditions above hold for the full model in class $\mathcal{M}$, then they hold for each submodel $M_{\mathcal{S}}$ in the same class.

Under assumptions A1, A2 and A3, Theorem 2 of Sanathanan (1972) implies that:

- $\hat{\boldsymbol{\theta}}_{\mathcal{S}} \xrightarrow{p} \boldsymbol{\theta}_{\mathcal{S},0}$ and $\sqrt{N}(\hat{\boldsymbol{\theta}}_{\mathcal{S}} - \boldsymbol{\theta}_{\mathcal{S},0}) \xrightarrow{d} \mathcal{N}(\boldsymbol{0}, \boldsymbol{J}^{-1}_{\mathcal{S},0})$
- $\hat{N}_{\mathcal{S}}/N \xrightarrow{p} 1$ and $(\hat{N}_{\mathcal{S}} - N)/\sqrt{N} \xrightarrow{d} \mathcal{N}(0, \eta^2_{\mathcal{S}})$

for each model $M_{\mathcal{S}}$, with

$$\eta^2_{\mathcal{S}} = \frac{1 - m_{\mathcal{S},0}}{m_{\mathcal{S},0}} + \frac{1}{m^2_{\mathcal{S},0}}\frac{\partial m(\boldsymbol{\theta}_{\mathcal{S},0})}{\partial \boldsymbol{\theta}'_{\mathcal{S}}}\boldsymbol{J}^{-1}_{\mathcal{S},0}\frac{\partial m(\boldsymbol{\theta}_{\mathcal{S},0})}{\partial \boldsymbol{\theta}_{\mathcal{S}}}. \tag{1}$$

We recall that $\xrightarrow{p}$ means convergence in probability and $\xrightarrow{d}$ means convergence in distribution as $N \to \infty$. Moreover, in the above expression, we use

$$\frac{\partial m(\boldsymbol{\theta}_{\mathcal{S},0})}{\partial \boldsymbol{\theta}_{\mathcal{S}}}$$

to denote the derivative of $m_{\mathcal{S}}$ with respect to $\boldsymbol{\theta}_{\mathcal{S}}$ evaluated at $\boldsymbol{\theta}_{\mathcal{S},0}$. A similar convention is used to denote a derivative of a function of $\boldsymbol{\theta}_{\mathcal{S}}$ evaluated at $\hat{\boldsymbol{\theta}}_{\mathcal{S}}$.

The asymptotic results above imply that we can compute the standard errors for $\hat{\boldsymbol{\theta}}_{\mathcal{S}}$ and $\hat{N}_{\mathcal{S}}$ from an estimate of $\boldsymbol{J}_{\mathcal{S},0}$. To this aim consider the following assumption:

A4. for every capture configuration $\boldsymbol{r}$, $p_{\mathcal{S}}(\boldsymbol{r})$ admits continuous second-order derivatives at every admissible value of $\boldsymbol{\theta}_{\mathcal{S}}$.

Under A4, $\boldsymbol{F}(\boldsymbol{\theta}_{\mathcal{S}})$ is equal to the (unconditional) expected value of

$$\boldsymbol{H}(\boldsymbol{\theta}_{\mathcal{S}}) = -\frac{\partial^2 \ell(\boldsymbol{\theta}_{\mathcal{S}})}{\partial \boldsymbol{\theta}_{\mathcal{S}} \partial \boldsymbol{\theta}'_{\mathcal{S}}} = \frac{\partial \boldsymbol{q}'_{\mathcal{S}}}{\partial \boldsymbol{\theta}_{\mathcal{S}}} \text{diag}(\boldsymbol{y}) \text{diag}(\boldsymbol{q}_{\mathcal{S}})^{-2} \frac{\partial \boldsymbol{q}_{\mathcal{S}}}{\partial \boldsymbol{\theta}'_{\mathcal{S}}} - \sum_{\boldsymbol{r}} \frac{y(\boldsymbol{r})}{q_{\mathcal{S}}(\boldsymbol{r})} \frac{\partial q_{\mathcal{S}}(\boldsymbol{r})}{\partial \boldsymbol{\theta}_{\mathcal{S}} \partial \boldsymbol{\theta}'_{\mathcal{S}}},$$

which is the observed information matrix. We also have that $\|\boldsymbol{H}(\hat{\boldsymbol{\theta}}_{\mathcal{S}})/N - \boldsymbol{J}_{\mathcal{S},0}\| \xrightarrow{p} 0$ and then $\boldsymbol{J}_{\mathcal{S},0}$ may be consistently estimated by

$$\hat{\boldsymbol{J}}_{\mathcal{S}} = \boldsymbol{H}(\hat{\boldsymbol{\theta}}_{\mathcal{S}})/\hat{N}_{\mathcal{S}}. \tag{2}$$

The standard error for $\hat{\boldsymbol{\theta}}_{\mathcal{S}}$ directly follows from the inverse of $\boldsymbol{H}(\hat{\boldsymbol{\theta}}_{\mathcal{S}})$; because of (1), the standard error for $\hat{N}_{\mathcal{S}}$ is given by

$$\sqrt{\hat{N}_{\mathcal{S}} \frac{1 - \hat{m}_{\mathcal{S}}}{\hat{m}_{\mathcal{S}}} + \frac{\hat{N}_{\mathcal{S}}^2}{\hat{m}_{\mathcal{S}}^2} \frac{\partial m(\hat{\boldsymbol{\theta}}_{\mathcal{S}})}{\partial \boldsymbol{\theta}'_{\mathcal{S}}} \boldsymbol{H}(\hat{\boldsymbol{\theta}}_{\mathcal{S}})^{-1} \frac{\partial m(\hat{\boldsymbol{\theta}}_{\mathcal{S}})}{\partial \boldsymbol{\theta}_{\mathcal{S}}}}.$$

2.3 Information criteria for model selection

One of the most used criteria for selecting a model for the analysis of a capture-recapture dataset is the AIC (Akaike, 1973). It is based on an index which provides a measure of the distance between the assumed model $M_{\mathcal{S}}$ and the true model and is defined as

$$AIC_{\mathcal{S}} = -2\ell(\hat{\boldsymbol{\theta}}_{\mathcal{S}}) + 2k_{\mathcal{S}},$$

where $k_{\mathcal{S}}$ is the number of the model parameters. Consequently, among the available models, we prefer the one with the smallest value of this index.

Other criteria for model selection are available in the literature which may be seen as versions of the above one; for a review in the capture-recapture context see Burnham *et al.* (1995) and Stanley and Burnham (1998). Among these criteria, it is worth mentioning those referred to as AIC_c and CAIC which are based on the indices

$$\begin{aligned} AIC_{c,\mathcal{S}} &= AIC_{\mathcal{S}} + \frac{2k_{\mathcal{S}}(k_{\mathcal{S}}+1)}{n-k_{\mathcal{S}}-1}, \\ CAIC_{\mathcal{S}} &= -2\ell(\hat{\boldsymbol{\theta}}_{\mathcal{S}}) + k_{\mathcal{S}}[\log(n)+1], \end{aligned}$$

which make use of a penalization term taking the sample size into account.

In the capture-recapture context, the model selected according to one of the above criteria is used to obtain an estimate of the population size (*single-model inference*). A different approach is that based on model averaging (*multimodel inference*) which consists of estimating the population by

$$\hat{N}_{\boldsymbol{w}} = \sum_{\mathcal{S}} w_{\mathcal{S}} \hat{N}_{\mathcal{S}}, \tag{3}$$

where $\sum_{\mathcal{S}}$ denotes the sum over all the models in $\mathcal{M}$ and

$$w_{\mathcal{S}} = \frac{\exp(-AIC_{\mathcal{S}}/2)}{\sum_{\mathcal{S}'} \exp(-AIC_{\mathcal{S}'}/2)} \tag{4}$$

is the AIC-weight for model $M_{\mathcal{S}}$. A similar estimator may be based on the other selection criteria mentioned above. In particular, the formula for computing AIC_c and CAIC-weights is the same as (4), with $AIC_{\mathcal{S}}$ substituted by $AIC_{c,\mathcal{S}}$ and $CAIC_{\mathcal{S}}$, respectively. For a general discussion on this type of inference see Buckland *et al.* (1997).

3 Asymptotic properties under the true model

Similarly to Claeskens and Hjort (2003) and Hjort and Claeskens (2003), we now assume that the full model holds with $\boldsymbol{\beta} = \boldsymbol{\beta}_0$ and $\boldsymbol{\gamma} = \boldsymbol{\gamma}_0 + \boldsymbol{\delta}/\sqrt{N}$, where $\boldsymbol{\delta}$ denotes a fixed vector of suitable dimension. This model is referred to as *true model* and the corresponding probability of the capture configuration $\boldsymbol{r}$ is denoted by $p_{\text{true}}(\boldsymbol{r}) = f(\boldsymbol{r}; \boldsymbol{\beta}_0, \boldsymbol{\gamma}_0 + \boldsymbol{\delta}/\sqrt{N})$. These probabilities are collected in the vector $\boldsymbol{p}_{\text{true}}$ for every $\boldsymbol{r} \neq \boldsymbol{0}$. We also denote by $p_0(\boldsymbol{r}) = f(\boldsymbol{r}; \boldsymbol{\beta}_0, \boldsymbol{\gamma}_0)$ the probability of $\boldsymbol{r}$ when $\boldsymbol{\beta} = \boldsymbol{\beta}_0$ and $\boldsymbol{\gamma} = \boldsymbol{\gamma}_0$ and by $\boldsymbol{p}_0$ the corresponding probability vector. Similarly, we use $q_0(\boldsymbol{r})$ to denote the conditional probability

$q(\boldsymbol{r})$ under this model, $\boldsymbol{q}_0$ to denote the corresponding probability vector and m_0 to denote the probability of being captured at least once.

Within this framework, we first study the asymptotic properties of the CML estimator of $\boldsymbol{\theta}_{\mathcal{S}}$, $\hat{\boldsymbol{\theta}}_{\mathcal{S}}$, and those of the corresponding estimator of N, $\hat{N}_{\mathcal{S}}$. We then consider the asymptotic properties of the estimator $\hat{N}_{\boldsymbol{w}}$.

3.1 Model parameters estimator

Let $\boldsymbol{f} = \boldsymbol{y}/N$ be the vector of the relative frequencies of all capture configurations $\boldsymbol{r} \neq \boldsymbol{0}$ and consider the following assumption:

A5. for every capture configuration $\boldsymbol{r}$,

$$p_{\text{true}}(\boldsymbol{r}) = p_0(\boldsymbol{r}) + \frac{\partial p_0(\boldsymbol{r})}{\partial \boldsymbol{\gamma}'}\boldsymbol{\delta}/\sqrt{N} + O(N^{-1}).$$

Under this assumption, the Central Limit Theorem implies that

$$\sqrt{N}(\boldsymbol{f} - \boldsymbol{p}_0) \xrightarrow{d} \mathcal{N}\left(\frac{\partial \boldsymbol{p}_0}{\partial \boldsymbol{\gamma}'}\boldsymbol{\delta}, \boldsymbol{\Omega}_0\right), \tag{5}$$

as $N \to \infty$, where $\boldsymbol{\Omega}_0 = \text{diag}(\boldsymbol{p}_0) - \boldsymbol{p}_0\boldsymbol{p}_0'$. On the basis of this result we prove Theorem 1 below, where we use the following decomposition of the average information matrix under the full model

$$\boldsymbol{J}_{\mathcal{F},0} = \begin{pmatrix} \boldsymbol{J}_{00} & \boldsymbol{J}_{01} \\ \boldsymbol{J}_{01} & \boldsymbol{J}_{11} \end{pmatrix}$$

and $\boldsymbol{P}_{\mathcal{S}}$ denotes the block of rows of an identity matrix of size c such that $\boldsymbol{\gamma}_{\mathcal{S}} = \boldsymbol{P}_{\mathcal{S}}\boldsymbol{\gamma}$. We also recall that convergence in probability and distribution are for $N \to \infty$.

Theorem 1 *Under the true model and provided that assumptions A1, A2, A3 and A5 hold, for each model* $M_{\mathcal{S}}$

- $\hat{\boldsymbol{\theta}}_{\mathcal{S}} \xrightarrow{p} \boldsymbol{\theta}_{\mathcal{S},0}$
- $\sqrt{N}(\hat{\boldsymbol{\theta}}_{\mathcal{S}} - \boldsymbol{\theta}_{\mathcal{S},0}) \xrightarrow{d} \mathcal{N}\left[\boldsymbol{J}_{\mathcal{S},0}^{-1}\begin{pmatrix} \boldsymbol{J}_{01} \\ \boldsymbol{P}_{\mathcal{S}}\boldsymbol{J}_{11} \end{pmatrix}\boldsymbol{\delta}, \boldsymbol{J}_{\mathcal{S},0}^{-1}\right].$

Proof. See Appendix A1.

3.2 Population size estimator

Let

$$T_{\mathcal{S}} = \frac{\hat{N}_{\mathcal{S}} - N}{\sqrt{N}}$$

and note that this quantity may also be expressed as

$$\sqrt{N}\left(\frac{\mathbf{1}'\boldsymbol{f}}{\hat{m}_{\mathcal{S}}} - 1\right). \tag{6}$$

The following Theorem holds.

Theorem 2 *Under the true model and provided that assumptions A1, A2, A3 and A5 hold, for each model* $M_{\mathcal{S}}$

- $\hat{N}_{\mathcal{S}}/N \xrightarrow{p} 1$
- $T_{\mathcal{S}} \xrightarrow{d} \mathcal{N}(\mu_{\mathcal{S}}, \sigma^2_{\mathcal{S}})$,

where

$$\mu_{\mathcal{S}} = \frac{1}{m_0}\boldsymbol{b}'_{\mathcal{S}}\boldsymbol{\delta} \qquad \textit{and} \qquad \sigma^2_{\mathcal{S}} = \frac{1-m_0}{m_0} + \frac{1}{m_0^2}\frac{\partial m(\boldsymbol{\theta}_{\mathcal{S},0})}{\partial \boldsymbol{\theta}'_{\mathcal{S}}}\boldsymbol{J}^{-1}_{\mathcal{S},0}\frac{\partial m(\boldsymbol{\theta}_{\mathcal{S},0})}{\partial \boldsymbol{\theta}_{\mathcal{S}}},$$

with

$$\boldsymbol{b}_{\mathcal{S}} = \left[\frac{\partial m_0}{\partial \boldsymbol{\gamma}'} - \frac{\partial m(\boldsymbol{\theta}_{\mathcal{S},0})}{\partial \boldsymbol{\theta}'_{\mathcal{S}}}\boldsymbol{J}^{-1}_{\mathcal{S},0}\begin{pmatrix}\boldsymbol{J}_{01}\\ \boldsymbol{P}_{\mathcal{S}}\boldsymbol{J}_{11}\end{pmatrix}\right]'. \tag{7}$$

Proof. See Appendix A1.

This Theorem implies that the asymptotic MSE of $T_{\mathcal{S}}$ may be expressed as

$$MSE(T_{\mathcal{S}}) = \mu^2_{\mathcal{S}} + \sigma^2_{\mathcal{S}} = \frac{1-m_0}{m_0} + FIC(T_{\mathcal{S}}),$$

with

$$FIC(T_{\mathcal{S}}) = \frac{1}{m_0^2}\boldsymbol{b}'_{\mathcal{S}}\boldsymbol{\delta}\boldsymbol{\delta}'\boldsymbol{b}_{\mathcal{S}} + \frac{1}{m_0^2}\frac{\partial m(\boldsymbol{\theta}_{\mathcal{S},0})}{\partial \boldsymbol{\theta}'_{\mathcal{S}}}\boldsymbol{J}^{-1}_{\mathcal{S},0}\frac{\partial m(\boldsymbol{\theta}_{\mathcal{S},0})}{\partial \boldsymbol{\theta}_{\mathcal{S}}}. \tag{8}$$

It is worth noting that the asymptotic MSE of $T_{\mathcal{S}}$ is equal to the index $FIC(T_{\mathcal{S}})$ plus a term which is constant with respect to $\mathcal{S}$. Therefore, as suggested by Claeskens and Hjort (2003), it is reasonable to choose the model with the smallest value of this index, once it has been properly estimated (see Section 4). Also note that the above formula for $FIC(T_{\mathcal{S}})$ does not closely resemble the general formula provided by Claeskens and Hjort (2003). The

latter, however, is based on certain simplifications which we do not consider since they do not affect the model selection strategy and are not so relevant in the present context.

Now consider the average estimator $\hat{N}_{\boldsymbol{w}}$ defined in (3) in the case in which the weights $w_{\mathcal{S}}$ are a priori fixed. Let

$$T_{\boldsymbol{w}} = \frac{\hat{N}_{\boldsymbol{w}} - N}{\sqrt{N}}$$

which may also be expressed as

$$\sqrt{N}\left(\sum_{\mathcal{S}} w_{\mathcal{S}} \frac{\mathbf{1}'\boldsymbol{f}}{\hat{m}_{\mathcal{S}}} - 1\right).$$

The following Theorem holds, where $\boldsymbol{w}$ denotes a column vector with elements $w_{\mathcal{S}}$ for every model $M_{\mathcal{S}}$.

Theorem 3 *Under the true model and provided that assumptions A1, A2, A3 and A5 hold,*

- $\hat{N}_{\boldsymbol{w}}/N \xrightarrow{p} 1$
- $T_{\boldsymbol{w}} \xrightarrow{d} \mathcal{N}(\mu_{\boldsymbol{w}}, \sigma^2_{\boldsymbol{w}})$,

with

$$\mu_{\boldsymbol{w}} = \frac{1}{m_0}\boldsymbol{w}'\boldsymbol{B}\boldsymbol{\delta} \qquad \text{and} \qquad \sigma^2_{\boldsymbol{w}} = \frac{1-m_0}{m_0} + \frac{1}{m_0^2}\boldsymbol{w}'\boldsymbol{C}\boldsymbol{J}_{\mathcal{F},0}\boldsymbol{C}'\boldsymbol{w},$$

where $\boldsymbol{B}$ is a matrix with rows $\boldsymbol{b}'_{\mathcal{S}}$ and $\boldsymbol{C}$ is a matrix with rows $\boldsymbol{c}'_{\mathcal{S}}$ for every model $M_{\mathcal{S}}$, with $\boldsymbol{b}'_{\mathcal{S}}$ defined in (7) and

$$\boldsymbol{c}_{\mathcal{S}} = \left[\frac{\partial m(\boldsymbol{\theta}_{\mathcal{S},0})}{\partial \boldsymbol{\theta}'_{\mathcal{S}}} \boldsymbol{J}^{-1}_{\mathcal{S},0} \begin{pmatrix} \boldsymbol{I} & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{P}_{\mathcal{S}} \end{pmatrix}\right]'.$$

Proof. See Appendix A1.

From the above expression, we find that the asymptotic MSE of $T_{\boldsymbol{w}}$ may be expressed as

$$MSE(T_{\boldsymbol{w}}) = \mu^2_{\boldsymbol{w}} + \sigma^2_{\boldsymbol{w}} = \frac{1-m_0}{m_0} + FIC(T_{\boldsymbol{w}}),$$

with

$$FIC(T_{\boldsymbol{w}}) = \frac{1}{m_0^2}\boldsymbol{w}'\boldsymbol{B}\boldsymbol{\delta}\boldsymbol{\delta}'\boldsymbol{B}'\boldsymbol{w} + \frac{1}{m_0^2}\boldsymbol{w}'\boldsymbol{C}\boldsymbol{J}_{\mathcal{F},0}\boldsymbol{C}'\boldsymbol{w}. \tag{9}$$

In the following, we show how to estimate $FIC(T_{\boldsymbol{w}})$ and suggest an optimal procedure for choosing $\boldsymbol{w}$. This procedure is based on a different idea with respect to the standard procedure used in connection with AIC and similar selection criteria.

4 Proposed model selection strategy

We now introduce an estimator of $FIC(T_{\mathcal{S}})$ which follows the basic idea of that of Claeskens and Hjort (2003). We then discuss an alternative estimator of the same quantity based on a different approach. Finally, we deal with estimation of $FIC(T_{\boldsymbol{w}})$ and multimodel inference on the population size.

4.1 Single-model inference

First of all note that, if A4 holds, we can consistently estimate all the quantities involved in (8), with the exception of the matrix $\boldsymbol{\delta}\boldsymbol{\delta}'$. Let

$$\hat{\boldsymbol{m}}_{\mathcal{F}} = \frac{\partial m(\hat{\boldsymbol{\theta}}_{\mathcal{F}})}{\partial \boldsymbol{\theta}_{\mathcal{F}}}$$

and partition this vector into the subvectors $\hat{\boldsymbol{m}}_1$, referred to the parameters $\boldsymbol{\beta}$, and $\hat{\boldsymbol{m}}_2$, referred to the parameters $\boldsymbol{\gamma}$. We then have that

$$\hat{\boldsymbol{m}}_{\mathcal{S}} = \begin{pmatrix} \hat{\boldsymbol{m}}_1 \\ \\ \boldsymbol{P}_{\mathcal{S}}\hat{\boldsymbol{m}}_2 \end{pmatrix}$$

is a consistent estimator of the derivative of $m(\boldsymbol{\theta}_{\mathcal{S},0})$ with respect to $\boldsymbol{\theta}_{\mathcal{S}}$ and

$$\hat{\boldsymbol{b}}_{\mathcal{S}} = \left[\hat{\boldsymbol{m}}_2' - \hat{\boldsymbol{m}}_{\mathcal{S}}'\tilde{\boldsymbol{J}}_{\mathcal{S}}^{-1}\begin{pmatrix} \hat{\boldsymbol{J}}_{01} \\ \\ \boldsymbol{P}_{\mathcal{S}}\hat{\boldsymbol{J}}_{11} \end{pmatrix}\right]' \tag{10}$$

is a consistent estimator of $\boldsymbol{b}_{\mathcal{S}}$, where $\hat{\boldsymbol{J}}_{01}$ and $\hat{\boldsymbol{J}}_{11}$ are suitable blocks of $\hat{\boldsymbol{J}}_{\mathcal{F}}$, with the latter defined according to (2). Also $\boldsymbol{J}_{\mathcal{S},0}$ may be consistently estimated by using a suitable block of $\hat{\boldsymbol{J}}_{\mathcal{F}}$, denoted in this case by $\tilde{\boldsymbol{J}}_{\mathcal{S}}$.

To derive an estimator of $\boldsymbol{\delta}\boldsymbol{\delta}'$, we rely on the Lemma below which directly follows from Theorem 1 considering that $\hat{N}_{\mathcal{F}}/N \xrightarrow{p} 1$ as $N \to \infty$.

Lemma 1 *Under the true model and provided that assumptions A1, A2, A3 and A5 hold,*

$$\sqrt{\hat{N}_{\mathcal{F}}}(\hat{\boldsymbol{\gamma}}_{\mathcal{F}} - \boldsymbol{\gamma}_0) \xrightarrow{d} \mathcal{N}(\boldsymbol{\delta}, \boldsymbol{K}),$$

with $\boldsymbol{K}$ denoting the block of $\boldsymbol{J}_{\mathcal{F},0}^{-1}$ corresponding to $\boldsymbol{J}_{11}$.

This implies that an asymptotically unbiased estimator of $\boldsymbol{\delta}\boldsymbol{\delta}'$ is

$$\hat{N}_{\mathcal{F}}(\hat{\boldsymbol{\gamma}}_{\mathcal{F}} - \boldsymbol{\gamma}_0)(\hat{\boldsymbol{\gamma}}_{\mathcal{F}} - \boldsymbol{\gamma}_0)' - \hat{\boldsymbol{K}},$$

with $\hat{\boldsymbol{K}}$ taken from $\hat{\boldsymbol{J}}_{\mathcal{F}}^{-1}$. By substituting the above expression into (8) and recalling the definition of $\hat{N}_{\mathcal{F}}$, we obtain an estimator of $FIC(T_{\mathcal{S}})$ which is given by

$$\widehat{FIC}(T_{\mathcal{S}}) = \frac{n}{\hat{m}_{\mathcal{F}}^3}\hat{\boldsymbol{b}}_{\mathcal{S}}'(\hat{\boldsymbol{\gamma}}_{\mathcal{F}} - \boldsymbol{\gamma}_0)(\hat{\boldsymbol{\gamma}}_{\mathcal{F}} - \boldsymbol{\gamma}_0)'\hat{\boldsymbol{b}}_{\mathcal{S}} + \frac{1}{\hat{m}_{\mathcal{F}}^2}\hat{\boldsymbol{m}}_{\mathcal{S}}'\tilde{\boldsymbol{J}}_{\mathcal{S}}^{-1}\hat{\boldsymbol{m}}_{\mathcal{S}} - \frac{1}{\hat{m}_{\mathcal{F}}^2}\hat{\boldsymbol{b}}_{\mathcal{S}}'\hat{\boldsymbol{K}}\hat{\boldsymbol{b}}_{\mathcal{S}}.$$

Note that, in the formula above, the quantities m_0, $\boldsymbol{b}_{\mathcal{S}}$ and $\boldsymbol{J}_{\mathcal{S},0}$ are estimated by using the CML estimator $\hat{\boldsymbol{\theta}}_{\mathcal{F}}$ under the full model, but they could also be consistently estimated on the basis of the CML estimator $\hat{\boldsymbol{\theta}}_{\emptyset}$ under the null model (see Claeskens and Hjort, 2003).

An alternative estimator of $FIC(T_{\mathcal{S}})$, which has a nice interpretation, may be built as follows. Let

$$D_{\mathcal{S}} = \frac{\hat{N}_{\mathcal{S}} - \hat{N}_{\mathcal{F}}}{\sqrt{\hat{N}_{\mathcal{F}}}}$$

and consider the following Theorem.

Theorem 4 *Under the true model and provided that assumptions A1, A2, A3 and A5 hold,* $D_{\mathcal{S}} \xrightarrow{d} \mathcal{N}(\nu_{\mathcal{S}}, \tau_{\mathcal{S}}^2)$*, with*

$$\nu_{\mathcal{S}} = \frac{1}{m_0}\boldsymbol{b}_{\mathcal{S}}'\boldsymbol{\delta} \quad \textit{and} \quad \tau_{\mathcal{S}}^2 = \frac{1}{m_0^2}\left[\frac{\partial m(\boldsymbol{\theta}_{\mathcal{F},0})}{\partial \boldsymbol{\theta}_{\mathcal{F}}'}\boldsymbol{J}_{\mathcal{F},0}^{-1}\frac{\partial m(\boldsymbol{\theta}_{\mathcal{F},0})}{\partial \boldsymbol{\theta}_{\mathcal{F}}} - \frac{\partial m(\boldsymbol{\theta}_{\mathcal{S},0})}{\partial \boldsymbol{\theta}_{\mathcal{S}}'}\boldsymbol{J}_{\mathcal{S},0}^{-1}\frac{\partial m(\boldsymbol{\theta}_{\mathcal{S},0})}{\partial \boldsymbol{\theta}_{\mathcal{S}}}\right].$$

Proof. See Appendix A1.

Since $\nu_{\mathcal{S}} = \mu_{\mathcal{S}}$, we propose to estimate $FIC(T_{\mathcal{S}})$ with

$$\widehat{FIC}^*(T_{\mathcal{S}}) = D_{\mathcal{S}}^2 - \hat{\tau}_{\mathcal{S}}^2 + \tilde{\sigma}_{\mathcal{S}}^2 = \frac{(\hat{N}_{\mathcal{S}} - \hat{N}_{\mathcal{F}})^2}{\hat{N}_{\mathcal{F}}} - \frac{1}{\hat{m}_{\mathcal{F}}^2}\hat{\boldsymbol{m}}_{\mathcal{F}}'\hat{\boldsymbol{J}}_{\mathcal{F}}^{-1}\hat{\boldsymbol{m}}_{\mathcal{F}} + \frac{2}{\hat{m}_{\mathcal{F}}^2}\hat{\boldsymbol{m}}_{\mathcal{S}}'\tilde{\boldsymbol{J}}_{\mathcal{S}}^{-1}\hat{\boldsymbol{m}}_{\mathcal{S}},$$

where by $\tilde{\sigma}_{\mathcal{S}}^2$ we mean an estimate of $\sigma_{\mathcal{S}}^2 - (1 - m_0)/m_0$.

Regardless of the estimator used for $FIC(T_{\mathcal{S}})$, we suggest to select the model with the smallest estimate of this index and then taking the corresponding estimate of N. In the following, this selection criterion will be indicated by FIC when it is based on the estimator $\widehat{FIC}(T_{\mathcal{S}})$ and by FIC* when it is based on the alternative estimator $\widehat{FIC}^*(T_{\mathcal{S}})$. It may obviously happen that FIC and FIC* do not lead to choosing the same model.

4.2 Multimodel inference

Given (9), we derive a first estimator of $FIC(T_{\boldsymbol{w}})$ which closely recalls the structure of the estimator $\widehat{FIC}(T_{\mathcal{S}})$ derived above. This estimator has the following structure

$$\widehat{FIC}(T_{\boldsymbol{w}}) = \frac{n}{\hat{m}_{\mathcal{F}}^3}\boldsymbol{w}'\hat{\boldsymbol{B}}(\hat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}_0)(\hat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}_0)'\hat{\boldsymbol{B}}'\boldsymbol{w} - \frac{1}{\hat{m}_{\mathcal{F}}^2}\boldsymbol{w}'\hat{\boldsymbol{B}}\hat{\boldsymbol{K}}\hat{\boldsymbol{B}}'\boldsymbol{w} + \frac{1}{\hat{m}_{\mathcal{F}}^2}\boldsymbol{w}'\hat{\boldsymbol{C}}\hat{\boldsymbol{J}}_{\mathcal{F}}\hat{\boldsymbol{C}}'\boldsymbol{w}, \quad (11)$$

where $\hat{\boldsymbol{B}}$ is made of rows $\hat{\boldsymbol{b}}_{\mathcal{S}}'$ defined in (10) and $\hat{\boldsymbol{C}}$ is a matrix with rows $\hat{\boldsymbol{c}}_{\mathcal{S}}'$, with

$$\hat{\boldsymbol{c}}_{\mathcal{S}} = \left[\hat{\boldsymbol{m}}_{\mathcal{S}}'\tilde{\boldsymbol{J}}_{\mathcal{S}}^{-1}\begin{pmatrix}\boldsymbol{I} & \boldsymbol{O}\\ \boldsymbol{O} & \boldsymbol{P}_{\mathcal{S}}\end{pmatrix}\right]'.$$

We can use an alternative estimator based on the asymptotic properties of

$$D_{\boldsymbol{w}} = \frac{\hat{N}_{\boldsymbol{w}} - \hat{N}_{\mathcal{F}}}{\sqrt{\hat{N}_{\mathcal{F}}}} = \sum_{\mathcal{S}} w_{\mathcal{S}} D_{\mathcal{S}}.$$

Theorem 5 *Under the true model and provided that assumptions A1, A2, A3 and A5 hold,* $D_{\boldsymbol{w}} \xrightarrow{d} \mathcal{N}(\nu_{\boldsymbol{w}}, \tau_{\boldsymbol{w}}^2)$*, with*

$$\nu_{\boldsymbol{w}} = \frac{1}{m_0}\boldsymbol{w}'\boldsymbol{B}\boldsymbol{\delta} \quad \textit{and} \quad \tau_{\boldsymbol{w}}^2 = \frac{1}{m_0^2}\boldsymbol{w}'\boldsymbol{E}\boldsymbol{J}_{\mathcal{F},0}\boldsymbol{E}'\boldsymbol{w},$$

where $\boldsymbol{E}$ *is a matrix with rows* $\boldsymbol{e}_{\mathcal{S}}'$ *for every model* $M_{\mathcal{S}}$*, where*

$$\boldsymbol{e}_{\mathcal{S}} = \left[\frac{\partial m(\boldsymbol{\theta}_{\mathcal{F},0})}{\partial \boldsymbol{\theta}_{\mathcal{F}}'}\boldsymbol{J}_{\mathcal{F},0}^{-1} - \frac{\partial m(\boldsymbol{\theta}_{\mathcal{S},0})}{\partial \boldsymbol{\theta}_{\mathcal{S}}'}\boldsymbol{J}_{\mathcal{S},0}^{-1}\begin{pmatrix}\boldsymbol{I} & \boldsymbol{O}\\ \boldsymbol{O} & \boldsymbol{P}_{\mathcal{S}}\end{pmatrix}\right]'.$$

Proof. See Appendix A1.

From this result, we have that $FIC(T_{\boldsymbol{w}})$ may be estimated by

$$\widehat{FIC}^*(T_{\boldsymbol{w}}) = D_{\boldsymbol{w}}^2 - \hat{\tau}_{\boldsymbol{w}}^2 + \tilde{\sigma}_{\boldsymbol{w}}^2 = D_{\boldsymbol{w}}^2 - \frac{1}{\hat{m}_{\mathcal{F}}^2}\boldsymbol{w}'\hat{\boldsymbol{E}}\hat{\boldsymbol{J}}_{\mathcal{F}}\hat{\boldsymbol{E}}'\boldsymbol{w} + \frac{1}{\hat{m}_{\mathcal{F}}^2}\boldsymbol{w}'\hat{\boldsymbol{C}}\hat{\boldsymbol{J}}_{\mathcal{F}}\hat{\boldsymbol{C}}'\boldsymbol{w}, \quad (12)$$

with $\tilde{\sigma}_{\boldsymbol{w}}^2$ denoting the estimator of $\sigma_{\boldsymbol{w}}^2 - (1-m_0)/m_0$ and $\hat{\boldsymbol{E}}$ denoting a matrix with rows $\hat{\boldsymbol{e}}_{\mathcal{S}}'$, where

$$\hat{\boldsymbol{e}}_{\mathcal{S}} = \left[\hat{\boldsymbol{m}}_{\mathcal{F}}'\hat{\boldsymbol{J}}_{\mathcal{F}}^{-1} - \hat{\boldsymbol{m}}_{\mathcal{S}}\tilde{\boldsymbol{J}}_{\mathcal{S}}^{-1}\begin{pmatrix}\boldsymbol{I} & \boldsymbol{O}\\ \boldsymbol{O} & \boldsymbol{P}_{\mathcal{S}}\end{pmatrix}\right]'.$$

Regardless of the estimator used for $FIC(T_{\boldsymbol{w}})$, we suggest to choose the vector of weights $\boldsymbol{w}$ by minimizing the corresponding expression, which is of type (11) or (12), under the constraints $w_{\mathcal{S}} \geqslant 0$, $\forall \mathcal{S}$, and $\sum_{\mathcal{S}} w_{\mathcal{S}} = 1$; see also Hansen (2007). This optimization may be

performed by iterative algorithms which are available in most statistical and mathematical packages. We suggest to initialize these algorithms by using CAIC-weights, since these weights usually represent a good solution in terms of efficiency of the estimator of N. The weights obtained from this optimization are used to compute a multimodel estimate of N on the basis of (3) and, in the following, are indicated by FIC or FIC*-weights according to which estimator is chosen for $FIC(T_{\boldsymbol{w}})$.

5 Simulation study

In order to compare the model selection criteria illustrated above with those based on AIC and CAIC indices, we carried out a simulation study along the same lines as Stanley and Burnham (1998). In the following, we illustrate the class of models considered in the study, the simulation design and then the results we obtained.

5.1 A class of models for capture-recapture data

The full model in the class we considered allows for heterogeneity, time and behavior effects on the capture probabilities. Using a notation taken from Otis *et al.* (1978), this model is then of type M_{htb}; see also Agresti (1994), Pledger (2000) and Dorazio and Royle (2003). The heterogeneity effect is introduced via a random effect z having standard normal distribution. Therefore, the model assumes that, given this random effect and for any $j > 1$, r_j depends on $r_1, \ldots, r_{j-1}$ only through c_{j-1}, where c_{j-1} is a dummy variable equal to 1 if the subject has already been captured (i.e. $\sum_{h<j} r_h > 0$) and to 0 otherwise. This implies that

$$p_{\mathcal{F}}(\boldsymbol{r}) = \int_{\mathbb{R}} p_{\mathcal{F}}(\boldsymbol{r}|z)\phi(z)dz, \tag{13}$$

where

$$p_{\mathcal{F}}(\boldsymbol{r}|z) = \prod_j \lambda_{\mathcal{F}}(j|z, c_{j-1})^{r_j}[1 - \lambda_{\mathcal{F}}(j|z, c_{j-1})]^{1-r_j}$$

and $\lambda_{\mathcal{F}}(j|z, c_{j-1})$ denotes the probability of being captured at the j-th occasion given z and c_{j-1}. Moreover, $\phi(z)$ denotes the density function of the standard normal distribution. The

model also assumes that

$$\log \frac{\lambda_{\mathcal{F}}(j|z, c_{j-1})}{1 - \lambda_{\mathcal{F}}(j|z, c_{j-1})} = \beta_1 + z\beta_2 + \boldsymbol{t}_j'\boldsymbol{\gamma}_1 + c_{j-1}\gamma_2,$$

with $c_0 \equiv 0$ and where $\boldsymbol{t}_j$ is a $(J-1)$-dimensional column vector of all zeros, apart form its $(j-1)$-th element that, when $j > 1$, is equal to r_j. In this way, the parameters in $\boldsymbol{\gamma}_1$ measure the differential effect of each capture occasion with respect to the first occasion (time effect), whereas γ_2 measures the effect of a previous capture (behavior effect). Moreover, β_1 is the intercept and β_2 is a scale parameter for the random effect (heterogeneity effect).

Several submodels may be conceived by setting to 0 suitable parameters of the full model. Using the notation of Otis *et al.* (1978), these models may be indicated by M_{tb}, M_{hb}, M_{ht}, M_b, M_t, M_h, M_0. The first model does not consider the heterogeneity effect, the second does not consider the time effects and so on. Therefore, the last model does not consider any of the mentioned effects. Estimation of each model may be carried out by using the EM algorithm described in Section 2. Technical details useful for its implementation are given in Appendix A2. In our implementation, the integral in (13) is computed by a quadrature method based on a suitable grid of points.

Note that the application of our approach is made difficult by the fact that the conditions on which it is based do not hold. This is because, in order to ensure identifiability and to make possible that the full model specifies into a model without heterogeneity, we require $\beta_2 \geqslant 0$ and then the range of the admissible values of this parameter is not an open set. In order to circumvent this difficult we apply FIC and FIC* (also in the case of multimodel inference) as follows:

- We first test the hypothesis $H_0 : \beta_2 = 0$ (absence of heterogeneity) by a likelihood ratio statistic between models M_{htb} and M_{tb}; this statistic has null asymptotic distribution of type $\chi^2_1/2$.

- If H_0 is rejected, we restrict our attention only to the models including heterogeneity effect so that it is possible to exclude from the parameter space the points with $\beta_2 = 0$ and then the conditions on which our approach is based hold. The parameter vector of the full model M_{htb} is $\boldsymbol{\theta}_{\mathcal{F}} = (\boldsymbol{\beta}', \boldsymbol{\gamma}')'$, with $\boldsymbol{\beta} = (\beta_1, \beta_2)'$ and $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_1', \gamma_2)'$. The

submodels we consider are M_{hb}, M_{ht} and M_h. In our notation, these are indicated by $M_{\mathcal{S}}$ with, respectively, $\mathcal{S} = \{J\}$, $\mathcal{S} = \{1, \ldots, J-1\}$ and $\mathcal{S} = \emptyset$.

- If H_0 is not rejected, we only consider models without heterogeneity effect. The full model is in this case M_{tb} with parameter vector $\boldsymbol{\theta}_{\mathcal{F}} = (\beta, \boldsymbol{\gamma}')'$, where $\beta = \beta_1$. The corresponding parameter space is $\mathbb{R}^J$ and then the conditions on which our approach is based hold. In this case, the submodels we consider are M_b, M_t and M_0, which may also be indicated by $M_{\mathcal{S}}$ with $\mathcal{S}$ again defined as $\mathcal{S} = \{J\}$, $\mathcal{S} = \{1, \ldots, J-1\}$ and $\mathcal{S} = \emptyset$, respectively.

5.2 Simulation design

For $N = 200, 400$ and $J = 5, 7$, we drew 1,000 samples from the full model M_{htb} with $\beta_2 = 0, 0.5$, parameters in $\boldsymbol{\gamma}_1$ generated, for every sample, from the distribution $N(0, \sigma^2_{\boldsymbol{\gamma}_1})$, $\sigma^2_{\boldsymbol{\gamma}_1} = 0, 0.1$, and $\gamma_2 = -0.5, 0, 0.5$. Note that letting $\beta_2 = 0$ is equivalent to removing the heterogeneity effect from M_{htb}, letting $\sigma^2_{\boldsymbol{\gamma}_1} = 0$ is equivalent to removing the time effect and letting $\gamma_2 = 0$ is equivalent to removing the behavior effect. Therefore, when $\beta_2 = \sigma^2_{\boldsymbol{\gamma}_1} = \gamma_2 = 0$, model M_0 results. Overall, we considered 48 scenarios corresponding to two choices for J, two for N and twelve for the model parameters. Under each scenario, the intercept β_1 is chosen in order to have an expected sample size close to the 75% of the population size. We then set $\beta_1 = -1.1$ when $J = 5$ and $\beta_1 = -1.5$ when $J = 7$.

For each generated sample, we fitted all the available models and we took, as estimate of N, the one computed on the basis of the model selected with AIC, CAIC, FIC (based on the estimator $\widehat{FIC}(T_{\mathcal{S}})$) and FIC* (based on the estimator $\widehat{FIC}^*(T_{\mathcal{S}})$). Note that, taking the structure of the parameter space into account, the last two criteria are applied as described at the end of the previous section. To assess the quality of the multimodel inference proposed here, for each sample we also computed the weighted average estimate of N based on AIC, CAIC, FIC and FIC*-weights. In this way we considered height different estimators of N. Note that we did not consider AIC_c because, according to Stanley and Burnham (1998), the estimator of N based on this criterion performs very closely to that based on AIC.

For each estimator of N described above, we computed the Relative Mean Squared Error (RMSE) as

$$\frac{1}{1000}\sum_{h=1}^{1000}\left(\frac{\hat{N}^{(h)}-N}{N}\right)^2,$$

where $\hat{N}^{(h)}$ is the estimate obtained from the h-th simulated sample. The results are displayed in Tables 1 to 4 which also show the RMSE of $\hat{N}_{\mathcal{F}}$ (the estimator of N based on the full model M_{htb}) together with the average percentage of captured population (CP), the percentage of anomalous samples (RS) and the percentage of times the test of the hypothesis of absence of heterogeneity ($H_0 : \beta_2 = 0$) is rejected (ET). In particular, RS is the percentage of samples for which the maximum among the estimates of N obtained from M_{htb}, M_{hb}, M_{ht} and M_h (when absence of heterogeneity is rejected) or the maximum among those obtained from M_{tb}, M_b, M_t and M_0 (when absence of heterogeneity is not rejected) is larger than four times n. These samples are discharged and regenerated within the simulation, since they lead to an unrealistic estimate of N. Note that procedures like this are common to other simulation studies performed in the capture-recapture literature, as the one of Stanley and Burnham (1998). They, in particular, adopted a procedure which discharges samples whenever at least one of the estimates of N is greater than three times the true value of this parameter. However, we consider our procedure more realistic, since it does require to know the true population size.

5.3 Results

We first consider the results in Tables 1 and 2 of the simulations carried out in absence of heterogeneity ($\beta_2 = 0$). The first of these tables is referred to the case of $J = 5$ trapping occasions. We can observe that, in terms of RMSE, FIC* is the best criterion for 9 of the 12 scenarios considered in this table. For the remaining 3 scenarios (corresponding to the generating models M_0 and M_b, which have not a high level of complexity), FIC* is the second best criterion. Moreover, FIC results to be the second best criterion 7 times on 12. A similar pattern is observed for $J = 7$ (see Table 2). In this case FIC* is the best criterion 8 times on 12 and the second best for the 4 remaining times. FIC is second best criterion 7 times

on 12. Overall FIC and FIC* behave much better than AIC and CAIC and the advantage is slightly more evident with $N = 200$ than with $N = 400$.

On the basis of Tables 1 and 2, we can draw similar conclusions for what concerns multimodel inference. In particular, the average estimators of N based on FIC*-weights is the best, in terms of RMSE, 8 times on 12 when $J = 5$ and when $J = 7$. In the remaining cases (essentially when samples are generated from models M_0 and M_b), it is the second best multimodel estimator of N. Moreover, the average estimators of N based on FIC-weights is the second best estimator 7 times on 12 when $J = 5$ and when $J = 7$. We also observe that, in terms of RMSE, a multimodel estimator always behaves better than the corresponding single-model estimator and then, estimating N by a weighted average based on FIC*-weights, results to be the best strategy in most of the scenarios without heterogeneity.

Under the scenarios based on models including heterogeneity effect, see Tables 3 and 4, FIC and FIC* have very good performances, even if slightly worse in comparison to the scenarios without heterogeneity. We can observe that, when $J = 5$ (Table 3), FIC* is the best criterion in terms of RMSE for 7 cases on 12 and in the remaining cases (essentially under the generating model M_h and M_{hb}) it is the second best criterion. Moreover, FIC is the best selection criterion one time and the second best 5 times on 12. When $J = 7$, FIC* is the best selection criterion 7 times on 12. FIC is the second best criterion 4 times on 12; the same happens for FIC*. For what concerns multimodel inference, weighted estimators based on FIC and FIC*-weights behave well, even if the advantage with respect to estimators based on AIC and CAIC-weights is less evident than in the case of absence of heterogeneity.

Overall, we can conclude that, both in absence and in presence of heterogeneity, FIC and FIC* usually outperform AIC and CAIC in terms of the quality of the inference on the population size. In particular, the advantage of FIC and FIC* over AIC and CAIC seems to increase with the complexity of the generating model, whereas this advantage seems to slightly decrease as N and J increase and in presence of heterogeneity. We can also observe that the estimator based on FIC* is almost always more efficient than that based on FIC. Similar considerations may be drawn for multimodel inference and we have also to note that weighted average estimators are always more efficient than the corresponding single-model

estimators. This was already noticed by Stanley and Burnham (1998) for the estimators based on AIC and CAIC-weights and is also evident for the average estimators based on FIC and FIC*-weights.

We repeated the simulation study above with smaller values for the intercepts, so as to understand what happens when the proportion of captured population is less than 75%. For instance, with $J = 5$, we considered -1.6 instead of -1.1 as true value of the parameter β_1 so to have a percentage of capture population (CP) around 60%. In this case, all the estimators seem to worse in the same direction with an increase of the RMSE roughly proportional under each scenario we considered. Therefore, these new set of simulations lead to the same conclusions in terms of comparison between the selection criteria. For this reason, and also because these new simulations gave less stable results, we prefer to avoid reporting detailed results. The greater instability is directly due to the smaller percentage of captured population and is confirmed by the fact that the percentage of anomalous samples (RS) considerably increases. For instance, under model M_0, with $J = 5$ and $N = 400$, this percentage increases from 1.5% to around 15% when $\beta_1 = -1.6$.

6 An application

In order to illustrate the approach proposed in this paper, we analyze a dataset collected in a live-trapping study of meadow voles (*Microtus pennsylvanicus*) based on $J = 5$ consecutive daily trapping sessions, which took place in June 1981. The number of animals captured at least once is $n = 104$. For details about the survey design see Nichols *et al.* (1984); see also Bartolucci and Pennoni (2007) and the references therein.

We first fitted the models described in the previous section (M_0, M_t, M_b, M_{tb}, M_h, M_{ht}, M_{hb}, M_{htb}) to the data at hand. The results, in terms of maximum log-likelihood, estimate of N, AIC and CAIC indices of each model are reported in Table 5. The table also shows the AIC and CAIC-weights and the corresponding single-model and multimodel estimates of N. In particular, AIC and CAIC lead to choosing the same model, M_h, and then to the same single-model estimate of N equal to 123.87. However, AIC does not give a large weight

to M_h and so the resulting multimodel estimate of N is a compromise among the estimates obtained from the models with heterogeneity; this estimate is equal to 131.99. On the other hand, the strategy based on CAIC-weights gives a very large weight to model M_h and then the resulting multimodel estimate, equal to 124.38, is close to the single-model one.

In order to apply FIC and FIC* we initially tested the hypothesis of absence of heterogeneity ($H_0 : \beta_2 = 0$) by a likelihood ratio statistic between models M_{htb} and M_{tb}. This test statistic is equal to 23.42 with a p-value less than 10^{-3} which clearly leads to rejecting H_0. The presence of a heterogeneity effect is confirmed by very low AIC and CAIC-weights for models which ignore this effect. Among these models there is M_{tb} which gives an estimate surprisingly large, i.e. 632.27. Then, following the approach illustrated at the end of Section 5.1, FIC and FIC* are applied only to models with heterogeneity (M_h, M_{ht}, M_{hb}, M_{htb}). For each model, Table 6 shows the value of $\widehat{FIC}$ and $\widehat{FIC}^*$ together with the single-model estimates of N and the corresponding weights for multimodel inference. It may be observed that FIC leads to choosing model M_{htb}, whereas FIC* leads to choosing the same model as AIC and CAIC, which is M_h. The single-model estimates of N are then 153.64 (with FIC) and 123.87 (with FIC*). Finally, the multimodel estimate based on FIC-weights is a compromise between those under models M_{hb} and M_{htb} and is equal to 141.91, whereas the multimodel estimate based on FIC*-weights is a compromise between those under models M_h and M_{htb} and is equal to 137.27.

Overall, the single-model estimate which seems to be preferred is $\hat{N} = 123.87$ deriving from model M_h; in fact, 3 criteria on 4 prefer this model, included FIC*. On the other hand, multimodel estimates obtained on the basis of FIC and FIC*-weights are larger than the AIC and CAIC multimodel estimates. However, on the basis of our simulation study which show the superiority of the strategy based on FIC*-weights over the other estimation strategies (especially when the population size is small) we take 137.27 as estimate of N.

7 Conclusions

Following the intuition of Claeskens and Hjort (2003), in this paper we introduce a selection criterion for capture-recapture models which takes into account the MSE of the resulting estimator of the population size. To this aim, we first extend the main asymptotic results of Hjort and Claeskens (2003) to the CML method of Sanathanan (1972) which is currently used for estimating capture-recapture models for closed populations. On the basis of these results, we derive an expression for the asymptotic MSE of the estimator of the population size which is also valid when the assumed model is different from the true one. This expression is given by a constant term plus a term depending on the model, which is a FIC index in the sense of Hjort and Claeskens (2003). We then introduce two different methods for estimating the FIC index. The first method recalls that developed by Claeskens and Hjort (2003), whereas the second seems to have a nicer interpretation because it derives from the asymptotic properties of the difference between the estimators of N computed under different models. Through a simulation study, carried out along the same lines as Stanley and Burnham (1998), we show that the proposed selection criterion generally outperforms AIC and CAIC in terms of quality of the inference induced on the population size, especially when we use the second estimation method for the FIC index.

In this paper, we also deal with multimodel inference for the population size on the basis of a weighted average estimator. To this aim, we first develop a weighted FIC index which may again be estimated in two different ways. We then propose a procedure for choosing weights which consists of minimizing the estimate of the weighted FIC index. Also in this case, we established by simulation that the proposed strategy generally outperforms that based on AIC and CAIC.

The proposed approach can be generalized to a hierarchical class of capture-recapture models which also include individual covariates. CML estimation of these models was introduced by Alho (1990) and may be implemented by using an EM algorithm similar to that illustrated in Section 2; see also Bartolucci and Forcina (2006). The asymptotic properties of this estimation method, when the assumed model is different from the true one, has first

to be studied along the same lines as we did in Section 3. On the basis of these results, it should then be possible to derive a FIC index and a weighted FIC index, similar to those derived in Section 4, which have to be properly estimated in order to implement a FIC selection strategy. This selection strategy also involves the choice of the best set of covariates to include in the model in order to minimize the MSE of the estimator of the population size induced by the model. Finally, a worthwhile extension is that to open-population models which, together with closed-population models, represent an important class of models for capture-recapture data. This class of models is suitable when the capture-recapture experiment takes long time and then modification of the population size, due essentially to births and deaths, must be taken into account. Model selection strategies for these models are studied, in particular, by Burnham *et al.* (1995).

Appendix

A1: Proofs

Proof of Theorem 1. Consistency can easily be proved by following the first part of the proof of Theorem 2 of Sanathanan (1972) and considering that $p_{\text{true}}(\boldsymbol{r}) \to p_0(\boldsymbol{r})$ as $N \to \infty$.

For what concerns asymptotic normality, consider first that

$$\boldsymbol{s}(\hat{\boldsymbol{\theta}}_{\mathcal{S}}) = N\frac{\partial \boldsymbol{q}(\hat{\boldsymbol{\theta}}_{\mathcal{S}})'}{\partial \boldsymbol{\theta}_{\mathcal{S}}}\text{diag}(\hat{\boldsymbol{q}}_{\mathcal{S}})^{-1}\boldsymbol{f} = \boldsymbol{0} \qquad \text{and} \qquad N\frac{\partial \boldsymbol{q}(\hat{\boldsymbol{\theta}}_{\mathcal{S}})'}{\partial \boldsymbol{\theta}_{\mathcal{S}}}\text{diag}(\hat{\boldsymbol{q}}_{\mathcal{S}})^{-1}\hat{\boldsymbol{p}}_{\mathcal{S}} = \boldsymbol{0}$$

and then

$$N\frac{\partial \boldsymbol{q}(\hat{\boldsymbol{\theta}}_{\mathcal{S}})'}{\partial \boldsymbol{\theta}_{\mathcal{S}}}\text{diag}(\hat{\boldsymbol{q}}_{\mathcal{S}})^{-1}(\boldsymbol{f} - \boldsymbol{p}_0) = N\frac{\partial \boldsymbol{q}(\hat{\boldsymbol{\theta}}_{\mathcal{S}})'}{\partial \boldsymbol{\theta}_{\mathcal{S}}}\text{diag}(\hat{\boldsymbol{q}}_{\mathcal{S}})^{-1}(\hat{\boldsymbol{p}}_{\mathcal{S}} - \boldsymbol{p}_0). \tag{14}$$

We also have that

$$\hat{\boldsymbol{p}}_{\mathcal{S}} - \boldsymbol{p}_0 = \frac{\partial \boldsymbol{p}(\bar{\boldsymbol{\theta}}_{\mathcal{S}})}{\partial \boldsymbol{\theta}'_{\mathcal{S}}}(\hat{\boldsymbol{\theta}}_{\mathcal{S}} - \boldsymbol{\theta}_{\mathcal{S},0}), \tag{15}$$

where $\bar{\boldsymbol{\theta}}_{\mathcal{S}}$ is an intermediate point between $\boldsymbol{\theta}_{\mathcal{S},0}$ and $\hat{\boldsymbol{\theta}}_{\mathcal{S}}$. Now substitute (15) in (14), divide the resulting expression by $\sqrt{N}$ and consider that

$$\boldsymbol{J}(\boldsymbol{\theta}_{\mathcal{S}}) = \frac{\partial \boldsymbol{p}'_{\mathcal{S}}}{\partial \boldsymbol{\theta}_{\mathcal{S}}}\text{diag}(\boldsymbol{q}_{\mathcal{S}})^{-1}\frac{\partial \boldsymbol{q}_{\mathcal{S}}}{\partial \boldsymbol{\theta}'_{\mathcal{S}}}$$

and that $\hat{\boldsymbol{\theta}}_{\mathcal{S}} \xrightarrow{p} \boldsymbol{\theta}_{\mathcal{S},0}$ and $\bar{\boldsymbol{\theta}}_{\mathcal{S}} \xrightarrow{p} \boldsymbol{\theta}_{\mathcal{S},0}$ as $N \to \infty$. It results that

$$\sqrt{N}(\hat{\boldsymbol{\theta}}_{\mathcal{S}} - \boldsymbol{\theta}_{\mathcal{S},0}) \xrightarrow{p} \sqrt{N}\boldsymbol{J}^{-1}_{\mathcal{S},0}\frac{\partial \boldsymbol{q}(\boldsymbol{\theta}_{\mathcal{S},0})'}{\partial \boldsymbol{\theta}_{\mathcal{S}}}\text{diag}(\boldsymbol{q}_0)^{-1}(\boldsymbol{f} - \boldsymbol{p}_0). \tag{16}$$

Finally, the Theorem follows because of (5) and since

$$\frac{\partial \boldsymbol{q}(\boldsymbol{\theta}_{\mathcal{S},0})'}{\partial \boldsymbol{\theta}_{\mathcal{S}}}\mathrm{diag}(\boldsymbol{q}_0)^{-1}\frac{\partial \boldsymbol{p}_0}{\partial \boldsymbol{\gamma}'} = \begin{pmatrix} \boldsymbol{J}_{01} \\ \boldsymbol{P}_{\mathcal{S}}\boldsymbol{J}_{11} \end{pmatrix} \tag{17}$$

$$\frac{\partial \boldsymbol{q}(\boldsymbol{\theta}_{\mathcal{S},0})'}{\partial \boldsymbol{\theta}_{\mathcal{S}}}\mathrm{diag}(\boldsymbol{q}_0)^{-1}\boldsymbol{\Omega}_0\mathrm{diag}(\boldsymbol{q}_0)^{-1}\frac{\partial \boldsymbol{q}(\boldsymbol{\theta}_{\mathcal{S},0})}{\partial \boldsymbol{\theta}'_{\mathcal{S}}} = \boldsymbol{J}_{\mathcal{S},0}. \tag{18}$$

Proof of Theorem 2. Consistency may be proved by extending the first part of the proof of Theorem 2 of Sanathanan (1972). Consider in particular that, as $N \to \infty$, both n/N and $\hat{m}_{\mathcal{S}}$ converge in probability to m_0.

To prove asymptotic normality, take the following first-order expansion of (6) around $\boldsymbol{\theta}_{\mathcal{S},0}$

$$T_{\mathcal{S}} = \sqrt{N}\left[\frac{\mathbf{1}'\boldsymbol{f}}{m_0} - 1 - \frac{\mathbf{1}'\boldsymbol{f}}{\bar{m}_{\mathcal{S}}^2}\frac{\partial m_{\mathcal{S}}(\bar{\boldsymbol{\theta}}_{\mathcal{S}})}{\partial \boldsymbol{\theta}'_{\mathcal{S}}}(\hat{\boldsymbol{\theta}}_{\mathcal{S}} - \boldsymbol{\theta}_{\mathcal{S},0})\right],$$

with $\bar{\boldsymbol{\theta}}_{\mathcal{S}}$ defined as above. As $N \to \infty$, $\mathbf{1}'\boldsymbol{f} \xrightarrow{p} m_0$, $\bar{\boldsymbol{\theta}}_{\mathcal{S}} \xrightarrow{p} \boldsymbol{\theta}_{\mathcal{S},0}$ and therefore $\bar{m}_{\mathcal{S}} \xrightarrow{p} m_0$. Considering also (16) and after some algebra, we have

$$T_{\mathcal{S}} \xrightarrow{p} \frac{\sqrt{N}}{m_0}\boldsymbol{a}'_{\mathcal{S}}(\boldsymbol{f} - \boldsymbol{p}_0), \qquad \boldsymbol{a}_{\mathcal{S}} = \left[\mathbf{1}' - \frac{\partial m(\boldsymbol{\theta}_{\mathcal{S},0})}{\partial \boldsymbol{\theta}'_{\mathcal{S}}}\boldsymbol{J}_{\mathcal{S},0}^{-1}\frac{\partial \boldsymbol{q}(\boldsymbol{\theta}_{\mathcal{S},0})'}{\partial \boldsymbol{\theta}_{\mathcal{S}}}\mathrm{diag}(\boldsymbol{q}_0)^{-1}\right]'. \tag{19}$$

The result then follows from (5). In particular, the expression given for $\mu_{\mathcal{S}}$ derives from

$$\frac{\partial \boldsymbol{p}'_0}{\partial \boldsymbol{\gamma}}\boldsymbol{a}_{\mathcal{S}} = \boldsymbol{b}_{\mathcal{S}}, \tag{20}$$

which holds because of (17). The expression given for $\sigma^2_{\mathcal{S}}$ derives from (18) and because

$$\mathbf{1}'\boldsymbol{\Omega}_0\mathbf{1} = m_0(1 - m_0) \qquad \text{and} \qquad \frac{\partial \boldsymbol{q}(\boldsymbol{\theta}_{\mathcal{S},0})'}{\partial \boldsymbol{\theta}_{\mathcal{S}}}\mathrm{diag}(\boldsymbol{q}_0)^{-1}\boldsymbol{\Omega}_0\mathbf{1} = \mathbf{0}.$$

Proof of Theorem 3. Proceeding as in the proof of Theorem 2, we find that

$$T_{\boldsymbol{w}} \xrightarrow{p} \frac{\sqrt{N}}{m_0}\boldsymbol{w}'\boldsymbol{A}(\boldsymbol{f} - \boldsymbol{p}_0),$$

with $\boldsymbol{A}$ denoting a matrix with rows $\boldsymbol{a}'_{\mathcal{S}}$ for each model $M_{\mathcal{S}}$. The result then follows from (5). In particular, the expression for $\mu_{\boldsymbol{w}}$ derives from (20). The expression for $\sigma^2_{\boldsymbol{w}}$ derives from $\boldsymbol{A}\boldsymbol{\Omega}_0\boldsymbol{A}' = m_0(1 - m_0)\mathbf{1}\mathbf{1}' + \boldsymbol{C}\boldsymbol{J}_{\mathcal{F},0}\boldsymbol{C}'$, which holds because

$$\frac{\partial \boldsymbol{q}(\boldsymbol{\theta}_{\mathcal{S},0})'}{\partial \boldsymbol{\theta}_{\mathcal{S}}}\mathrm{diag}(\boldsymbol{q}_0)^{-1}\boldsymbol{\Omega}_0\mathrm{diag}(\boldsymbol{q}_0)^{-1}\frac{\partial \boldsymbol{q}(\boldsymbol{\theta}_{\mathcal{S}',0})}{\partial \boldsymbol{\theta}'_{\mathcal{S}'}} = \begin{pmatrix} \boldsymbol{I} & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{P}_{\mathcal{S}} \end{pmatrix}\boldsymbol{J}_{\mathcal{F},0}\begin{pmatrix} \boldsymbol{I} & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{P}'_{\mathcal{S}'} \end{pmatrix}$$

for every pair of models $M_{\mathcal{S}}$ and $M_{\mathcal{S}'}$.

Proof of Theorem 4. First of all consider that

$$D_{\mathcal{S}} \xrightarrow{p} \frac{\hat{N}_{\mathcal{S}} - N}{\sqrt{N}} - \frac{\hat{N}_{\mathcal{F}} - N}{\sqrt{N}}.$$

From (19) it then follows that

$$D_{\mathcal{S}} \xrightarrow{p} \frac{\sqrt{N}}{m_0}(\boldsymbol{a}_{\mathcal{S}} - \boldsymbol{a}_{\mathcal{F}})'(\boldsymbol{f} - \boldsymbol{p}_0).$$

As usual, the result follows from (5). For what concerns $\nu_{\mathcal{S}}$, we have to consider that

$$\frac{\partial \boldsymbol{p}_0'}{\partial \boldsymbol{\gamma}}(\boldsymbol{a}_{\mathcal{S}} - \boldsymbol{a}_{\mathcal{F}}) = \boldsymbol{b}_{\mathcal{S}}$$

because of (20) and since $\boldsymbol{b}_{\mathcal{F}} = 0$. Now let

$$\boldsymbol{d}_{\mathcal{S}} = \left[\frac{\partial m(\boldsymbol{\theta}_{\mathcal{S},0})}{\partial \boldsymbol{\theta}_{\mathcal{S}}'}\boldsymbol{J}_{\mathcal{S},0}^{-1}\frac{\partial \boldsymbol{q}'(\boldsymbol{\theta}_{\mathcal{S},0})}{\partial \boldsymbol{\theta}_{\mathcal{S}}}\right]'$$

and note that $(\boldsymbol{a}_{\mathcal{S}} - \boldsymbol{a}_{\mathcal{F}})'\boldsymbol{\Omega}_0(\boldsymbol{a}_{\mathcal{S}} - \boldsymbol{a}_{\mathcal{F}}) = m_0^2(\boldsymbol{d}_{\mathcal{F}} - \boldsymbol{d}_{\mathcal{S}})'\text{diag}(\boldsymbol{p}_0)^{-1}(\boldsymbol{d}_{\mathcal{F}} - \boldsymbol{d}_{\mathcal{S}})$. The expression for $\tau_{\mathcal{S}}^2$ derives from

$$\begin{aligned}
m_0^2\boldsymbol{d}_{\mathcal{S}}'\text{diag}(\boldsymbol{p}_0)^{-1}\boldsymbol{d}_{\mathcal{S}} &= \frac{\partial m(\boldsymbol{\theta}_{\mathcal{S},0})}{\partial \boldsymbol{\theta}_{\mathcal{S}}'}\boldsymbol{J}_{\mathcal{S},0}^{-1}\frac{\partial m(\boldsymbol{\theta}_{\mathcal{S},0})}{\partial \boldsymbol{\theta}_{\mathcal{S}}} \\
m_0^2\boldsymbol{d}_{\mathcal{S}}'\text{diag}(\boldsymbol{p}_0)^{-1}\boldsymbol{d}_{\mathcal{F}} &= \frac{\partial m(\boldsymbol{\theta}_{\mathcal{S},0})}{\partial \boldsymbol{\theta}_{\mathcal{S}}'}\boldsymbol{J}_{\mathcal{S},0}^{-1}\begin{pmatrix} \boldsymbol{J}_{00} & \boldsymbol{J}_{01} \\ \boldsymbol{P}_{\mathcal{S}}\boldsymbol{J}_{11} & \boldsymbol{P}_{\mathcal{S}}\boldsymbol{J}_{11} \end{pmatrix}\boldsymbol{J}_{\mathcal{F},0}^{-1}\frac{\partial m(\boldsymbol{\theta}_{\mathcal{F},0})}{\partial \boldsymbol{\theta}_{\mathcal{F}}} = \frac{\partial m(\boldsymbol{\theta}_{\mathcal{S},0})}{\partial \boldsymbol{\theta}_{\mathcal{S}}'}\boldsymbol{J}_{\mathcal{S},0}^{-1}\frac{\partial m(\boldsymbol{\theta}_{\mathcal{S},0})}{\partial \boldsymbol{\theta}_{\mathcal{S}}}.
\end{aligned}$$

Proof of Theorem 5. First of all consider that

$$D_{\boldsymbol{w}} \xrightarrow{p} \sum_{\mathcal{S}} w_{\mathcal{S}}\left(\frac{\hat{N}_{\mathcal{S}} - N}{\sqrt{N}} - \frac{\hat{N}_{\mathcal{F}} - N}{\sqrt{N}}\right).$$

Using the notation of the proofs of Theorems 3 and 4, we then have that

$$D_{\boldsymbol{w}} \xrightarrow{p} \frac{\sqrt{N}}{m_0}(\boldsymbol{A} - \mathbf{1}\boldsymbol{a}_{\mathcal{F}}')(\boldsymbol{f} - \boldsymbol{p}_0).$$

The result then follows because of (5) and (20) and because $\boldsymbol{A} - \mathbf{1}\boldsymbol{a}_{\mathcal{F}}' = (\mathbf{1}\boldsymbol{d}_{\mathcal{F}}' - \boldsymbol{D})\text{diag}(\boldsymbol{q}_0)^{-1}$, where $\boldsymbol{D}$ is a matrix with rows $\boldsymbol{d}_{\mathcal{S}}'$ for every model $M_{\mathcal{S}}$. Moreover,

$$(\mathbf{1}\boldsymbol{d}_{\mathcal{F}}' - \boldsymbol{D})\text{diag}(\boldsymbol{q}_0)^{-1}\boldsymbol{\Omega}_0\text{diag}(\boldsymbol{q}_0)^{-1}(\mathbf{1}\boldsymbol{d}_{\mathcal{F}}' - \boldsymbol{D})' = m_0^2(\mathbf{1}\boldsymbol{d}_{\mathcal{F}}' - \boldsymbol{D})\text{diag}(\boldsymbol{p}_0)^{-1}(\mathbf{1}\boldsymbol{d}_{\mathcal{F}}' - \boldsymbol{D})'$$

and the row of the matrix $\mathbf{1}\boldsymbol{d}_{\mathcal{F}}' - \boldsymbol{D}$ corresponding to model $M_{\mathcal{S}}$ is

$$\boldsymbol{d}_{\mathcal{F}} - \boldsymbol{d}_{\mathcal{S}} = \frac{\partial \boldsymbol{q}(\boldsymbol{\theta}_{\mathcal{F},0})}{\partial \boldsymbol{\theta}_{\mathcal{F}}'}\boldsymbol{e}_{\mathcal{S}}.$$

A2: Partial derivatives of the probabilities $p_{\mathcal{S}}(\boldsymbol{r})$ and $q_{\mathcal{S}}(\boldsymbol{r})$

For each model $M_{\mathcal{S}}$ including heterogeneity effect, the following expressions derive from (13) for the first derivative vector and the second derivative matrix of the probability $p_{\mathcal{S}}(\boldsymbol{r})$:

$$\frac{\partial p_{\mathcal{S}}(\boldsymbol{r})}{\partial \boldsymbol{\theta}_{\mathcal{S}}} = \int_{\mathbb{R}} \frac{\partial p_{\mathcal{S}}(\boldsymbol{r}|z)}{\partial \boldsymbol{\theta}_{\mathcal{S}}} \phi(z) dz, \qquad \frac{\partial^2 p_{\mathcal{S}}(\boldsymbol{r})}{\partial \boldsymbol{\theta}_{\mathcal{S}} \partial \boldsymbol{\theta}'_{\mathcal{S}}} = \int_{\mathbb{R}} \frac{\partial^2 p_{\mathcal{S}}(\boldsymbol{r}|z)}{\partial \boldsymbol{\theta}_{\mathcal{S}} \partial \boldsymbol{\theta}'_{\mathcal{S}}} \phi(z) dz.$$

We also have that

$$\frac{\partial p_{\mathcal{S}}(\boldsymbol{r}|z)}{\partial \boldsymbol{\theta}_{\mathcal{S}}} = p_{\mathcal{S}}(\boldsymbol{r}|z) \sum_j [r_j - \lambda_{\mathcal{S}}(j|z, c_{j-1})] \boldsymbol{x}_{\mathcal{S}}(j|z, c_{j-1}),$$

where $\boldsymbol{x}_{\mathcal{S}}(j|z, c_{j-1})$ is a subvector of $\boldsymbol{x}_{\mathcal{F}}(j|z, c_{j-1}) = (1, z, \boldsymbol{t}'_j, c_{j-1})'$ such that

$$\lambda_{\mathcal{S}}(j|z, c_{j-1}) = \frac{\exp[\boldsymbol{x}_{\mathcal{S}}(j|z, c_{j-1})' \boldsymbol{\theta}_{\mathcal{S}}]}{1 + \exp[\boldsymbol{x}_{\mathcal{S}}(j|z, c_{j-1})' \boldsymbol{\theta}_{\mathcal{S}}]}.$$

Using the matrix notation, we can also write

$$\frac{\partial p_{\mathcal{S}}(\boldsymbol{r}|z)}{\partial \boldsymbol{\theta}_{\mathcal{S}}} = p_{\mathcal{S}}(\boldsymbol{r}|z) \boldsymbol{X}_{\mathcal{S}}(\boldsymbol{r}|z)' \boldsymbol{u}_{\mathcal{S}}(\boldsymbol{r}|z), \tag{21}$$

where $\boldsymbol{X}_{\mathcal{S}}(\boldsymbol{r}|z)$ is a matrix with rows $\boldsymbol{x}_{\mathcal{S}}(j|z, c_{j-1})'$ and $\boldsymbol{u}_{\mathcal{S}}(\boldsymbol{r}|z)$ is a vector with elements $r_j - \lambda_{\mathcal{S}}(j|z, c_{j-1})$, for $j = 1, \ldots, J$. For what concerns the second derivative, we have

$$\frac{\partial^2 p_{\mathcal{S}}(\boldsymbol{r}|z)}{\partial \boldsymbol{\theta}_{\mathcal{S}} \partial \boldsymbol{\theta}'_{\mathcal{S}}} = p_{\mathcal{S}}(\boldsymbol{r}|z) \boldsymbol{X}_{\mathcal{S}}(\boldsymbol{r}|z)' \{\boldsymbol{u}_{\mathcal{S}}(\boldsymbol{r}|z) \boldsymbol{u}_{\mathcal{S}}(\boldsymbol{r}|z)' - \mathrm{diag}[\boldsymbol{v}_{\mathcal{S}}(\boldsymbol{r}|z)]\} \boldsymbol{X}_{\mathcal{S}}(\boldsymbol{r}|z), \tag{22}$$

where $\boldsymbol{v}_{\mathcal{S}}(\boldsymbol{r}|z)$ is a vector with elements $\lambda_{\mathcal{S}}(j|z, c_{j-1})[1 - \lambda_{\mathcal{S}}(j|z, c_{j-1})]$, $j = 1, \ldots, J$.

For a model $M_{\mathcal{S}}$ not including heterogeneity effect, the first and second derivatives of $p_{\mathcal{S}}(\boldsymbol{r})$ can be simply deduced from (21) and (22) considering that the random effect z must be ignored.

References

Agresti, A. (1994). Simple capture recapture models permitting unequal catchability and variable sampling effort. *Biometrics* **50**, 494–500.

Akaike, H. (1973). Information theory as an extension of the maximum likelihood principle. *Second International symposium on information theory.* Petrov, B. N. and Csaki F. (eds), 267–281. Budapest: Akademiai Kiado.

Alho, J. M. (1990). Logistic Regression in Capture-Recapture Models. *Biometrics* **46**, 623–635.

Amstrup, S. C., McDonald, T. L. and Manly, B. F. J. (2005). *Handbook of Capture-Recapture Analysis*, Princeton University Press.

Bartolucci, F. and Forcina, A. (2006). A class of latent marginal models for capture-recapture data with continuous covariate. *Journal of the American Statistical Association*, **101**, 786–794.

Bartolucci, F. and Pennoni, F. (2007), A class of latent Markov models for capture-recapture data allowing for time, heterogeneity and behavior effects, *Biometrics*, **63**, 568–578.

Borchers, D. L., Buckland, S. T. and Zucchini, W. (2004). *Estimating Animal Abundance*, Springer.

Buckland, S. T., Burnham, K. P. and Augustin, N. H. (1997). Model Selection: An integral Part of Inference. *Biometrics* **53**, 603–618.

Burnham, K. P., White, G. C. and Anderson, D. R. (1995). Model selection strategy in the Analysis of capture-recapture data. *Biometrics* **51**, 888–898.

Burnham, K. P. and Anderson, D. R. (2002). *Model selection and multimodel inference.* Springer, New York.

Claeskens, G. and Hjort, N. L. (2003). The focused information criterion. *Journal of the American Statistical Association* **98**, 900–916.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society, Series B* **39**, 1–22.

Dorazio, R. M. and Royle, J. A.(2003). Mixture models for estimating the size of a closed population when capture rates vary among individuals. *Biometrics* **59**, 351–364.

Fienberg, S. E., Johnson, M. S. and Junker, B. W. (1999). Classical multilevel and Bayesian approaches to population size estimation using multiple lists. *Journal of the Royal Statistical Society A* **162**, 383–405.

Hansen, B. E. (2007). Least Squares Model Averaging. *Econometrica* **75**, 1175–1189.

Hjort, N. L. and Claeskens, G. (2003). Frequentist model average estimators. *Journal of the American Statistical Association* **98**, 879–899.

Nichols, J. D., Pollock, K. H., and Hines J. E. (1984). The use of a robust capture-recapture design in small mammal population studies: A field example with *Microtus pennsylvanicus. Acta Theriologica* **29**, 357–365.

Otis, D. L., Burham, K. P., White, G. C. and Anderson, D. R. (1978). Statistical inference from capture data on closed animal populations. *Wildlife Monographs* **62**, 1–135.

Pledger, S. (2000). Unified maximum likelihood estimates for closed capture-recapture models using mixtures. *Biometrics* **56**, 434–442.

Rao, C. R. (1965). *Linear statistical inference and its applications.* New York : Wiley & Sons.

Sanathanan, L. (1972). Estimating the size of a multinomial population. *The Annals of Mathematical Statistics* **43**, 142–152.

Stanley, T. and Burnham, K. P. (1998). Information-theoretic model selection and model averaging for closed-population capture-recapture studies. *Biometrics Journal* **40**, 475–494.

Schwarz, C. J. and Seber, A. F. (1999). Estimating animal abundance: review III. *Statistical Science* **14**, 427–456.

Yip, P. S. F., Bruno, G., Tajima, N., Seber, G. A. F., Buckland, S. T., Cormack, R. M., Unwin, N., Chang, Y.-F., Fienberg, S. E., Junker, B. W., LaPorte, R. E., Libman, I. M. and McCarty D. J. (1995a). Capture-recapture and multiple-record systems estimation I: history and theoretical development. *American Journal of Epidemiology* **142**, 1047–1058.

Yip, P. S. F., Bruno, G., Tajima, N., Seber, G. A. F., Buckland, S. T., Cormack, R. M., Unwin, N., Chang, Y.-F., Fienberg, S. E., Junker, B. W., LaPorte, R. E., Libman, I. M. and McCarty D. J. (1995b). Capture-recapture and multiple-record systems estimation II: applications in human diseases. *American Journal of Epidemiology* **142**, 1059–1068.

Tables of a simulation study described in Section 5

		$N = 200$											
$\sigma^2_{\gamma_1}$	γ_2	FULL	AIC	CAIC	FIC	FIC*	AICw	CAICw	FICw	FICw*	CP%	RS%	ET%
0.0	0.0	10.32	1.30	0.53	0.99	0.81	0.99	0.50	0.75	0.59	76.18	4.3	5.3
0.0	0.5	11.02	3.09	2.34	1.94	1.57	2.36	1.94	1.50	1.18	76.35	5.4	4.2
0.0	-0.5	13.20	3.96	4.75	4.06	3.69	3.43	3.82	3.81	3.47	76.19	6.2	5.4
0.1	0.0	10.35	4.22	4.11	1.64	0.86	2.82	3.13	1.34	0.86	77.11	6.9	4.0
0.1	0.5	9.30	8.95	7.22	4.04	1.62	5.72	6.01	2.57	1.39	77.08	6.2	4.4
0.1	-0.5	9.89	7.39	5.44	4.66	3.45	5.51	4.79	4.12	3.39	77.12	7.0	3.6
		$N = 400$											
$\sigma^2_{\gamma_1}$	γ_2	FULL	AIC	CAIC	FIC	FIC*	AICw	CAICw	FICw	FICw*	CP%	RS%	ET%
0.0	0.0	7.47	1.26	0.23	0.51	0.47	0.75	0.22	0.40	0.34	76.20	1.5	3.6
0.0	0.5	6.81	2.20	1.04	1.95	1.00	1.71	0.93	1.35	0.75	76.18	1.7	3.6
0.0	-0.5	8.12	2.03	3.46	2.75	2.55	2.22	2.68	2.50	2.38	76.20	2.2	4.6
0.1	0.0	6.75	4.08	2.22	1.27	1.00	2.40	2.04	1.04	0.74	77.16	2.2	4.4
0.1	0.5	7.55	7.77	5.74	3.37	1.77	6.31	4.54	2.18	1.16	77.22	2.3	4.3
0.1	-0.5	8.30	4.31	4.48	3.27	2.88	4.04	3.75	3.08	2.73	77.21	3.6	3.5

Table 1: *RMSE% of the estimators of N considered in the simulation study (*w *stands for weighted average estimator) with* $J = 5$ *for* $N = 200$ *and* $N = 400$ *in absence of heterogeneity (*$\beta_1 = -1.1$, $\beta_2 = 0$*).*

		$N = 200$											
$\sigma^2_{\gamma_1}$	γ_2	FULL	AIC	CAIC	FIC	FIC*	AICw	CAICw	FICw	FICw*	CP%	RS%	ET%
0.0	0.0	8.94	1.06	0.50	1.07	0.84	0.91	0.46	0.68	0.59	75.57	4.1	3.4
0.0	0.5	10.00	2.29	2.17	1.88	1.68	1.87	1.75	1.46	1.32	75.49	3.2	5.5
0.0	-0.5	13.09	4.43	5.76	4.59	3.98	3.76	4.48	4.25	3.93	75.53	5.1	4.5
0.1	0.0	8.87	5.60	4.76	2.09	0.86	2.94	3.80	1.43	0.95	76.66	4.8	4.8
0.1	0.5	8.53	8.82	8.19	4.22	1.71	5.91	6.89	2.50	1.35	76.90	3.8	4.2
0.1	-0.5	9.27	6.46	5.12	4.69	4.05	5.21	4.89	4.66	3.69	76.69	6.7	4.0
		$N = 400$											
$\sigma^2_{\gamma_1}$	γ_2	FULL	AIC	CAIC	FIC	FIC*	AICw	CAICw	FICw	FICw*	CP%	RS%	ET%
0.0	0.0	5.50	1.13	0.22	0.62	0.54	0.72	0.22	0.53	0.36	75.50	1.5	3.9
0.0	0.5	5.12	1.77	1.02	3.01	1.12	1.38	0.93	1.68	0.85	75.55	1.4	4.4
0.0	-0.5	5.71	1.93	3.53	2.57	2.37	1.81	2.67	2.29	2.16	75.51	1.5	3.2
0.1	0.0	5.36	3.80	1.96	1.27	0.70	2.47	1.84	1.10	0.70	76.71	2.4	5.2
0.1	0.5	3.80	4.21	6.16	4.13	1.83	3.53	5.46	2.40	1.23	76.49	2.2	5.5
0.1	-0.5	6.58	4.16	4.71	3.16	2.62	3.75	4.02	3.25	2.56	76.74	2.3	4.7

Table 2: *RMSE% of the estimators of N considered in the simulation study (*w *stands for weighted average estimator) with* $J = 7$ *for* $N = 200$ *and* $N = 400$ *in absence of heterogeneity (*$\beta_1 = -1.5$, $\beta_2 = 0$*).*

		$N = 200$											
$\sigma^2_{\gamma_1}$	γ_2	FULL	AIC	CAIC	FIC	FIC*	AICw	CAICw	FICw	FICw*	CP%	RS%	ET%
0.0	0.0	13.73	2.94	1.39	1.67	1.64	1.86	1.07	1.59	1.53	74.64	3.8	28.0
0.0	0.5	10.07	3.20	2.38	2.08	2.14	2.30	1.88	1.86	1.88	74.48	3.5	24.3
0.0	-0.5	9.56	3.74	2.80	3.24	2.66	2.45	1.95	3.02	2.65	74.59	6.7	19.5
0.1	0.0	10.48	4.86	2.88	2.28	1.53	2.82	2.26	1.85	1.65	75.39	6.3	23.2
0.1	0.5	10.25	8.96	6.56	3.41	2.20	5.88	5.37	2.82	1.95	75.44	6.2	27.8
0.1	-0.5	10.91	6.12	3.93	4.38	3.50	4.22	3.13	3.99	3.34	75.53	8.2	22.5
		$N = 400$											
$\sigma^2_{\gamma_1}$	γ_2	FULL	AIC	CAIC	FIC	FIC*	AICw	CAICw	FICw	FICw*	CP%	RS%	ET%
0.0	0.0	5.72	1.90	0.92	1.01	0.98	1.22	0.70	0.87	0.84	74.51	1.3	40.5
0.0	0.5	5.52	2.50	1.29	2.39	1.30	1.90	1.12	1.53	1.13	74.50	1.1	48.9
0.0	-0.5	7.49	2.56	2.01	2.59	2.44	1.88	1.39	2.34	2.23	74.71	1.4	32.3
0.1	0.0	6.80	5.01	2.17	1.43	1.10	2.77	1.66	1.23	0.99	75.21	1.6	40.3
0.1	0.5	7.43	7.00	3.85	3.22	1.86	5.03	3.28	2.15	1.47	75.38	2.2	47.2
0.1	-0.5	5.15	4.83	3.52	3.12	2.25	3.20	2.73	2.68	2.24	75.37	3.2	30.5

Table 3: *RMSE% of the estimators of N considered in the simulation study (*w *stands for weighted average estimator) with* $J = 5$ *for* $N = 200$ *and* $N = 400$ *in the presence of heterogeneity (*$\beta_1 = -1.1$, $\beta_1 = 0.5$*).*

		$N = 200$											
$\sigma^2_{\gamma_1}$	γ_2	FULL	AIC	CAIC	FIC	FIC*	AICw	CAICw	FICw	FICw*	CP%	RS%	ET%
0.0	0.0	8.91	1.90	1.39	1.63	1.57	1.40	1.12	1.45	1.37	73.92	3.9	33.6
0.0	0.5	8.48	2.46	1.95	2.16	1.91	1.87	1.65	1.78	1.69	73.96	3.7	42.5
0.0	-0.5	10.17	3.72	2.62	3.34	3.04	2.51	1.95	3.04	2.91	73.92	4.3	23.2
0.1	0.0	8.29	5.38	3.13	1.98	1.48	3.06	2.57	1.72	1.43	74.95	4.2	34.8
0.1	0.5	8.96	8.21	4.87	3.73	2.12	6.20	4.29	2.37	1.79	74.90	4.1	46.0
0.1	-0.5	9.46	4.84	2.80	3.50	2.96	3.90	2.45	3.30	2.87	75.07	5.6	24.8
		$N = 400$											
$\sigma^2_{\gamma_1}$	γ_2	FULL	AIC	CAIC	FIC	FIC*	AICw	CAICw	FICw	FICw*	CP%	RS%	ET%
0.0	0.0	5.65	1.33	1.02	1.16	0.99	1.04	0.77	1.02	0.85	73.93	1.8	57.2
0.0	0.5	3.88	1.57	1.26	2.92	1.31	1.45	1.13	1.92	1.10	73.90	2.0	65.9
0.0	-0.5	6.37	2.08	2.34	2.72	2.50	1.82	1.55	2.36	2.27	73.88	1.2	41.1
0.1	0.0	3.90	2.97	2.18	1.63	1.21	2.05	1.81	1.17	0.92	74.75	2.5	56.3
0.1	0.5	3.10	3.31	2.99	2.66	1.73	2.73	2.54	1.81	1.23	74.83	1.9	64.5
0.1	-0.5	6.07	3.99	2.67	3.14	2.48	3.21	2.17	2.65	2.54	74.89	1.7	39.3

Table 4: *RMSE% of the estimators of N considered in the simulation study (*w *stands for weighted average estimator) with* $J = 7$ *for* $N = 200$ *and* $N = 400$ *in presence of heterogeneity (*$\beta_1 = -1.5$, $\beta_2 = 0.5$*).*

Tables of application discussed in Section 6

	M_0	M_t	M_b	M_{tb}	M_h	M_{ht}	M_{hb}	M_{htb}	$\hat{N}$
$k_{\mathcal{S}}$	1	5	2	6	2	6	3	7	-
$\ell(\hat{\boldsymbol{\theta}}_{\mathcal{S}})$	-339.02	-336.48	-333.01	-321.95	-314.78	-311.56	-314.22	-310.24	-
$\hat{N}_{\mathcal{S}}$	104.67	104.64	106.79	632.27	123.87	123.67	129.28	153.64	-
AIC	680.03	682.97	670.03	655.90	**633.56**	635.12	634.43	634.48	123.87
CAIC	683.68	701.19	677.32	677.77	**640.85**	656.99	645.36	659.99	123.87
AIC-weights	0.000	0.000	0.000	0.005	0.365	0.167	0.237	0.231	131.99
CAIC-weights	0.000	0.000	0.000	0.000	0.905	0.000	0.095	0.063	124.38

Table 5: *Preliminary results for the meadow voles dataset. For each model $M_{\mathcal{S}}$, $k_{\mathcal{S}}$ is the number of parameters, $\ell(\hat{\boldsymbol{\theta}}_{\mathcal{S}})$ is the maximum log-likelihood and $\hat{N}_{\mathcal{S}}$ is the estimate of the population size. Rows AIC and CAIC contain the value of these indices for each model (minimum values in bold) and the corresponding estimate of N (last column). Rows AIC-weights and CAIC-weights contain the corresponding weights and multimodel estimate of N.*

	M_h	M_{ht}	M_{hb}	M_{htb}	$\hat{N}$
$k_{\mathcal{S}}$	2	6	3	7	-
$\ell(\hat{\boldsymbol{\theta}}_{\mathcal{S}})$	-314.78	-311.56	-314.22	-310.24	-
$\hat{N}_{\mathcal{S}}$	123.87	123.67	129.28	153.64	-
FIC	9.66	9.94	5.13	**4.98**	153.64
FIC*	**4.40**	4.59	4.87	4.98	123.87
FIC-weights	0.000	0.000	0.482	0.518	141.91
FIC*-weights	0.550	0.000	0.000	0.450	137.27

Table 6: *Final results for the meadow voles dataset for the models with heterogeneity, with $k_{\mathcal{S}}$, $\ell(\hat{\boldsymbol{\theta}}_{\mathcal{S}})$ and $\hat{N}_{\mathcal{S}}$ defined as in Table 5. Rows FIC and FIC* contain the value of these indices for each model (minimum value in bold) and the corresponding estimate of N (last column). Rows FIC-weights and FIC*-weights contain the corresponding weights and multimodel estimate of N.*