Model-based adaptive spatial sampling for occurrence map construction

Nathalie Dubois Peyrard Peyrard, Régis Sabbadin, Danny Spring, Barry Brook, Ralph Mac Nally

Stat Comput
DOI 10.1007/s11222-011-9287-3

Model-based adaptive spatial sampling for occurrence map construction

Nathalie Peyrard · Régis Sabbadin · Daniel Spring · Barry Brook · Ralph Mac Nally

Received: 1 September 2010 / Accepted: 28 August 2011

Abstract In many environmental management problems, the construction of occurrence maps of species of interest is a prerequisite to their effective management. However, the construction of occurrence maps is a challenging problem because observations are often costly to obtain (thus incomplete) and noisy (thus imperfect). It is therefore critical to develop tools for designing efficient spatial sampling strategies and for addressing data uncertainty. Adaptive sampling strategies are known to be more efficient than non-adaptive strategies. Here, we develop a model-based adaptive spatial sampling method for the construction of occurrence maps. We apply the method to estimate the occurrence of one of the world's worst invasive species, the red imported fire ant, in and around the city of Brisbane, Australia. Our contribution is threefold: (i) a model of uncertainty about invasion maps using the classical image analysis probabilistic framework of Hidden Markov Random Fields (HMRF), (ii) an original exact method for optimal spatial sampling with HMRF and approximate solution algorithms for this problem, both in the static and adaptive sampling cases, (iii) an empirical evaluation of these methods on simulated problems inspired by the fire ants case study. Our analysis demonstrates that the adaptive strategy can lead to substantial improvement in occurrence mapping.

Keywords Hidden Markov random fields · Optimal sampling approximation · Fire ant sampling for mapping

N. Peyrard (✉) · R. Sabbadin
Unité de Biométrie et Intelligence Artificielle UR875, INRA-Toulouse, BP 52627, 31326 Castanet-Tolosan, France
e-mail: peyrard@toulouse.inra.fr

R. Sabbadin
e-mail: sabbadin@toulouse.inra.fr

D. Spring · R. Mac Nally
School of Biological Sciences, Monash University, Wellington Road, Clayton, Victoria, 3800, Australia

D. Spring
e-mail: spring@sci.monash.edu.au

R. Mac Nally
e-mail: macnally@sci.monash.edu.au

B. Brook
Research Institute for Climate Change and Sustainability, The University of Adelaide, Adelaide, South Australia, 5005, Australia
e-mail: barry.brook@adelaide.edu.au

1 Introduction

In many environmental management problems, estimation of occurrence maps of species of interest, including endangered and invasive species, is a prerequisite to their effective management (Elith and Leathwick 2009). Map estimation is a complex problem because observations are imperfect (detectability of individuals is usually imperfect) and incomplete (it may be infeasible to survey the entire area that might contain individuals). There is often a prohibitive cost of conducting surveillance with perfect sensitivity in all locations that might contain individuals. Therefore, there is a need for methodological tools for designing efficient sampling strategies and for using the resulting imperfect and incomplete observations to estimate occurrence maps.

In adaptive spatial sampling, a set of locations to sample is built sequentially, taking the results of previous sampling steps into account. Such a strategy, which takes into account intermediate observations to monitor sampling, is more efficient than non adaptive methods (Thompson and Seber 1996). In addition, to deal with the uncertainty of the observation, a model-based approach (Gruijter et al.

Published online: 22 September 2011

2006) for sampling should be preferred. Geostatistical models and tools (Chiles and Delfiner 1999), such as kriging, have been applied to model and solve problems of sampling design for map reconstruction (Buesco et al. 1998; Fuentes et al. 2007). However those methods are adapted to continuous data such as pollution levels or temperatures. The application of these tools is not straightforward if the variable to sample and to map is of presence/absence type (1/0 variable), and when observations are noisy (see Bonneau et al. 2010 for a proposition of modeling in the geostatistical framework). In the problem we consider, we are interested in occurrence maps and the only data available are located on a regular grid of spatial sampling units. Therefore, rather than applying commonly used geostatistical models and tools, we propose to adopt a classical image analysis probabilistic framework: Hidden Markov Random Fields (HMRF, Geman and Geman 1984). In addition to being suited to occurrence data on a regular grid of sampling units, another advantage of the HMRF approach is that it can represent dependencies which are not linked to space (for example social networks, transportation networks) while in geostatistics, correlations are strongly linked to the notion of spatial distances.

Image reconstruction from imperfect data is a classical problem tackled by HMRF (Li 1995; Winkler 1995) even with missing data (Blanchet and Vignes 2009). Estimation of HMRF parameters has also been widely studied and efficient algorithms are available (Chalmond 1989; Comer and Delp 2000; Celeux et al. 2003). This model has recently been used in the context of static sampling and spatial decision making when taking into account the value of information (Bhattacharjya et al. 2010). In this article, we propose to use the HMRF framework not only for map construction from an incomplete observation set but also to build efficient adaptive sampling strategies for the purpose of mapping. We present an original model-based adaptive spatial sampling method and we illustrate its performance on a case study focusing on an invasive species management problem. The campaign to eradicate the Red Imported Fire Ant from around Brisbane, Australia, which we considered, involves one of the world's 100 worst invasive species (Lowe et al. 2000). For comparison purposes, we consider both adaptive and static variants of the optimization problem. We use the Maximum Posterior Marginal criterion to measure map quality and sample values. Under this approach, solving the optimization problems (both static and adaptive) requires the evaluation of conditional marginal probabilities for each possible output of each sampling strategy. Those problems are intractable in most realistic circumstances, including those considered here, and, therefore, we propose an approximation of the optimal strategy in both the static and dynamic cases.

The paper is organized as follows. In Sect. 2, we provide background information on the fire ant sampling problem which motivated the methodological work presented in this article. In Sect. 3, we describe the HMRF model that we propose for modeling uncertainty about fire ants occurrence maps. The exact formulation of the optimization problems (static and adaptive) and their approximate resolution are derived in Sect. 4. In Sect. 5, we analyze the performance of the adaptive sampling method and we compare it to the static method and two classical sampling methods, using simulated data inspired by the fire ants problem. We also illustrate map reconstruction on the fire ants mapping problem. Possible extensions of our work are identified in the concluding section (Sect. 6).

2 Fire ants detection problem and data

The red imported fire ant (*Solenopsis invicta*) was first discovered in Australia near Brisbane in February 2001 and the National Fire Ant Eradication Program formally commenced in September 2001. Two forms of treatment are applied. Injection of poison directly into fire ant nests is the method applied when nests are detected with targeted surveillance by trained personnel. The effectiveness of this method depends on the proportion of nests that are detected during surveillance operations. The second method used is to apply a corn-based bait several times across general areas of infestation, with the bait then taken into the nest by foraging individuals. Targeted surveillance activity is conducted primarily in areas near nests detected by private citizens near their residences, business places and public spaces. To distinguish between targeted surveillance and citizen monitoring we refer to those surveillance methods as active and passive surveillance, respectively. Active surveillance is discretionary surveillance, that is, surveillance whose placement is determined by the eradication program manager, the Biosecurity Queensland Control Center (BQCC). In contrast, there is no discretion regarding the placement of citizen monitoring because that form of monitoring occurs primarily in urban areas whose locations are fixed. Since no decision is made on the placement of citizen monitoring, it can be described as a form of passive surveillance.

The method used by BQCC to estimate the current spatial distribution of fire ants in Brisbane is a variant of the Adaptive Cluster Sampling (ACS) method (Thompson and Seber 1996). Nests detected by passive surveillance are used as an initial sample. Then, locations neighboring infected locations in the initial sample are explored. New infected locations are added to the sampling set. Their neighboring locations are sampled and so on until no more nests are found. This is classical ACS. Here, information on the locations

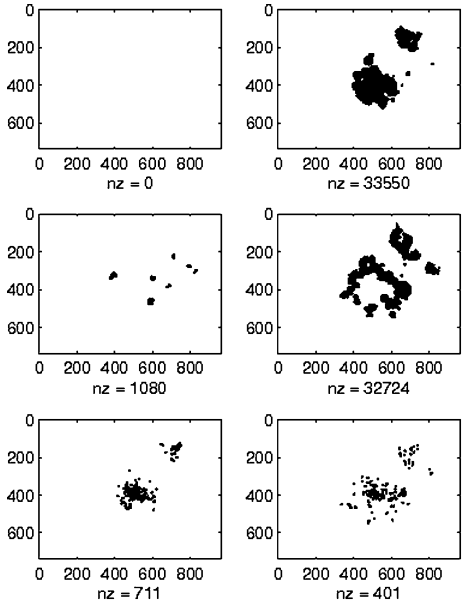

Fig. 1 *Top line*: eradication for years 2000 and 2001 (no eradication in 2000). *Middle line*: search actions for years 2001 and 2002. *Bottom line*: observations for years 2001 and 2002. The value nz indicates the number of non zero cells in the image

where surveillance activity occurred (both passive and active) and information on locations where treatment occurred have also been used to increase the sampling set at each step.

The study region is a 73.7 km × 96.2 km rectangle, including the city of Brisbane and surrounding rural areas around the city. It is represented by a grid of cells of size 100 m × 100 m, thus the complete zone comprises $n = 737 \times 962$ cells. Detection and treatment efforts occurred each year since 2001. The cells which are actively searched during year t are listed in a search action vector, a^t: $a_i^t = 1$ if cell number i was actively searched during year t, and $a_i^t = 0$ otherwise. A list of detected nests is also maintained for each year t. These observations are represented in an observation vector o^t, where $o_i^t = 1$ if ants nests were found in cell i during year t, and $o_i^t = 0$ otherwise. If $o_i^t = 1$, it may be that nests were actively searched for ($a_i^t = 1$), but it is possible as well that they were discovered accidentally ($a_i^t = 0$, passive search). If $o_i^t = 0$, either there were no nests in cell i or they were not detected. Information about treatment actions is also maintained in the form of treatment vectors e^t, where $e_i^t = 1$ if cell i is eradicated at the end of year t, and $e_i^t = 0$ otherwise. A given year, treatment occurs after observation. It is possible to observe $o_i^{t+1} = 1$ even when $e_i^t = 1$, either because the eradication treatment failed or because cell i was colonized again by invasion from the neighboring cells. Figure 1 shows the treatment, search and observation informations for the whole area under study for the first two years of the campaign.

3 A HMRF model of the invasion map

In this section we present our model of uncertainty on invasion map knowledge. This model is based on the HMRF framework (Geman and Geman 1984), which allows to represent the conditional probability distribution of a map, given observations (obtained by sampling). Here and in the following, upper-case letters represent random variables and lower-case letters represent realizations of the same random variables.

In the fire ant problem, a graph $G = (V, E)$ is associated to the n cells of the regular grid dividing the area under study. The set of sites is $V = \{1, \ldots, n\}$ and the set E of edges is defined by the neighborhood system. A first order neighborhood is chosen: for any cell i, the neighborhood $N(i)$ is composed of the four closest cells to cell i (except on the edge of the grid). Other neighborhood systems could be considered: 8-closest cells, or non regular neighborhood systems in the case where an irregular network of locations is considered. The choice of a grid-based first order system is arbitrary and is made for illustrative purpose only. If we were to provide an actual decision-making tool for the management of fire ants, we would have to compare different models. However, this is out of the scope of this paper, which aim is to demonstrate the feasibility of the sampling approach we propose. A random variable X_i is associated to cell i and can take two values: 0 if there are no ants nests in the corresponding cell, 1 if there is at least one. The set $X = \{X_i, i \in 1, \ldots, n\}$ is referred to as the set of hidden variables. The objective is to recover their values from observations. If e is the vector representing the treatment actions applied the year before on all cells, then $P_e(X = x)$ will be modeled as a 2-state Potts model with external field (Wu 1982), defined by:

$$\forall x \in \{0,1\}^n,$$

$$P_e(X = x|\alpha, \beta) = \frac{1}{Z} \exp\left(\sum_{i \in V} \alpha_{e_i} x_i + \sum_{(i,j) \in E} \beta \, \mathrm{eq}(x_i, x_j)\right), \quad (1)$$

where $\mathrm{eq}(x_i, x_j)$ is the Kronecker function, equal to 1 if $x_i = x_j$ and 0 otherwise.

We consider an external field in the Potts model in order to take into account the available information on eradication treatments. Indeed, the eradication treatment applied in year $t-1$ in a given cell is correlated with the presence of ants in the same cell in year t. For a given treatment vector e, values $\alpha_{e_i} \in \alpha = \{\alpha_0, \alpha_1\}$ model different "strength levels" of invasion, depending on whether treatment was performed or not on the cell. We should expect that in treated areas the density of occupied cells is lower than in non-treated areas, modeling causal influence of the treatment. However,

it may happen (see Sect. 5.2) that the density is higher in treated area (which is a "correlation" effect: areas are treated because it is expected that there are nests there, and eradication is not entirely efficient). Note that the external field could also help modelling soil features favoring (or not) ant colonization, etc.

The parameter β, when positive, leads to higher probability for maps x where neighboring cells are in the same state, as expected when there is spatial aggregation of nests. Z is a normalizing constant, ensuring that P_e sums to one. If the state of neighbor cells is known, the probability of a cell infection is independent of the state of the other cells (conditional independences are represented on Fig. 2). If $x_{N(i)} = \{x_j, j \in N(i)\}$, then the conditional distribution is defined by:

$$P_e(X_i = 1 \mid x_{N(i)}, \alpha, \beta) = \frac{\exp(\alpha_{e_i} + \beta N_i^1)}{\exp(\beta N_i^0) + \exp(\alpha_{e_i} + \beta N_i^1)}.$$

N_i^1 counts the number of neighbors of cell i in state 1, while N_i^0 counts those in state 0. They can be computed as $N_i^1 = \sum_{j \in N(i)} x_j$ and $N_i^0 = \text{card}(N(i)) - N_i^1$.

A second variable, O_i is attached to a cell i. It can take values in $\{0, 1\}$ and represents the result of the sampling: an ant nest has been found (1) or not (0) in cell i. A classical assumption in HMRF is that the conditional distribution of observations given hidden variables admits the following decomposition (again, see Fig. 2):

$$P(o \mid x) = \prod_{i \in V} P(o_i \mid x_i). \tag{2}$$

In the fire ants problem, these probabilities depend on whether active or passive search occurred on the cell. This information is represented by the search action vector $a = \{a_i, i \in V\}$ (see Sect. 2). Let $\theta = \{\theta_0, \theta_1\}$ denote the respective probabilities that a nest present in an arbitrary cell i be discovered, either passively or actively, then:

$$\begin{aligned} &P_{a_i}(O_i = 1 \mid X_i = 1, \theta) = \theta_{a_i}, \\ &P_{a_i}(O_i = 1 \mid X_i = 0, \theta) = 0 \quad \text{and} \\ &P_a(O = o|x, \theta) = \prod_{i \in V} P_{a_i}(O_i = o_i \mid x_i, \theta). \end{aligned} \tag{3}$$

Probability θ_1 of discovering a nest after a search action was applied is naturally assumed to be larger than θ_0, the probability if no active search was performed. Expression (3) of the conditional distribution $P_a(O = o|x)$ relies on several assumptions. First, observation probabilities (θ_0, θ_1) are independent of the precise cell which is searched. Then, we assume that observation probabilities do not depend on whether ants were eradicated in the preceding year. We also assume that when the state of a hidden variable is 0, the corresponding observation is 0. Finally, observation conditional

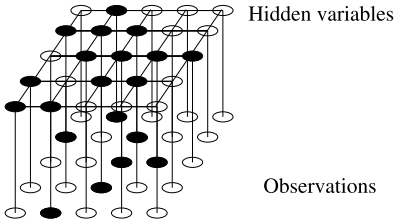

Fig. 2 Hidden Markov random field for the fire ants invasion map, with hidden variables (*top*) and observed variables (*bottom*). Hidden variables can take values 0 (absence, *white sites*) or 1 (presence, *black sites*). Observations can take values 0 (no nest detected, *white sites*) or 1 (nests detected, *black sites*). If the state of a hidden variable is 0, the corresponding observation is 0

probabilities are purely local and do not depend on whether ants nests are present in neighbor cells. The three first assumptions could be relaxed by increasing the number of parameters. Modifying the fourth one would imply changes in the structure of the HMRF and the sampling methods described in Sect. 4 would no longer be applicable as such.

Let us now express the joint distribution of X conditionally to o, a and e. In the fire ants model $P_{a_i}(O_i = 1|X_i = 0, \theta) = 0$ because we assume that there are no false positive observations. Consequently, if $\lambda = \{\alpha, \beta, \theta\}$, $P_{e,a}(x|\lambda, o) = 0$ as soon as there exists i such that $o_i = 1$ and $x_i = 0$. Therefore we can write:

$$P_{e,a}(x|\lambda, o) \propto \left(\prod_{i, o_i = 1} x_i \right) P_e(x|\alpha, \beta) P_a(o|x, \theta). \tag{4}$$

Exploiting (1), (2) and (4), we get

$$\begin{aligned} &P_{e,a}(x|o, \lambda) \\ &= \frac{1}{Z'(\lambda)} \prod_{i, o_i = 1} x_i \times \exp\left(\sum_{i \in V} \alpha_{e_i} x_i \right. \\ &\quad \left. + \beta \sum_{(i,j) \in E} \text{eq}(x_i, x_j) + \sum_{i, o_i = 0} \log(1 - \theta_{a_i}) x_i \right). \end{aligned} \tag{5}$$

where $Z'(\lambda)$ is a normalizing constant, function of the model's parameters λ. Equation (5) defines how the initial knowledge $P_e(x|\lambda)$ about the invasion map is updated when observation actions are applied and resulting observations are taken into account. In this section and in the following ones we omit reference to time, for sake of simplicity. In the above conditional distribution, if x is the hidden map at time t, then e, o and a stand respectively for e^{t-1}, o^t, and a^t.

4 Spatial sampling policies

We now define the problem of designing a spatial sampling policy (strategy) for fire ants map construction as a problem

of optimization under uncertainty. To do so, we first need to define the value of the uncertain knowledge about the actual invasion map, $P = P_{e,a}(x|o,\lambda)$, as well as an estimator of the invasion map associated to this value (Sect. 4.1). The optimization problem can be modeled as non-sequential (static, Sect. 4.2) or sequential (adaptive, Sect. 4.3), depending on the conditions of the search process.

A sampling policy is static if the cells chosen for search are chosen once and for all at the beginning of the year, and active search is limited to them. In the adaptive spatial sampling problem, only a few cells are chosen for active search at the beginning of the year. Then, given the results of the active search in those cells (presence or absence of ants nests), new cells are chosen for active search. This process is repeated until a specified stopping criterion is met (for example the total budget, expressed in terms of number of cells that can be searched, is exhausted). Note that in adaptive spatial sampling problems, cells can be searched more than once, unlike in the static case. In both cases, we assume that a first arbitrary sample (a^0, o^0) is available (for example, a few regularly spaced cells will be sampled before the sampling policy is computed).

4.1 Information value of a map distribution

In spatial sampling problems, it is important to define the "information value" of a probability distribution over maps, describing current knowledge. Sampling strategies will aim at maximizing a criterion based on this information value.

Let us assume that x^* is an unknown map, and that the only available knowledge about x^* is modeled by distribution P. The Maximum Posterior Marginal (MPM) estimator (Besag 1986) of x^* is the configuration x^{MPM} verifying:

$$x^{MPM} = \left\{ x_i^{MPM}, x_i^{MPM} = \arg\max_{x_i} P_i(X_i = x_i) \right\}. \quad (6)$$

The information value of P is defined as $V^{MPM}(P)$, the sum of the marginal probabilities of the most probable state for all sites:

$$V^{MPM}(P) = \sum_{i \in V} \max_{x_i} P_i(X_i = x_i). \quad (7)$$

This value is equal to the expected number of correctly "classified" sites. It is a direct measure of the information value of P. Other information value criteria could be considered, such as the mode of distribution P (Maximum a Posteriori criterion, MAP, Guyon 1995; Li 1995), or its entropy. The former is a valid alternative and the corresponding (static) optimal sampling problem has been studied from a computational complexity perspective (Peyrard et al. 2010). Using MPM does not lead to a simpler computational problem. However, MPM should be more discriminant than MAP since the mode of a joint distribution with large state space may not be very "peaked". We did not consider the entropy criterion since it does not directly lead to an estimator of the hidden map.

4.2 Static spatial sampling

In the static spatial sampling problem, a typical sampling sequence can be decomposed into the following steps:

1. An initial arbitrary sample (a^0, o^0) is performed, which will be used both as prior information and for estimating the HMRF parameters $\lambda = \{\beta, \alpha, \theta\}$.
2. A search action vector a representing the set of cells which will be explored is chosen on the basis of $P_{e,a^0}(x|o^0,\lambda)$. The size of this set is constrained: $|\{i \in V, a_i = 1\}| \leq A_{max}$, where $|\cdot|$ denotes the cardinality of a set and A_{max} the maximum affordable sample size.
3. A set of observations is produced. It can be completed by passive observations, leading to the observation vector o.
4. The a priori knowledge is updated, using (5), providing a new distribution $P_{e,a^0,a}(x|o^0,o,\lambda)$ representing knowledge about fire ants nests presence after sampling.
5. Finally, the MPM value $V^{MPM}(P_{e,a^0,a}(.|o^0,o,\lambda))$ of the new MRF is computed and the corresponding MPM map x^{MPM} is returned.

4.2.1 Exact optimization problem

The optimal sampling strategy will be defined as follows. First, since the results of search actions are not deterministic, a set a of searched cells may result in many different observations o. This implies that the output observations (active and passive) o are only determined through their probability distribution $P_{e,a^0,a}(o|o^0,\lambda)$. The value of a sampling action a can therefore be defined as the expected value U of the updated MRF of step 5, according to that probability distribution:

$$U_{e,a^0,o^0}(a) = \sum_o P_{e,a^0,a}(o \mid o^0,\lambda) V^{MPM}(P_{e,a^0,a}(.|o^0,o,\lambda)). \quad (8)$$

The probability $P_{e,a^0,a}(o \mid o^0,\lambda)$ is obtained as:

$$P_{e,a^0,a}(o \mid o^0) = \sum_x P_a(o|x,\theta) P_{e,a^0}(x|o^0,\lambda).$$

Solving the static spatial sampling problem amounts to finding the sampling vector a^* which maximizes $U_{e,a^0,o^0}(a)$ under constraints $|\{i \in V, a_i = 1\}| \leq A_{max}$.

4.2.2 Approximate static spatial sampling

Computing the static sampling action a^* is infeasible in practice for large problems. When replacing MPM with the

MAP criterion, which does not make the problem more complex, it has been shown that the latter problem is NP-hard (Peyrard et al. 2010). NP-hard optimization problems (Cook 1971) are problems for which it is highly unlikely that efficient solution algorithms (that is with time complexity increasing only polynomially with problem size) can be designed.[1] Computing a^* requires a maximization over the set of possible search action vectors of an expression involving summation over the set of possible observations. Both of these state spaces are of size exponential in the number of sites. In addition, it involves computations of $V^{MPM}(P_{e,a^0,a}(.|\lambda, o^0, o))$ for all pairs (a, o), an operation of exponential complexity as well. Given the size of the problems we wish to address (tens of thousands of cells), we must turn to approximation methods for computing the set of cells that will be explored given the a priori knowledge about invasion.

The approximation method we suggest relies on the following simplifying assumptions:

A1 Current observations are reliable (the state of searched cells is perfectly known after the search) and there are no passive observations, i.e. $\theta_0 = 0$ and $\theta_1 = 1$ (this assumption is made only for the current sampling action to choose, a, and not for the initial observation step (a^0, o^0)).

A2 The states of cells are independent given initial sampling results. This leads to the following approximation:

$$P_{e,a^0}(x|o^0, \lambda) \sim \prod_{i=1}^{n} P_{e,a^0}(X_i = x_i|o^0, \lambda)$$

where $P_{e,a^0}(X_i = x_i|\lambda, o^0)$ is the marginal distribution of the resulting MRF on cell i, given initial observation result (a^0, o^0).

With these two assumptions, it can be shown that optimizing a spatial sample amounts to choosing the cells whose marginal occupation probabilities $P_{e,a^0}(X_i = x_i|o^0, \lambda)$ are the closest to 0.5, that is, the cells whose occupation status is most uncertain (a proof is given in the Appendix). Computing exactly a marginal occupation probability is costly since it involves the marginalization of the joint distribution (5) over all variables except x_i. This cannot be performed in reasonable time. Therefore, we use a belief propagation algorithm (Pearl 1988; Yedidia et al. 2000) in order to approximate these marginal probabilities. This algorithm requires only a time polynomial in the number of cells to compute the approximate marginals.

[1]Proving impossibility of this fact is one of the seven Millennium problems proposed by the Clay institute: http://www.claymath.org/millennium/P_vs_NP/.

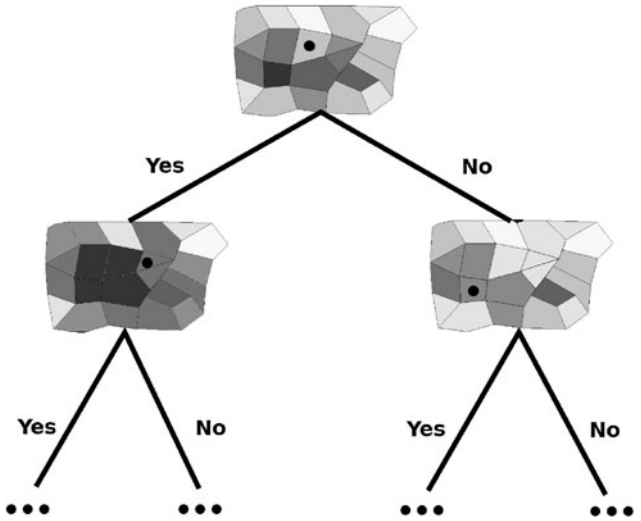

Fig. 3 Part of an adaptive sampling strategy. *Levels of gray* represent estimates of marginal occupation probabilities. *Black dots* represent the current cell chosen for exploration

4.3 Adaptive spatial sampling

In the adaptive spatial sampling problem, we assume that the A_{max} cells we explore can be decomposed into successive small groups, the next one being chosen taking into account observations of previously sampled cells. For illustration purpose, we describe exact adaptive sampling in the case where one cell is chosen (and explored) at each step. Thus the number of steps is exactly A_{max}. For this particular case, one step of adaptive sampling is represented on Fig. 3. One cell is chosen for exploration (black dot in the top figure) and then, depending on whether ants are detected (Yes branch) or not (No branch), the MRF is updated in different ways. Therefore, the next cell to explore according to the strategy can be different (black dots in bottom maps). In the following, since the action vector a contains only one cell in state one, it will be identified to the index of that cell ($a \in V$). Similarly, o is identified to the value (0 or 1) observed on that cell.

4.3.1 Adaptive sampling strategy

As Fig. 3 suggests, an adaptive sampling strategy may well lead to many different sets of cells being sampled, depending on the observations obtained. Thus the sampling strategy can no more be represented as a subset a of V of size A_{max}. It is now a tree, δ, which vertices are cells chosen for sampling and edges represent observations (0/1 or Yes/No outputs when a single cell is sampled). A part of such a tree is represented in Fig. 4. Let $a^k, 1 \leq k \leq A_{max}$ denote the cell which is explored during the k^{th} sampling phase: a^k is chosen as a function of past samples results $(o^1, \ldots, o^{k-1})$. From δ we can define δ_k, a function specifying the k^{th} cell to

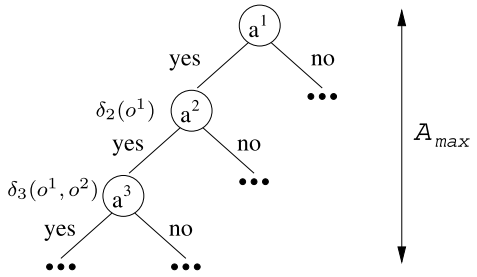

Fig. 4 Part of an adaptive sampling strategy tree

sample, as a function of the $k-1$ observations which were obtained from past sampling steps. For example, on Fig. 4, $a^3 = \delta_3(o^1, o^2)$.

4.3.2 Optimal sampling strategy computation

As in the static case, an initial arbitrary sample (a^0, o^0) is used as prior information. The value of an adaptive sampling policy is defined by extension of the value of a static sampling policy (see (8)), taking into account the fact that δ is a tree:

$$U^{A_{max}}_{e,a^0,o^0}(\delta) = \sum_{(o^1 \dots o^{A_{max}}) \in \tau_\delta} P_{e,a^0,\delta}(o^1 \dots o^{A_{max}} \mid o^0, \lambda) \times V^{MPM}(P_{e,a^0,\delta}(.|\lambda, o^0, o^1, \dots, o^{A_{max}})). \quad (9)$$

In (9), τ_δ denotes the set of possible observation sequences given δ, i.e. the set of paths from the root to a leaf of the policy tree. The knowledge of δ enables to recover the sequence of sampled cells: $P_{e,\delta}(o^1 \dots o^{A_{max}} \mid \lambda, a^0, o^0) =_{def} P_{e,a^0,a^1,\dots,a^{A_{max}}}(o^0, o^1 \dots o^{A_{max}} \mid \lambda)$, with $a^i = \delta_i(o^1, \dots, o^{i-1})$ the action defined by the sampling policy at step i, given past observations. More precisely, $\delta_i(o^1, \dots, o^{i-1})$ can be read from the policy tree representation of δ, as the last node of the partial branch defined by $o^1, \dots, o^{i-1}$.

Since multiple samplings at a same site are possible in adaptive sampling, $P_{e,a^0,\delta}(x|\lambda, o^0, o^1, \dots, o^{A_{max}})$ is obtained from a slight modification of (5), taking into account repeated samplings of cells:

$$P_{e,a^0,\delta}(x|\lambda, o^0, o^1, \dots, o^{A_{max}}) \propto \prod_{h \in S_1} x_{a^h} \times \exp\left(\sum_{i \in V} \alpha_{e_i} x_i + \beta \sum_{(i,j) \in E} \mathrm{eq}(x_i, x_j) + \sum_{h \in S_0} \log(1 - \theta_{a^h}) x_{a^h} \right),$$

where $S_0 = \{h, 0 \le h \le A_{max} \text{ and } o^h = 0\}$ and $S_1 = \{h, 0 \le h \le A_{max} \text{ and } o^h = 1\}$ are respectively the sets of observation steps h where the sampled cell a^h was found unoccupied or occupied. In the set S_0 (or S_1) a same cell index can appear more than once if the correspond cell is explored several times.

The problem of optimizing δ with A_{max} cells to sample can be solved recursively *backwards*, noting that

$$U^{A_{max}}_{e,a^0,o^0}(\delta^*) = \max_{a^1} \left\{ \sum_{o^1} P_{e,a^0,a^1}(o^1|o^0, \lambda) \times U^{A_{max}-1}_{e,a^0,o^0,a^1,o^1}(\delta^*_{|a^1,o^1}) \right\},$$

where δ^* is an optimal policy (we have $U^{A_{max}}_{e,a^0,o^0}(\delta^*) \ge U^{A_{max}}_{e,a^0,o^0}(\delta), \forall \delta$) and $\delta^*_{|a^1,o^1}$ is the optimal sub-policy computed from the HMRF resulting from observations o^0, o^1 and with sampling budget $A_{max} - 1$.

Of course, the recursive algorithm explores a solution space of exponential size, which makes it unsuitable to solve realistic problems. This is all the more true if we can explore more than one cell in each sampling step. In the following section, we propose an approximate sampling algorithm which relies on the static sampling approximate algorithm and directly applies to the case where more than one cell is sampled at a time.

4.3.3 Approximate adaptive sampling

For our approximate adaptive algorithm, we propose to use a greedy algorithm (as is usually done in heuristic search problems), in conjunction with the approximation approach of the static sampling case. The set of cells a^k (now, a^k represents a set of cells indices and not a single index) which will be sampled during sample phase k will be computed on-line, by applying the method of Sect. 4.2.2 and considering that the initial sample is the sequence $(a^0, o^0, a^1, o^1, \dots, a^{k-1}, o^{k-1})$ of actions/observations obtained so far. More precisely the procedure is:

1. An initial arbitrary sample (a^0, o^0) is performed, from which the model parameter λ is estimated.
2. Evaluate the marginal probabilities for the conditional distribution $P_{e,a^0}(x|\lambda, o^0)$.
3. Explore the cells whose marginal probabilities are the closest to 0.5. This leads to (a^1, o^1).
4. Update the sampling informations: $(a^0, o^0) \leftarrow (a^0, o^0, a^1, o^1)$.
5. Go to step 2 while the number of sampled cells is less than A_{max}.

When we consider only two successive sample phases, this on-line procedure can be related to the two-phase adaptive method for optimal spatial sampling proposed by Chao and Thompson (2001) in the case of log-normal perfectly observable variables and a mean square error criterion.

5 Validation of the model-based sampling methods

In this section, we present a validation of the heuristic sampling approaches on simulated data. This validation can only be performed on simulated data since as far as real data is concerned, no validation with respect to the "true" invasion status of cells is possible, this "true" status being unobserved. However, the method is validated on simulated problems with various parameters sets, covering the range of likely parameters values for the fire ant problem. An illustration of parameters estimation and map reconstruction based on the available fire ant data (Sect. 5.2) is also presented.

5.1 Evaluation of the heuristic sampling methods

In order to evaluate the relative performances of the static and adaptive heuristic sampling methods we compared the methods using simulated data generated by a HMRF model whose parameters (α, β) were unknown to the sampling algorithms (Sect. 5.1.1). Since the ACS sampling method (Thompson and Seber 1996, and Sect. 2) was used to collect the fire ant data set, the static and adaptive heuristic sampling methods were compared with the ACS method. A comparison was also made with the purely random sampling method. ACS is a method originally developed for estimating global characteristics of spatially distributed populations under the hypothesis of perfect observation ($\theta = (0, 1)$). The random sampling method (Thompson and Seber 1996) consists in selecting a fixed number of cells to observe (in a non-adaptive way), with each cell having the same probability of being selected.

The evaluations presented below include a parameters estimation step. It is performed using the Simulated Field EM algorithm (SF-EM, Celeux et al. 2003), an approximation of the EM algorithm for parameters estimation in HMRF. In SF-EM, at each iteration, the MRF distribution is replaced by one of independent variables, built by setting the state of the neighborhood of each cell to a simulated value. We observed good performance of the SF-EM algorithm for the parameters values corresponding to an established epidemic, that is when the correlation coefficient β is high. On the contrary, when β is low, a low incidence (α) with high probability of detection (θ) is confused with a higher incidence with low probability of detection. This parameters identifiability problem can be easily intuitively understood: when variables are independent ($\beta = 0$) and we have no idea on incidence and detectability (α and θ), it is impossible to distinguish, in the model expression, between a highly incident but difficult to detect process and a low incidence, easily detected one. Of course, estimation of α poses no problem when θ is not estimated but fixed at its true value. In the fire ants problem, expert estimations of the detection probabilities with active and passive search are available, therefore, we assumed in the following experiments that θ was known and did not have to be estimated.

5.1.1 Comparison procedure

In order to compare the four sampling methods, we considered eight configurations for (α, β, θ), reported in Table 1. This corresponds to four different choices for (α, β), and for each choice, two values of θ were considered. In Fig. 5, a realization of the hidden Potts model for each of the four parameter choices is presented. The experimental protocol was the following. For each set of parameters values we simulated ten different hidden maps of 50×50 cells, according to $P(x)$. The grid was divided into four equal squares and treatment efforts were applied to the top-left and the bottom-right squares. For each map we started by applying an arbitrary regular sampling action a^0 comprising around 10% of the total number of cells (see first image of Fig. 7, top) and then simulated an observation set o^0 according to the hidden map and θ. This initial sample was used to compute an estimate $(\tilde{\alpha}, \tilde{\beta})$ of the MRF parameters, using the SF-EM algorithm. The same estimate was used for the static and adaptive heuristic methods (these estimates are also used in the random sampling and the ACS procedure but only in the map restoration step). Then, for each of the ten initial samples (x, a^0, o^0) we ran the four methods (heuristic static, heuristic adaptive, ACS and random) five times. The number of cells that could be sampled by the static heuristic method varied from 5% to 90% of the total number of cells. For the adaptive heuristic method, a maximum of 5% of the cells could be sampled at each time step and there were a maximum of 18 sampling steps, implying that a maximum of 90% of the cells could be sampled. Under the ACS method, the number of cells sampled during each sampling phase is not fixed in advance, nor is the total number of cells sampled. The random approach sampled from 5% to 90% of the total number of cells, as in the heuristic static method. After the list of sampled sites is established, the corresponding observations are simulated. Based on all observations (o^0 plus the ones obtained after sampling), the MPM restoration of the map is computed ((6) with $P(x)$ updated as described in Sect. 3 or Sect. 4.3.2). We compared, for each method, the average proportion of (i) misclassified empty cells (cells where there are no nests, incorrectly classified as occupied) (ii) misclassified occupied cells (invaded cells incorrectly classified as empty) (iii) misclassified cells (cells which are incorrectly classified as either invaded or empty).

5.1.2 Comparison of the methods performances

The parameters (α, β, θ), as well as the budget allocated to sampling influence the performances of the four methods. For configurations 3 and 4, none of the methods are efficient

Table 1 The eight configurations of (α, β, θ) tested

Config	α	β	θ
1	$(0, -2)$	0.8	$(0.5, 0.8)$
2	$(0, -2)$	0.8	$(0, 0.8)$
3	$(-2, -3)$	0.2	$(0.5, 0.8)$
4	$(-2, -3)$	0.2	$(0, 0.8)$
5	$(0, 0)$	0.5	$(0.5, 0.8)$
6	$(0, 0)$	0.5	$(0, 0.8)$
7	$(1, -1)$	0.4	$(0.5, 0.8)$
8	$(1, -1)$	0.4	$(0, 0.8)$

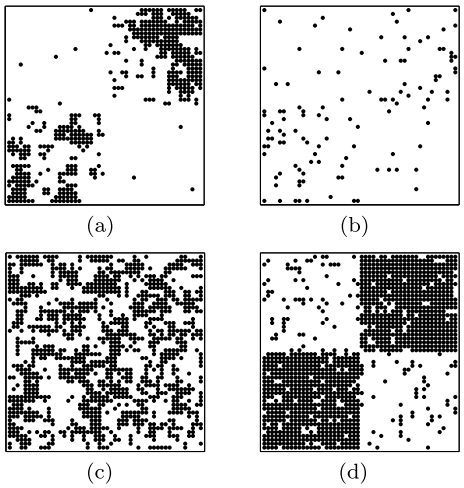

Fig. 5 Realizations of a two-state Potts model with external field on a 50×50 grid (obtained for 10000 iteration of the Gibbs Sampling). α_0 (resp. α_1) is attached to the *top-right* and the *bottom-left squares* (resp. to the *top-left* and the *bottom-right squares*) of the grid. (**a**) $\alpha = (0, -2), \beta = 0.8$, (**b**) $\alpha = (-2, -3), \beta = 0.2$, (**c**) $\alpha = (0, 0), \beta = 0.5$, (**d**) $\alpha = (1, -1), \beta = 0.4$

in reconstructing the map since the proportion of occupied cells is very low. For the other configurations, several general qualitative conclusions can be made. We discuss them and present numerical results for configurations 2, 6 and 8 (Fig. 6). The changes observed when θ_0 increases from 0 to 0.5 are discussed at the end of this section.

First, the ACS method is clearly dominated by the three other sampling methods in terms of quality of the restored invasion map. The ACS method is not designed to reconstruct maps of spatial processes, but rather to estimate global statistics of these processes, such as average densities of occupation. Thus this poor performance is not surprising. The random approach is dominated by the two model-based heuristic approaches. When sampling resources (percentage of cells sampled) increase, results of random sampling become closer to those of heuristic static sampling because in both cases almost all cells are sampled.

Another general conclusion is that the heuristic adaptive sampling method has superior performance than the static method, with the difference being small in two specific situations: low sampling resource and low spatial structure. First, when the sampling budget is low, the adaptive method selected cells to explore based on similar information to that which was available under the static approach, and, therefore, explored similar or identical cells. If the sampling budget is large enough, the adaptive sampling approach can exploit the first observations that were made, while the static approach does not. Therefore, under the adaptive approach, exploration is more informed, leading to a strategy for space exploration different to that of the static method. This is demonstrated in Fig. 7 representing the locations of sampled sites and the corresponding observations respectively for the heuristic static method and the heuristic adaptive method, for configuration 8 ($\alpha = (1, -1), \beta = 0.4, \theta = (0, 0.8)$). In the heuristic static approach, whatever the percentage of area sampled, the only information used is that illustrated on the top left image (initial arbitrary regular sample), while in the heuristic adaptive method, for a given percentage of sampled area, information on all intermediate images was also used. Under an adaptive strategy, it can be more informative to revisit a site that was previously sampled, if uncertainty remains high on this site, than to systematically explore new cells. The resulting estimated marginal probabilities of presence for a sampling size of 90% of the whole area are displayed on Fig. 8. In that case, despite the large sample size, uncertainty remains substantially higher with static than with adaptive sampling. The latter strategy eventually leads to an improved restoration of the hidden process. This example also illustrates that for both heuristic methods, sampling is preferably performed near detected occupied sites in low density areas (the top left and bottom right squares of the area under study are explored first in configuration 8): in these areas, a sampled cell with $o_i = 0$ has only few neighbor cells with $o_j = 1$: enough to maintain uncertainty (was presence missed or is it a true absence?) but not enough to influence belief strongly towards $x_i = 1$.

We also observed (Fig. 6) that the difference in performance between the heuristic static and the heuristic adaptive method increases with the hidden map structure. Both methods lead to similar results in configuration 6 ($\alpha = (0, 0), \beta = 0.5, \theta = (0, 0.8)$), but if the value of the spatial parameter β is increased then the adaptive method outperforms the static one (results not shown). Because treatment actions are applied to create a chessboard pattern if the difference of weights $(\alpha_1 - \alpha_0)$ increases, this creates large scale structure in the map. In that case we observed better results for the adaptive method.

Finally, when $\theta_0 = 0.5$, the number of invaded cells found after the initial arbitrary sample will be higher than when $\theta_0 = 0$, because it includes cells in passively sampled areas.

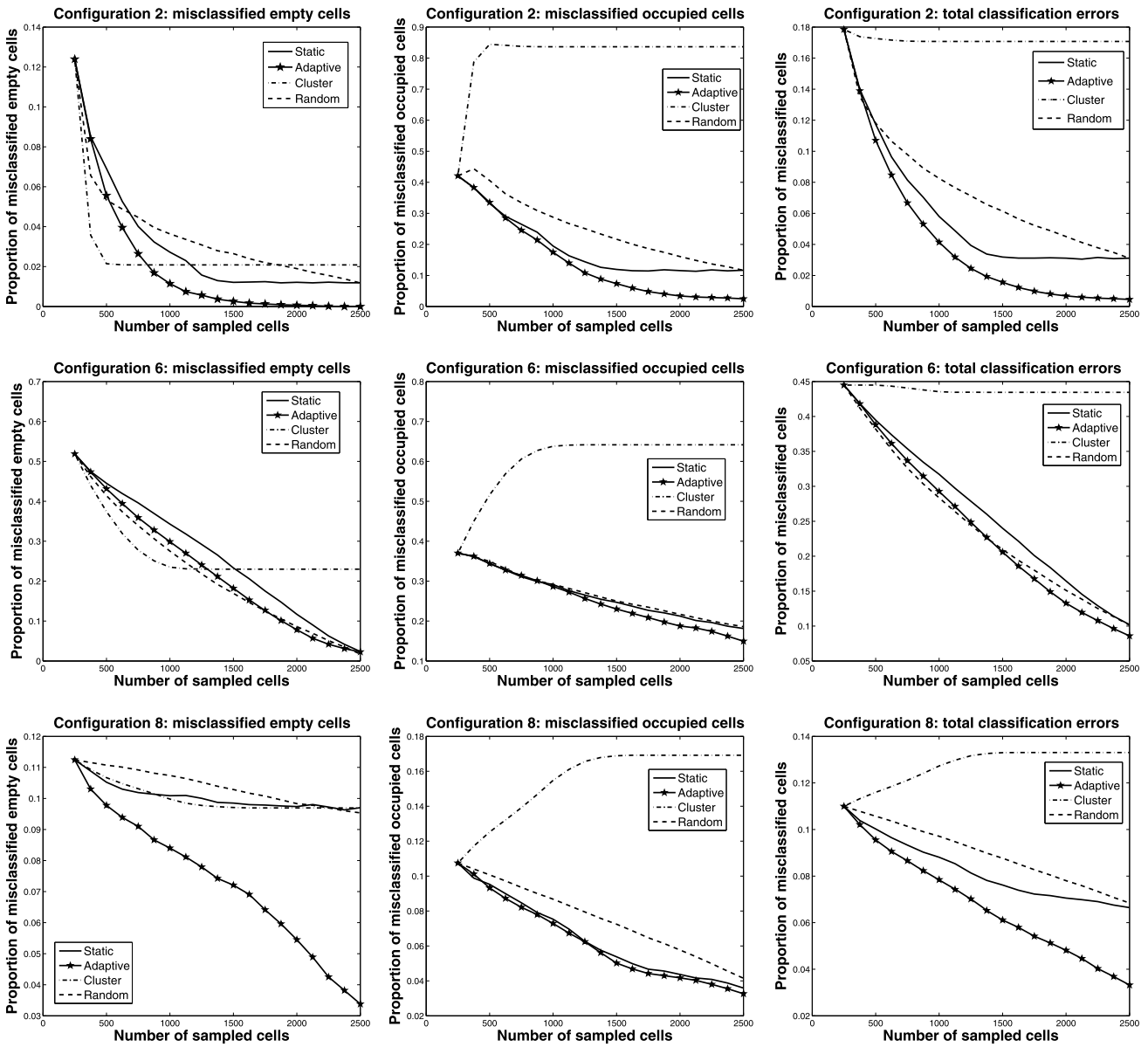

Fig. 6 Errors rates for the different sampling strategies and for different model parameters. *Left*: proportion of misclassified empty cells in the map restoration for the four sampling methods tested, *middle*: proportion of misclassified occupied cells, *right*: proportion of misclassified cells. *From top to bottom rows*, configurations 2, 6, and 8. Average number of infected cells are respectively 459.3, 1243 and 1249.6

The consequence of that, as expected, is that the classification errors will be lower. However, the conclusions on relative performances of the four methods are not significantly altered by the choice of θ_0.

5.2 Fire ants illustrative case study

We applied the methods for parameters estimation (SF-EM) and map reconstruction (MPM) based on the sampling actions actually applied and the resulting observations. We selected a sub-grid of the entire study region to ensure there was sufficient information in the sample. The region was selected on the basis of its low proportion of rural areas (where detection by passive search is estimated to be close to zero). Only years 2001, 2002 and 2003 were considered in the data set. Subsequent years (2004 to 2007) do not present enough detected nests to reliably fit a spatial model. Therefore it is pointless to apply our method to these. Years 2008 and 2009 show a new rise in infected cells, but not much structure. The selected grid was composed of $100 \times 100 = 10000$ cells. The statistics are summarized in Table 2 and the treatment actions, search actions and observed nests are illustrated in

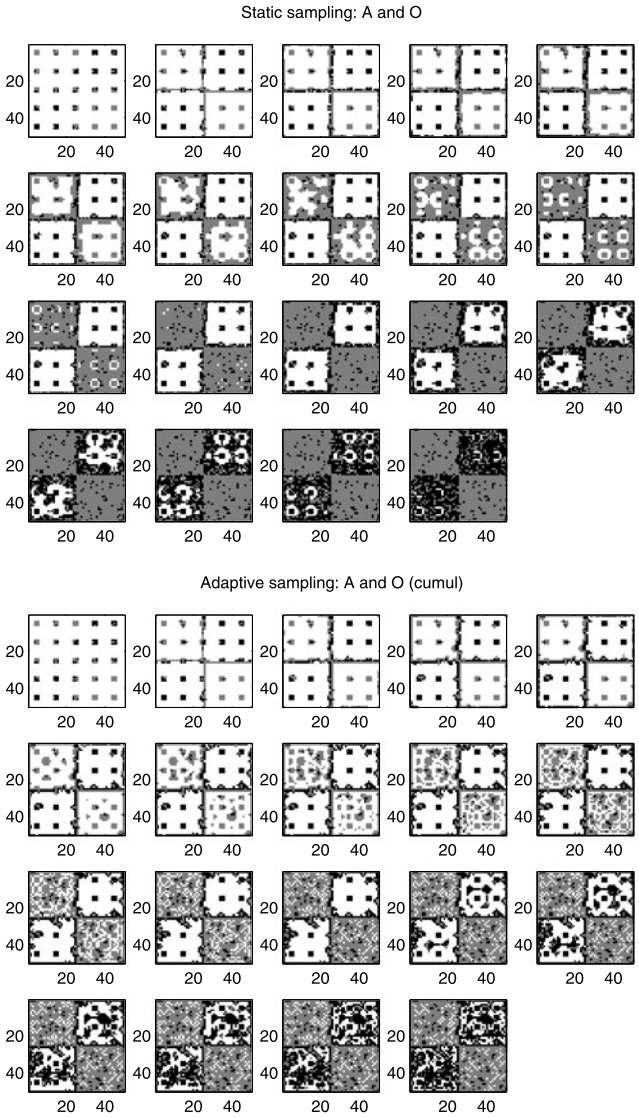

Fig. 7 Cumulative sampling locations and realization of corresponding observations for the heuristic static (*top*) and adaptive (*bottom*) methods. Observations were simulated for $\theta = (0, 0.8)$. The *first top-left images* corresponds to the initial regular sampling, then *from left to right* and *top to bottom images* correspond to a sample size increasing from 5% to 90 % of the whole area

Fig. 9. From Table 2 we can see that the number of actively searched cells increases with time and that the percentage of cells with observed nests is initially significant but declines with time. We recall that the eradication vector used in the HMRF model of year t is e^{t-1}.

Three HMRF models have been estimated, using treatment, sampling and observation data for years 2001 to 2003. The SF-EM algorithm was initialized with the following values: $\alpha = (0, -1)$ and $\beta = 0.5$. Parameter θ was not estimated, but fixed to the following "plausible" values: $\theta = (0.5, 0.8)$ in urban areas and $\theta = (0.01, 0.8)$ in rural areas. The value of θ_0 was not set to zero in rural areas, in order to account for the passive observations of nests which actually occurred, even though rare. The parameters estimation for the 3 years considered are reported in Table 3. In 2001, α_1 cannot be estimated (and is arbitrary fixed to -1) because no treatment was applied in 2000. In 2002 and 2003, the orderings of α_0 (without eradication) and α_1 (with eradication) are consistent with the orderings of the proportions of occupied cells in areas with and without treatment (see Table 3, bottom). The two estimations of α_0 in 2002 and in 2003 are also in agreement with the proportions of cells with observed nests in the area without treatment, namely 20% and 8 % in 2002 and 2003. This proportion was equal to 33% in 2001.

Figure 10 shows a restoration of the 2002 invasion map, as well as the estimated marginals probabilities of occupation based on the sole data a, o and e and of estimated parameters values $(\alpha_0, \alpha_1, \beta)$ (listed in Table 3). The restored map is a smoothed version of the observation map o, with clusters of occupied cells of larger size: After the restoration, 369 cells are considered likely to be invaded (marginal occupation probability greater than 0.5) while nests were only observed in 189 cells.

Table 2 Number and percentage of cells with active search, observed nests and eradication for year 2001 to 2003 on the sub-grid selected for analysis

	2001	2002	2003
$A = 1$	0	656	3593
%	0	6.4307	35.2220
$O = 1$	340	189	109
%	3.3330	1.8528	1.0685
$E = 1$	7548	9473	10142
%	73.9927	92.8634	99.4216

Table 3 Top: estimation of the HMRF parameters of the fire ants model. Bottom: percentage of observed nest in areas with and without treatment (right) on the sub-grid selected for analysis

	2001	2002	2003
α_0	0.0006	0.3907	−0.7548
α_1	−1	0.1867	0.1299
β	1.1619	1.3810	1.2641

	2002	2003
$e = 0$	1.9224	0.8242
$e = 1$	1.8283	1.0873

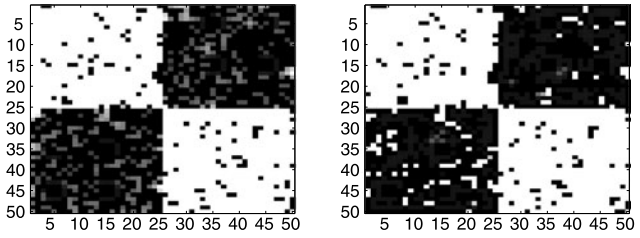

Fig. 8 Estimations of the marginals probabilities of presence for a sample size of 90% of the whole area (blackness increases with the probability of presence). *Left*, heuristic static sampling; *right*, heuristic adaptive sampling

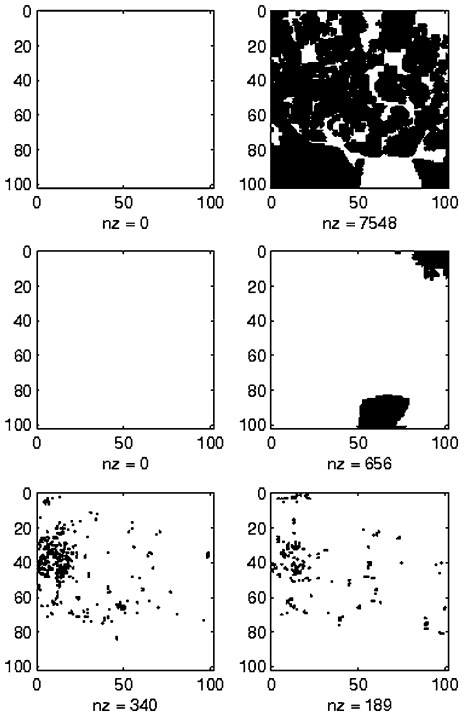

Fig. 9 *Top line*: eradication for years 2000 and 2001 (no eradication in 2000). *Middle line*: search actions for years 2001 and 2002. *Bottom line*: observations for years 2001 and 2002

6 Concluding remarks and discussion

In this article, we have presented an original method for designing approximate sampling strategies for estimating occurrence maps of spatial processes. The main innovation of our approach is that it is a model-based approach which embeds the objective of map reconstruction in the sample selection criterion. We formulated the problem within the HMRF framework (Geman and Geman 1984; Guyon 1995; Li 1995), the classical framework used in image analysis problems. More precisely, we formulated the problem of selecting sampling strategies as a combinatorial optimization problem in which the expectation of the value of the possible resulting MRF is to be maximized. We formulated static and adaptive versions of that approach. In practice both are

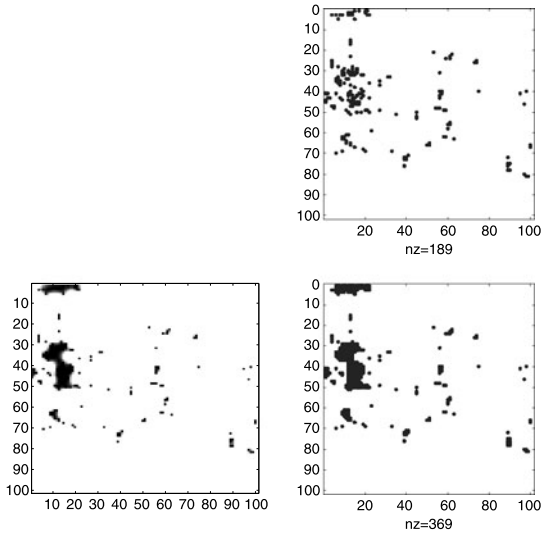

Fig. 10 Observed nests in 2002 (*top*), marginals (*bottom left*) and restored invasion map (*bottom right*)

too complex to be applied directly to problems of realistic size and, therefore, we proposed approximate variants of those methods. We simplified the methods in two ways: (i) approximating the computation of marginal probabilities by using the belief-propagation algorithm (Pearl 1988) and (ii) replacing the exact optimization problems (static and adaptive) with the computation of simpler criteria based on those approximate marginal probabilities.

Theoretical validation (for example, distance to the value of the true optimal sample) of the heuristic static and adaptive approaches remains difficult. Here, we presented an empirical validation approach based on simulated data. Our study demonstrated the superiority of the model-based approach over two standard sampling methods (random sampling and adaptive cluster sampling (Thompson and Seber 1996)) when the two following conditions are fulfilled: (i) in circumstances where spatial structure is present, as in our fire ants case study, and (ii) provided that sufficient sampling resources are available (at least 10% of the total area). We should insist on the fact that the ACS method and our methods do have different goals, even though they can be used interchangeably. Our methods aim at providing the most accurate map of a given spatially structured stochastic process, when bounded sampling resources are available. We illustrated their use with binary variables, but finite domain variables could be considered as well (at the price of increased complexity and certainly decreased performance, in practice). ACS, on the other hand, is dedicated to binary variables, and the sampling resource is not bounded a priori (while there still are newly found infected cells, sampling goes on). The ACS method does not provide a reconstruction tool for the whole map (it only classifies cells into three categories : occupied, empty and unexplored). For our comparison, we equipped the ACS method with the same

HMRF-based reconstruction tool as ours and we noticed that the reconstructed maps are of poorer quality than the ones reconstructed after our sampling methods are applied. This result is expected, since our heuristic methods aim at approaching "optimal" sampling for reconstruction, which is not the objective of the ACS method. In order to compare the efficiency of both types of method for eradication and not only for mapping, we could add, at the end of the HMRF sampling phase, an "eradication phase", consisting in eradicating all cells with marginal probability of occupation exceeding a given threshold (0.5, for example). However, this was not the objective of this study and is left for further research.

In our study we took constraints on resources into account only through a limit on the sample size. Constraints can be more complex: the cost of a sample could be related to the time spent during exploration. In adaptive sampling fixed sampling costs could be incurred whenever a new sampling phase starts, etc. Our assumption was that sampling costs are negligible compared to the cost of mapping errors. Introducing such costs in the optimization problem and evaluating the impact on the sampling designs remain open questions which are of crucial interest in environmental management problems. One question is of course how to scale costs and map quality.

This work is one of the first attempts to combine HMRF modeling and tools for sequential decision making under uncertainty in order to solve optimal sampling problems for occurrence map reconstruction. Our proposed method led to substantial improvement compared to classical design-based sampling methods, even with the simple approximation we used in this paper. These results confirm our approach is promising, particularly given that several improvements could be considered that are expected to strengthen the approach.

The two heuristic approaches we have presented can be improved in two different ways. Optimization can be improved. The spatial sampling problems we tackled are too complex to solve exactly. The approximation we proposed is the simplest and, a priori, least efficient, in the family of approximate algorithms that could be applied to sampling problems involving stochasticity (Spall 2003). A natural direction to derive more efficient algorithms is the exploration of simulation-based optimization methods. We are currently studying solutions using Reinforcement-Learning algorithms (Sutton and Barto 1998), which have been successful in the resolution of optimal sequential planning problems.

Parameters estimation can also be improved, in two different ways. In the adaptive version, data obtained during the sampling process can be used to improve the current estimation of the HMRF model. Thus, alternation of sampling and estimation phases would improve the method. In our case study, fire ant data are available for successive years of treatment, sample actions and nests observations. This information could also be taken into account to improve parameters estimation, provided that knowledge about the temporal dynamics of the ants propagation is available.

Acknowledgements The authors acknowledge the support of the Australian Research Council ("Discovery project" N° DP0771672).

Appendix

We demonstrate here that the approximate solution algorithm for static spatial sampling presented in Sect. 4.2.2 provides the exact solution when the HMRF model satisfies assumptions A1 and A2.

Let us recall the definition of V^{MPM}:

$$V^{MPM}(P_{e,a^0,a}(X|\lambda, o^0, o)) = \sum_{i=1}^{n} \max_{x_i} P_{e,a^0,a}(X_i = x_i \mid \lambda, o^0, o).$$

If we assume that current observations o, obtained after sampling actions a are reliable and that there is no passive observation (A1), and denoting $x_a = \{o_i : a_i = 1\}$, we have

$$\begin{aligned}
&\sum_{o} P_{e,a^0,a}(o \mid o^0, \lambda) V^{MPM}(P_{e,a^0,a}(X \mid o^0, o, \lambda)) \\
&= \sum_{x_a} P_{e,a^0,a}(x_a|o^0, \lambda) \\
&\quad \times \sum_{i=1}^{n} \max_{x_i} P_{e,a^0,a}(X_i = x_i \mid o^0, x_a, \lambda) \\
&= \sum_{i=1}^{n} \sum_{x_a} P_{ea^0,a}(x_a|o^0, \lambda) \\
&\quad \times \max_{x_i} P_{e,a^0,a}(X_i = x_i \mid o^0, x_a, \lambda).
\end{aligned}$$

If $a_i = 1$, then $\max_{x_i} P_{e,a^0,a}(X_i = x_i \mid o^0, x_a, \lambda) = 1$ (cell i has been observed and observation was reliable). If $a_i = 0$, from A2 x_i is independent of x_a conditionally to o^0 so that $P_{e,a^0,a}(X_i = x_i \mid o^0, x_a, \lambda) = P_{e,a^0,a}(X_i = x_i \mid o^0, \lambda)$. Finally, under A1 and A2:

$$\begin{aligned}
&\sum_{o} P_{e,a^0,a}(o|o^0, \lambda) V^{MPM}(P_{e,a^0,a}(X|\lambda, o^0, o)) \\
&\sim \sum_{i=1}^{n} v_i(a_i),
\end{aligned}$$

where

If $a_i = 0$, $v_i(a_i) = \max_{x_i} P_{e,a^0}(X_i = x_i|o^0, \lambda)$.

If $a_i = 1, v_i(a_i) = 1$.

The corresponding approximation $\tilde{a}(e, \lambda, a^0, o^0)$ of $a^*(e, \lambda, a^0, o^0)$ satisfies

$$\forall i, \quad \tilde{a}_i = 1 \quad \text{if } -c_i(1) + 1 > \max(v_i(0), 1 - v_i(0)) \quad (10)$$

which is equivalent to

$$c_i(1) < 1 - \max(v_i(0), 1 - v_i(0)) = \min(v_i(0), 1 - v_i(0)).$$

Computing $\tilde{a}(e, \lambda, a^0, o^0)$ defined in (10) consists in practice in ranking the cells i in decreasing order of $\{\nu(i) = \min(v_i(0), 1 - v_i(0)) - c_i(1)\}$. Then, all the cells with positive value $\nu(i)$ are sampled if sampling resources are sufficient. Fewer cells are sampled if sampling resources are not sufficient, the cells with higher heuristic values being sampled in priority, since the heuristic function models their contribution to map uncertainty reduction.

References

Besag, J.: On the statistical analysis of dirty pictures. Journal of the Royal Statistical Society B **48**(3), 259–302 (1986)

Bhattacharjya, D., Eidsvik, J., Mukerji, T.: The value of information in spatial decision making. Mathematical Geosciences **42**(2), 141–163 (2010)

Blanchet, J., Vignes, M.: A model-based approach to gene clustering with missing observations reconstruction in a Markov random field framework. Journal of Computational Biology **16**(3), 475–486 (2009)

Bonneau, M., Peyrard, N., Sabbadin, R.: Echantillonnage spatial basé sur le krigeage pour la reconstruction de cartes d'occurence. In: Proceedings of the 17th Conference on Reconnaissance de Forme et Intelligence Artificielle (RFIA) (2010)

Buesco, M., Angulo, J., Alonso, F.: A state-space model approach to optimum spatial sampling design based on entropy. Environmental and Ecological Statistics **5**(1), 29–44 (1998)

Celeux, G., Forbes, F., Peyrard, N.: EM procedures using mean field-like approximations for Markov model-based image segmentation. Pattern Recognition **36**(1), 131–144 (2003)

Chalmond, B.: An iterative Gibbsian technique for reconstruction of m-ary images. Pattern Recognition **22**(6), 747–761 (1989)

Chao, C.-T., Thompson, S.T.: Optimal adaptive selection of sampling sites. Environmetrics **12**, 517–538 (2001)

Chiles, J.-P., Delfiner, P.: Geostatistics: Modeling Spatial Uncertainty. Wiley, New York (1999)

Comer, M., Delp, E.: The EM/MPM algorithm for segmentation of texture images: analysis and further experimental results. IEEE Transactions on Image Processing **9**(10), 1731–1744 (2000)

Cook, S.A.: The complexity of theorem-proving procedures. In: Conference Record of Third Annual ACM Symposium on Theory of Computing (STOC), pp. 151–158 (1971)

Elith, J., Leathwick, J.: Species distribution models: ecological explanation and prediction across space and time. Annual Review of Ecology, Evolution, and Systematics **40**, 677–697 (2009)

Fuentes, M., Chaudhuri, A., Holland, D.: Bayesian entropy for spatial sampling design of environmental data. Environmental and Ecological Statistics **14**(3), 323–340 (2007)

Geman, S., Geman, D.: Stochastic relaxation, Gibbs distribution, and the Bayesian restoration of images. IEEE Transactions on Pattern Analysis and Machine Intelligence **6**, 721–741 (1984)

Gruijter, J.d., Brus, D., Bierkens, M., Knotters, M.: Sampling for Natural Resource Monitoring. Springer, Berlin (2006)

Guyon, X.: Random Fields on a Network—Modeling, Statistics and Applications. Probability and Its Applications. Springer, Berlin (1995)

Li, S.Z.: Markov Random Field Modeling in Computer Vision. Springer, Berlin (1995)

Lowe, S.J., Browne, M., Boudjelas, S.: 100 of the world's worst invasive alien species. Invasive Species Specialist Group (ISSG) 12(3), 12 p. (2000)

Pearl, J.: Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. Morgan Kaufmann, San Mateo (1988)

Peyrard, N., Sabbadin, R., Farrokh, U.: Decision-theoretic optimal sampling in hidden Markov random fields. In: 19th European Conference on Artificial Intelligence (ECAI), pp. 687–692 (2010)

Spall, J.C.: Introduction to Stochastic Search and Optimization. Wiley, New York (2003)

Sutton, R.S., Barto, A.G.: Reinforcement Learning: An Introduction. MIT Press, Cambridge (1998)

Thompson, S., Seber, G.: Adaptive Sampling. Series in Probability and Statistics. Wiley, New York (1996)

Winkler, G.: Image Analysis, Random Fields and Dynamics Monte Carlo Methods: A Mathematical Introduction. Springer, Berlin (1995)

Wu, F.: The Potts model. Reviews of Modern Physics **54**, 235–268 (1982)

Yedidia, J., Freeman, W., Weiss, Y.: Generalized belief propagation. In: Advances in Neural Information Processing Systems (NIPS), pp. 689–695 (2000)