Studying Species Demography and Distribution in Natural Conditions: Hidden Markov Models

Olivier Gimenez, Julie Louvrier, Valentin Lauret, Nina Luisa Santostasi

3

Studying species demographics and distributions in natural conditions : hidden Markov models

Olivier GIMENEZ[1], Julie LOUVRIER[2], Valentin LAURET[1] and Nina SANTOSTASI[3]

[1] *CEFE, Univ Montpellier, CNRS, EPHE, IRD, Univ Paul Valéry Montpellier 3, Montpellier, France*

[2] *Department of Ecological Dynamics, Leibniz Institute for Zoo and Wildlife Research, Berlin, Germany*

[3] *Department of Biology and Biotechnologies "Charles Darwin", University of Rome La Sapienza, Rome, Italy*

3.1. Introduction

Ecology may be defined as the study of living organisms in interaction with their environment. At the heart of this discipline lie two key questions : how many individuals are there in a population, and where are they ? In other terms, the first question relates to the dynamics of populations, while the second concerns the distribution of species. These questions have long attracted the interest of researchers ; for example, in the early 19th Century, Laplace attempted to estimate the size of the French population (Amorós 2014), while Grinnell, at the start of the 20th Century, focused on a formalization of the role of species in the operation of ecosystems (Grinnell 1917).

Statistical research in relation to these questions continues to this day, notably in terms of the analysis of data generated using new technologies (Gimenez, Buckland, Morgan, Bez, Bertrand, Choquet, Dray, Etienne, Fewster, Gosselin, Mérigot,

Monestiez, Morales, Mortier, Munoz, Ovaskainen, Pavoine, Pradel, Schurr, Thomas, Thuiller, Trenkel, de Valpine and Rexstad 2014). One issue which has attracted particular attention is the difficulty of observing individuals and species in natural conditions — essentially, a detection problem (Royle and Dorazio 2008). Given the imperfections inherent in the detection of individuals and species, variables such as whether an individual is dead or alive, or whether or not a species is present in a particular location, are only partially observable ; as such, they constitute hidden variables, in the sense defined in the introduction to this book.

In this chapter, we shall show how hidden Markov models (HMM) can be used to develop capture-recapture and occupancy models, traditionally used to study the dynamics of populations and the distribution of species in a context of imperfect detection. We shall show how the HMM formulation permits the estimation of hidden variables in two different case studies. The question of population dynamics will be illustrated through an estimation of the prevalence of wolf-dog hybrids in Italy, while the distribution of species will be illustrated by examining the distribution of wolves in France.

3.2. Overview of HMMs

Hidden Markov models (HMM) are a class of statistical models, generally used for analyzing data from systems with temporal dynamics. An ecological process may be modeled using a state process (or system process) of which the future states are solely dependent on current states : this is the Markov hypothesis. In an HMM, this process is not observed directly, but is hidden (latent). Observations are made based on a state-dependent process, controlled by the subjacent state process. These observations are essentially considered to be noisy measures of system states with a specific dependence structure. HMMs are a specific class of state space models with a finite number of states (Auger-Méthé *et al.* 2020 ; Gimenez *et al.* 2012).

In formal terms, an HMM consists of an observed state-dependent process $Y^1, Y^2, \ldots, Y^T$ and a non-observed (hidden) state process $Z^1, Z^2, \ldots, Z^T$. HMMs are often represented scehmatically in the way shown in Figure 2.1, which highlights the way in which observations are conditional on states, and illustrates the Markovian structure of the series of states.

Three components are needed to fully specify an HMM with N states. The first component is the initial distribution $\boldsymbol{\delta} = \big(\Pr(Z^1 = 1), \ldots, \Pr(Z^1 = N)\big)$ which combines the probabilities of being in different states at the start of the sequence. The second component is made up of the probabilities of transition $\gamma_{ij} = \Pr(Z^{t+1} = j | Z^t = i)$ between states i and j, generally grouped into an $N \times N$ transmission matrix, $\boldsymbol{\Gamma} = (\gamma_{ij})$. The third component is the distribution of the collected observations $f(y^t | Z^t = i)$, used to facilitate the calculation of likelihood in a diagonal matrix

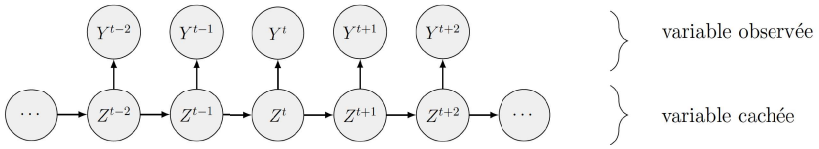

Figure 3.1 – Schematic illustration of a Hidden Markov Model.

of dimension $N \times N$, noted $\mathbf{P}(y^t) = \text{diag}\big(f(y^t|Z^t = 1), \ldots, f(y^t|Z^t = N)\big)$. In this chapter, only discrete and univariate distributions of observations will be addressed, but continuous distributions (Mews *et al.* 2020 ; Choquet *et al.* 2017) and multivariate distributions (Choquet *et al.* 2013 ; Laake *et al.* 2014 ; Johnson *et al.* 2016) may also be used.

The likelihood $\mathcal{L}(\boldsymbol{\theta} \mid y^1, \ldots, y^T)$ of the unknown parameters $(\boldsymbol{\theta})$ given an observed sequence $(Y^1, \ldots, Y^T)$ is expressed formally as :

$$\begin{aligned}\mathcal{L}(\boldsymbol{\theta} \mid y^1, \ldots, y^T) &= f_{\boldsymbol{\theta}}(y^1, \ldots, y^T) \\ &= \sum_{z^1=1}^{N} \cdots \sum_{z^T=1}^{N} f_{\boldsymbol{\theta}}(y^1, \ldots, y^T|z^1, \ldots, z^T) f_{\boldsymbol{\theta}}(z^1, \ldots, z^T) \\ &= \sum_{z^1=1}^{N} \cdots \sum_{z^T=1}^{N} \delta_{z^1} \prod_{t=1}^{T} f_{\boldsymbol{\theta}}(y^t|z^t) \prod_{t=2}^{T} \gamma_{z^{t-1},z^t}.\end{aligned}$$

The first step is obtained by applying the law of total probability ; the second step is a result of the Markovian dependency structure of the model.

The problem lies in the fact that the calculation of the likelihood of an HMM in this form requires N^T summations, making it time-consuming, if not impossible, to evaluate. One solution is to use a more efficient method to calculate likelihood. In this chapter, we have chosen to use the forward algorithm, which draws on the dependency structure of the model, instead of a the "brute force" approach which consists of summing all possible series of states.

Using the forward algorithm, likelihood is calculated as a matrix product :

$$\mathcal{L}(\boldsymbol{\theta} \mid y^1, \ldots, y^T) = \boldsymbol{\delta}\mathbf{P}(y^1)\boldsymbol{\Gamma}\mathbf{P}(y^2)\cdots\boldsymbol{\Gamma}\mathbf{P}(y^{T-1})\boldsymbol{\Gamma}\mathbf{P}(y^T)\mathbf{1}$$

where $\mathbf{1}$ is a column vector of ones. The complexity of this calculation is linear as a function of the number of observations, meaning that likelihood can be evaluated

rapidly in most of the cases encountered in ecology. The parameters $\boldsymbol{\theta}$ of an HMM can be calculated by maximum likelihood, using optimization routines (such as the Newton-Raphson method) to maximize likelihood numerically. This is the approach used here, implemented using R.

Once the parameters have been estimated, the next step is to infer the hidden states $z^1, \ldots, z^T$. In the context of HMM, this step is known as decoding. In this case, we use global decoding to look for the series of states $(g^1, \ldots, g^T)$ with the highest joint probability (this differs from local decoding, in which we search for the most likely value of z^t taken separately). In other terms, we wish to find :

$$(g^1, \ldots, g^T) = \underset{(z^1,\ldots,z^T)}{\arg\max} \Pr(Z^1 = z^1, \ldots, Z^T = z^T \mid y^1, \ldots, y^T).$$

This is a relatively complex optimization problem ; however, it can be solved efficiently using the Viterbi algorithm (Rabiner 1989).

For more details on HMMs in general, see (Zucchini *et al.* 2016) ; for the ecological context, see (McClintock *et al.* 2020).

3.3. HMM and demographics

3.3.1. *General overview*

The hidden (or partially hidden) variables encountered in the study of animal populations are living/dead ; developmental states, which are generally discrete, such as sexual maturity (Nichols *et al.* 1994) ; epidemiological states (Marescot *et al.* 2018) ; or social states (Dupont *et al.* 2015). These states can be hard to measure in the field. It is often impossible to track animals in their environment in an exhaustive manner, i.e. in the way human patients might be monitored in the context of a medical protocol. Data is often obtained in capture-recapture form, indicating whether or not an animal has been detected. If an individual is not detected, it may be possble to infer its state ; if an individual is detected, then its state may be known perfectly or imperfectly. HMMs are a natural choice for use in these contexts, as they can be used to formalize the analysis of noisy measures of demographic states.

One example involves the two states "dead" and "alive', with $Z^t = \mathrm{V}$ denoting "alive at time t" and $Z^t = \mathrm{M}$ "dead at time t" (Gimenez *et al.* 2007). The "dead" state here is absorbent insofar as an individual cannot leave the state once it has entered it (except in the context of zombie movies). An illustration of the corresponding HMM is given in Figure 2.2.

As we have seen, an HMM is defined using three components. The initial distribution is :

$$\begin{matrix} & \mathrm{V} & \mathrm{M} \end{matrix}$$
$$\boldsymbol{\delta} = \begin{pmatrix} 1 & 0 \end{pmatrix}.$$

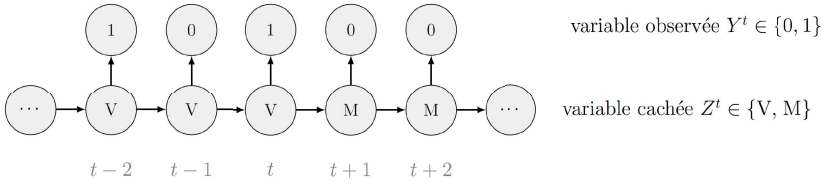

Figure 3.2 – Two-state capture-recapture model expressed in HMM form.

Let ϕ be the probability of survival over an interval of time. The transition probability matrix is given by :

$$\mathbf{\Gamma} = \begin{matrix} \\ \text{V} \\ \text{M} \end{matrix} \begin{matrix} \begin{matrix} \text{V} & \text{M} \end{matrix} \\ \begin{bmatrix} \phi & 1-\phi \\ 0 & 1 \end{bmatrix} \end{matrix}$$

Finally, the distribution of observations Y^t conditional on the states Z^t is a Bernoulli distribution of parameter p, where p is the probability of detection, if $Z^t = V$, or a Bernoulli distribution of parameter 0 if $Z^t = M$:

$$\mathbf{P}(y^t) = \begin{matrix} \\ \text{V} \\ \text{M} \end{matrix} \begin{matrix} \begin{matrix} \text{V} & \text{M} \end{matrix} \\ \begin{bmatrix} p^{y^t}(1-p)^{1-y^t} & 0 \\ 0 & 1-y^t \end{bmatrix} \end{matrix}.$$

Thus, if the individual is dead, $Z^t = M$, then the probability of observation is null, $\Pr(y^t = 1 | Z^t = M) = 1 - y^t = 0$, and the probability of it not being observed is 1, $\Pr(y^t = 0 | Z^t = M) = 1 - y^t = 1$. If the individual is a live, $Z^t = V$, the probability of observing it is $\Pr(y^t = 1 | Z^t = V) = p^{y^t}(1-p)^{1-y^t} = p$, and the probability of it not being observed $\Pr(y^t = 0 | Z^t = V) = p^{y^t}(1-p)^{1-y^t} = 1 - p$.

The contribution of each individual to the overall likelihood of the data set can then be calculated using these components. For example, consider a study which takes place over the course of $T = 3$ years, and let us take an individual observed in the first and third years, but not in the second year : $(y^1 = 1, y^2 = 0, y^3 = 1)$. This individual's contribution to the likelihood is written :

$$\mathcal{L}(\phi, p \mid y^1, y^2, y^3) = f_{\phi,p}(y^1, y^2, y^3) = \boldsymbol{\delta}\mathbf{P}(y^1)\mathbf{\Gamma}\mathbf{P}(y^2)\mathbf{\Gamma}\mathbf{P}(y^3)\mathbf{1}$$

with

$$\mathbf{P}(y^1) = \mathbf{P}(y^3) = \begin{matrix} \\ \text{V} \\ \text{M} \end{matrix} \begin{matrix} \begin{matrix} \text{V} & \text{M} \end{matrix} \\ \begin{bmatrix} p & 0 \\ 0 & 0 \end{bmatrix} \end{matrix}.$$

and

$$\mathbf{P}(y^2) = \begin{array}{c} \\ 1-p \\ 0 \end{array} \begin{array}{c} \begin{array}{cc} \text{V} & \text{M} \end{array} \\ \begin{bmatrix} 0 \\ 1 \end{bmatrix} \end{array}.$$

We can verify (with a little patience) that this matrix product is equal to $p\phi(1-p)\phi p$, generally conditioned with respect to the first capture, with an assigned value of 1, such that $f_{\phi,p}(y^1, y^2, y^3) = \phi(1-p)\phi p$.

Once the probabilities of survival and detection have been estimated, it should, in theory, be possible to calculate life expectancy based on the inferred dead/alive status of individuals, reconstructing the sequence of states for each individual.

This two-state example may be generalized to give a multi-state capture-recapture model (Lebreton *et al.* 2009), incorporating reproductive states. This model may be formulated as a three-state HMM, including a "dead" state and two reproductive states, R and NR : $Z^t = \text{R}$ for "alive and reproducing at time t" and $Z^t = \text{NR}$ "alive and not reproducing at time t". A schematic representation of this HMM is shown in Figure 2.3

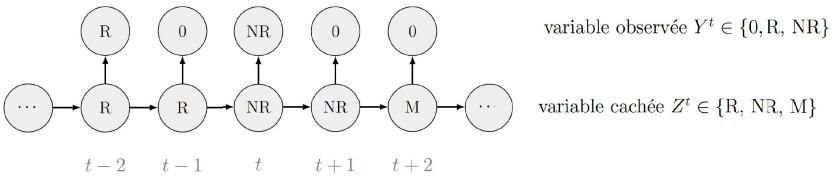

Figure 3.3 – Multi-state capture-recapture model expressed in HMM form.

The first component of the associated HMM is the initial distribution :

$$\boldsymbol{\delta} = \begin{array}{c} \\ \end{array} \begin{array}{c} \begin{array}{ccc} \text{R} & \text{NR} & \text{M} \end{array} \\ \begin{pmatrix} \delta_R & 1-\delta_R & 0 \end{pmatrix} \end{array}$$

Let ϕ_R be the probability of survival of reproducing individuals, ϕ_{NR} that of non-reproducing individuals, $\psi_{NR,R}$ the probability of an individual which is not reproducing at time t entering the reproductive state at time $t+1$, and $\psi_{R,NR}$ the probability that an individual which is in the reproductive state at time t will have left this state at

time $t+1$. The transition matrix is written :

$$\mathbf{\Gamma} = \begin{array}{c} \\ \text{R} \\ \text{NR} \\ \text{M} \end{array} \begin{array}{c} \begin{array}{ccc} \text{R} & \text{NR} & \text{M} \end{array} \\ \begin{bmatrix} \phi_R(1-\psi_{R,NR}) & \phi_R\psi_{R,NR} & 1-\phi_R \\ \phi_{NR}\psi_{NR,R} & \phi_{NR}(1-\psi_{NR,R}) & 1-\phi_{NR} \\ 0 & 0 & 1 \end{bmatrix} \end{array}$$

Finally, let p_R be the probability of detection of reproducing individuals (and p_{NR} that of non-reproducing individuals). The diagonal matrix giving the distribution of observations conditional on states is thus :

$$\mathbf{P}(y^t) = \begin{array}{c} \begin{array}{ccc} \text{R} & \text{NR} & \text{M} \end{array} \\ \begin{bmatrix} p_R^{I(y^t=R)}(1-p_R)^{I(y^t=0)}0^{I(y^t=NR)} & 0 & 0 \\ 0 & p_{NR}^{I(y^t=NR)}(1-p_{NR})^{I(y^t=0)}0^{I(y^t=R)} & 0 \\ 0 & 0 & I(y^t=0) \end{bmatrix} \end{array}$$

where $I(y^t = k)$ is the indicator function, taking a value of 1 when $y^t = k$ and 0 otherwise. The distribution implied here is a generalization of the Bernouilli distribution for more than two possible outcomes, i.e. a categorical (single-trial multinomial) distribution.

For example, to study reproduction costs, ecologists may compare the probability of reproducing in year $t+1$ based on the individual's reproductive, $(\psi_{R,R} = 1 - \psi_{R,NR})$, or non-reproductive, $(\psi_{NR,R})$, state in year t ; the differences in survival rates between reproducing (ϕ_R) and non-reproducing (ϕ_{NR}) individuals may also be studied in this way.

While multi-state models were originally developed for use in estimating demographic parameters (survival, movement, etc.) which depend on geographical sites (Brownie *et al.* 1993), there are few real limits to their application in ecology (Gimenez *et al.* 2012).

Once the parameters have been estimated, the subjacent states can be inferred. In this way, it becomes possible to calculate particularly interesting ecological quantities. Examples of this include the sex ratio (Pradel *et al.* 2008), where the states are the sex of individuals ; reproductive success over a lifetime (Rouan, Gaillard, Guédon and Pradel 2009 ; Gimenez *et al.* 2012 ; Desprez *et al.* 2018) ; or the number of sick individuals (Buzdugan *et al.* 2017) in the case of epidemiological states.

One tacit hypothesis which is inherent in multi-state levels is that the state of an individual can be measured without error. In practice, however, it can be difficult to assign a sure state to individuals, for example when observing reproduction in the field. A reproductive state can be confirmed if a female is seen with one or more

young, for example, but if a female is observed alone, status assignment is less certain. The HMM approach takes account of this element of uncertainty in the assignment of states to individuals (Dupuis 1995 ; Pradel 2005 ; Gimenez *et al.* 2012), as we shall see in the following example.

3.3.2. *Case study : estimating the prevalence of dog-wolf hybrids in a context of uncertain individual identification*

The points made above can be illustrated using an example, in this case relating to the estimation of the prevalence of hybrids in a wild animal population. Our case study concerns cross-breeding between dogs and wolves in the Tusco-Emilian Apennines National Park, Italy (Santostasi *et al.* 2019). The data was obtained using wolf feces collected from August 2016 and May 2017, from which DNA was extracted, amplified and sequenced (Caniglia *et al.* 2014) ; using this DNA data, a distinction can be made between wolves, hybrids and animals of uncertain status. There were 5 capture sessions, each spanning 2 months, featuring samples from 39 individuals (19 wolves, 12 hybrids and 8 uncertain). In the original study, the authors compared different models, including or ignoring the difference between hybrid and parental individuals in terms of detection and assignment probabilities.

The possible states included parental $Z^t = P$, hybrid $Z^t = H$ or dead $Z^t = M$, with observations noted $y^t = 0$ for undetected, $y^t = 1$ for observed parental, $y^t = 2$ for observed hybrid and $y^t = 3$ for observed, uncertain status. All parameters in the model used here are constant, except for survival, which is state dependent ; hence $\phi_P \neq \phi_H$.

The components used to write the likelihood of the HMM are the initial distribution

$$\begin{matrix} & \text{P} & \text{H} & \text{M} \\ \boldsymbol{\delta} = & (\delta_P & 1-\delta_P & 0\,), \end{matrix}$$

the transition matrix

$$\begin{matrix} & \text{P} & \text{H} & \text{M} \\ \boldsymbol{\Gamma} = & \begin{bmatrix} \phi_P & 0 & 1-\phi_P \\ 0 & \phi_H & 1-\phi_H \\ 0 & 0 & 1 \end{bmatrix} & \begin{matrix} \text{P} \\ \text{H} \\ \text{M} \end{matrix} \end{matrix}$$

and the diagonal matrix which gives the distribution of observations conditional to the states

$$\begin{matrix} & \text{P} & \text{H} & \text{M} \\ \mathbf{P}(y^t) = & \begin{bmatrix} f(y^t|Z^t=P) & 0 & 0 \\ 0 & f(y^t|Z^t=H) & 0 \\ 0 & 0 & I(y^t=0) \end{bmatrix} \end{matrix}$$

Tableau 3.1 – Prevalence of hybrids : observed and estimated using the Viterbi algorithm.

Prevalence	Occ. 1	Occ. 2	Occ. 3	Occ. 4	Occ. 5
Observed	0.27	0.33	0.20	0.46	0.27
Estimated	0.27	0.33	0.20	0.20	0.18
95% confidence interval	(0.09,0.61)	(0.10,0.61)	(0.00, 0.50)	(0.00, 0.50)	(0.00, 0.50)

where $f(y^t|Z^t = P) = (1-p)^{I(y^t=0)}(p\delta)^{I(y^t=1)}0^{I(y^t=2)}(p(1-\delta))^{I(y^t=3)}$ and $f(y^t|Z^t = H) = (1-p)^{I(y^t=0)}0^{I(y^t=1)}(p\delta)^{I(y^t=2)}(p(1-\delta))^{I(y^t=3)}$.

The important parameter here is δ, the probability of an individual being assigned to a state. If the genetic or morphological assessment is not sufficient to assign parental or hybrid status to an individual, then it will be classed as uncertain, with probability $1-\delta$.

As the hybridization test was carried out just once for each genotype, the assignment probability δ is estimated for the first capture alone. The assignment of parental or hybrid status to individuals in the uncertain category, and consequently the calculation of the prevalence of hybrids, is carried out using global decoding by means of the Viterbi algorithm.

The probability of survival for wolves ϕ_P is estimated at 0.63 (0.39-0.82), lower than the probability of survival for hybrids ϕ_H, estimated at 0.81 (0.59-0.93). The probability of detection p is estimated to be 0.46 (0.31-0.61) and the probability of assignment δ is estimated to be 0.85 (0.75-0.91).

The main result in this case is an estimation of the number of hybrid individuals. The estimated prevalence varies from 0.18 to 0.33, and is comparable to the observed prevalence (Table 2.1).

(Santostasi *et al.* 2019) compares several models ; the authors show that the estimated prevalence is systematically in excess of the observed prevalence, with important consequences in terms of species management. The HMM permits a confidence interval to be used in conjunction with the estimation of prevalence.

3.4. HMM and distribution

3.4.1. *General case*

Instead of working on an individual scale, a different perspective can be gained by using detection and non-detection data at species level. This data gives us access to

spatial information in relation to species and populations, e.g. occupancy. In concrete terms, datais obtained by monitoring several spatial units (such as breeding sites or photo traps) where a species may or may not be detected. Occupancy models are used to estimate the proportion of an area occupied by a species, with corrections for imperfect detectability (MacKenzie *et al.* 2018) ; in dynamic cases, the probabilities of local extinction ϵ and colonization κ are also included. Sites are treated in exactly the same way as individuals using the capture-recapture approach, and occupancy models are thus similar to the capture-recapture models presented in the previous section. Occupancy models can be seen as HMMs (Royle and Kéry 2007 ; Gimenez, Blanc, Besnard, Pradel, Doherty Jr, Marboutin and Choquet 2014) in which the state process governs the dynamics of site states, with $Z^t = \text{O}$ denoting "occupied site" and $Z^t = \text{NO}$ denoting "non-occupied site" for a year t. A species may be detected, $Y^{t,k} = 1$, or undetected, $Y^{t,k} = 0$, at each site on multiple visits k over the course of a year t. A schematic representation of the corresponding HMM is shown in Figure 2.4

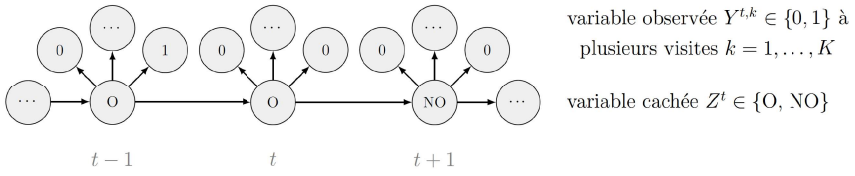

Figure 3.4 – Diagram of a dynamic occupancy model expressed as an HMM.

The components used in constructing the likelihood of the model are written as above. We begin with the initial distribution :

$$\begin{matrix} & \text{O} & \text{NO} \end{matrix}$$
$$\boldsymbol{\delta} = (\psi_1 \;\; 1-\psi_1)$$

where ψ_1 is the probability of initial occupancy (in the first year). The transition matrix is written :

$$\begin{matrix} & \text{O} & \text{NO} \end{matrix}$$
$$\boldsymbol{\Gamma} = \begin{bmatrix} 1-\epsilon & \epsilon \\ \kappa & 1-\kappa \end{bmatrix} \begin{matrix} \text{O} \\ \text{NO} \end{matrix}$$

Finally, the state-dependent matrix of the observation distribution is :

$$\mathbf{P}(\mathbf{y}^t) = \begin{array}{c} \begin{array}{cc} \text{O} & \text{NO} \end{array} \\ \begin{bmatrix} \prod_{k=1}^{K} p^{y^{t,k}}(1-p)^{1-y^{t,k}} & 0 \\ 0 & \prod_{k=1}^{K}(1-y^{t,k}) \end{bmatrix} \end{array}$$

where p is the probability of detection of the species.

One sepcial case is that of single-season (static) occupancy (MacKenzie *et al.* 2002) where $\epsilon = \kappa = 0$ (Gimenez, Blanc, Besnard, Pradel, Doherty Jr, Marboutin and Choquet 2014) and $T = 1$. The HMM formulation allows us not only to estimate the probabilities of occupancy, extinction and colonization, but also to estimate the sstate of a site if the species has not been detected (via global decoding). Thanks to the flexibility of the HMM formulation, the standard model can be extended to take account of differences in the probability of detecting a species via finite number mixtures (Louvrier, Chambert, Marboutin and Gimenez 2018) or a discrete measure of this heterogeneity such as population density or reproductive state (Gimenez, Blanc, Besnard, Pradel, Doherty Jr, Marboutin and Choquet 2014 ; Veran *et al.* 2015) ; it can also take account of the occurrence of false positives in the data due to erroneous species identification (Miller *et al.* 2011 ; Louvrier *et al.* 2019). As in the case of multi-state capture-recapture models, HMM occupancy models can be extended to include multiple "occupied" states, such as reproductive states (MacKenzie *et al.* 2009 ; Martin *et al.* 2009), epidemiological states (McClintock *et al.* 2010), or landscape-related states (Lamy *et al.* 2013). These models can also be extended to cases with multiple species in order to study predator-prey relationships (Fidino *et al.* 2019 ; Rota *et al.* 2016).

3.4.2. *Case study : Estimating the distribution of a wolf population in a case with species identification errors and heterogeneous detection*

In this case, an HMM will be used to model species distribution in a case featuring identification errors and heterogeneous deteciton. The data analyzed relates to the detection and non-detection of wolves in France, and was collected in 2013 (Louvrier, Duchamp, Lauret, Marboutin, Cubaynes, Choquet, Miquel and Gimenez 2018). Signs that the species was present, such as tracks, faeces, prey remains, dead animals, camera trap photographs and actual spottings were collected by a network of professional and amateur observers (Duchamp *et al.* 2012). The data for 2013 comprised 250 certain detections, 54 uncertain detections (cases of confusion with another species) and 12540 non-detections across a grid of 3211 sites over a 10 x 10 km space.

We have chosen to consider each month, from December to March, as a separate sampling occasion. These months correspond to a period between two dispersion events, in the fall and the spring (Louvrier, Duchamp, Lauret, Marboutin, Cubaynes, Choquet, Miquel and Gimenez 2018). This choice increases the chance of respecting an important hypothesis inherent to occupancy models, namely that the state of the site should stay the same over the course of the study. In a previous study, we found that the main explanatory factor for occupation was site altitude, but that the probability of detecting the species was mostly determined by the sampling effort, defined as the number of observers per site per year (Louvrier, Duchamp, Lauret, Marboutin, Cubaynes, Choquet, Miquel and Gimenez 2018). In this case, for illustrative purposes, we have chosen to focus on a model which takes account of identification errors and heterogeneous detection in the determination of detection probabilities (Louvrier, Chambert, Marboutin and Gimenez 2018). After estimating the parameters of the model, we constructed a map representing the 3211 sites in the study area, each associated with a heterogeneity class estimated using the Viterbi algorithm.

We considered two classes of site, A and B, with respective proportions π and $1-\pi$. The possible states were $Z^k = OA$ for an occupied site of class A, A, $Z^k = OB$ for an occupied site of class B, $Z^k = NOA$ for a non-occupied site of class A and $Z^k = NOB$ for a non-occupied site of class B. We constructed a single-season (static) model with $k = 1, \ldots, K$ visits. Observations were classed as $y^k = 0$ for a site where the species was not observed, $y^k = 1$ for a site with an unambiguous observation, and $y^k = 2$ for a site with an ambiguous observation. ambiguïté. In this case, we have chosen to work with a model in which all parameters are constant over time, but dependent on the site classification in terms of detection.

The components used in writing the likelihood of the HMM are the initial distribution :

$$\begin{matrix} & \text{NOA} & \text{NOB} & \text{OA} & \text{OB} \end{matrix}$$
$$\boldsymbol{\delta} = (\pi(1-\psi_A) \;\; (1-\pi)(1-\psi_B) \;\; \pi\psi_A \;\; (1-\pi)\psi_B),$$

and the diagonal matrix giving the distribution of observations conditional on states :

$$\mathbf{P}(y^k) = \begin{matrix} \text{NOA} \\ \text{NOB} \\ \text{OA} \\ \text{OB} \end{matrix} \begin{bmatrix} f(y^k|Z^k = NOA) & 0 & 0 & 0 \\ 0 & f(y^k|Z^k = NOB) & 0 & 0 \\ 0 & 0 & f(y^k|Z^k = OA) & 0 \\ 0 & 0 & 0 & f(y^k|Z^k = OB) \end{bmatrix}$$

where $f(y^k|Z^k = NOA) = (1 - p_{A10})^{I(y^k=0)} 0^{I(y^k=1)} p_{A10}^{I(y^k=2)}$, $f(y^k|Z^k = NOB) = (1 - p_{B10})^{I(y^k=0)} 0^{I(y^k=1)} p_{B10}^{I(y^k=2)}$, $f(y^k|Z^k = OA) = (1 - p_{A11})^{I(y^k=0)} (bp_{A11})^{I(y^k=1)} (1-b) p_{A11})^{I(y^k=2)}$ et $f(y^k|Z^k = OB) = (1 -$

$p_{B11})^{I(y^k=0)}(bp_{B11})^{I(y^k=1)}(1-b)p_{B11})^{I(y^k=2)}$ where p_{A11} is the probability of correctly detecting the species at a class A occupied site (respectively p_{B11} for class B), p_{A10} is the probability of wrongly detecting the species at a class A non-occupied site (resp. p_{B10} for B), and b is the probability of classifying a true positive as unambiguous or certain. As there is no dynamic element with respect to site state, the transition matrix is the identity matrix.

The model presents several local maxima in terms of likelihood, something which is common when using HMMs. It can be hard to pinpoint the reason for this problem ; our preferred approach is to apply multiple numerical optimizations, changing the initial values each time. In this case, 100 random drawings were carried out from a uniform distribution between 0 and 1 to provide initial values for the model parameters, which are all probabilities ; the model was then adjusted for each combination. The results are striking, featuring multiple optima, as shown in Figure 2.5.

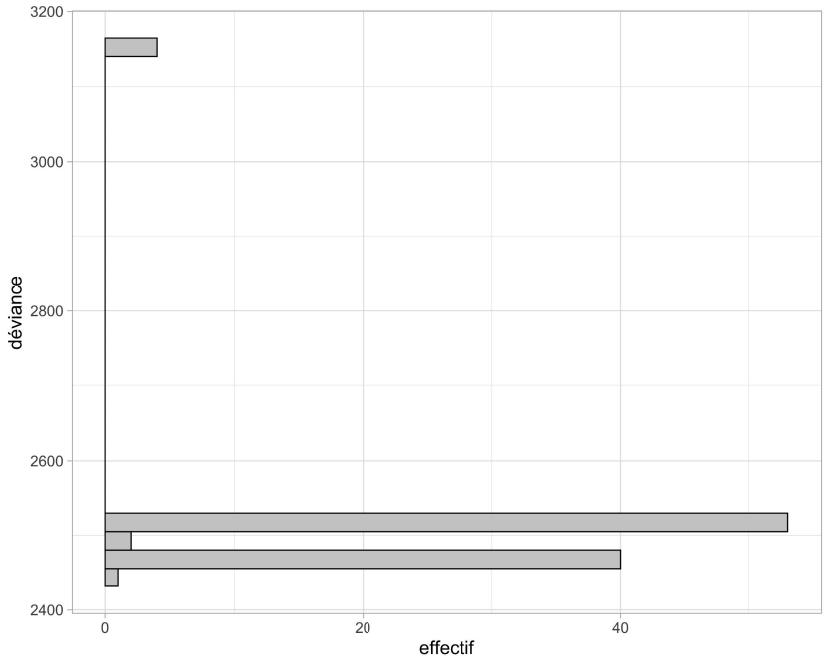

Figure 3.5 – Identification of local minima in the $-2\log(\mathcal{L}(\boldsymbol{\theta}))$ deviance of an HMM. Numerical optimization was carried out using 100 random drawings of initial values. The graph shows number of instances (x axis) against value (y axis). Several local minima are clearly visible.

The estimated probability of occupation is low, at 0.05 (0.04-0.06). According to the adjusted model, 94% have a zero probability of detection of a false positive p_{B10},

indicating that there are no identification errors for these sites. From the remaining 6% of sites, the estimated value of p_{A10} is also low, at 0.05 (0.03-0.08). These results suggest that the training procedure followed by observers in the network was effective, and/or that the data filtering process applied prior to analysis minimizes the number of false positives. The probability b of classifying a true positive as non-ambiguous is high, estimated at 0.93 (0.90-0.95). Taken in conjunction with the low risk of false positives, this result suggests that uncertain detections could be considered as certain. Finally, the probability of detection of true positives p_{11} was estimated at 1 for 6% of sites, and at 0.39 (0.35-0.43) for the remaining 94%.

Once the parameters of the model have been estimated, the Viterbi algorithm may be used to determine the most probable state for each site. Once the most probable state of each site has been determined, the results may be viewed on a map, such as that shown in Figure 2.6, showing the level of heterogeneity. Observed variations between sites are partly the result of spatial variations in sampling effort, defined as the number of active observers for a site (Louvrier, Chambert, Marboutin and Gimenez 2018). The interest of using HMMs in this case lies in the ability to take account of heterogeneity in the observation process. This is done using a hidden variable to account for belonging to a finite number of classes ; it is thus possible to avoid the need to measure sampling effort on the ground, a promising property for analyzing data obtained using participative approaches.

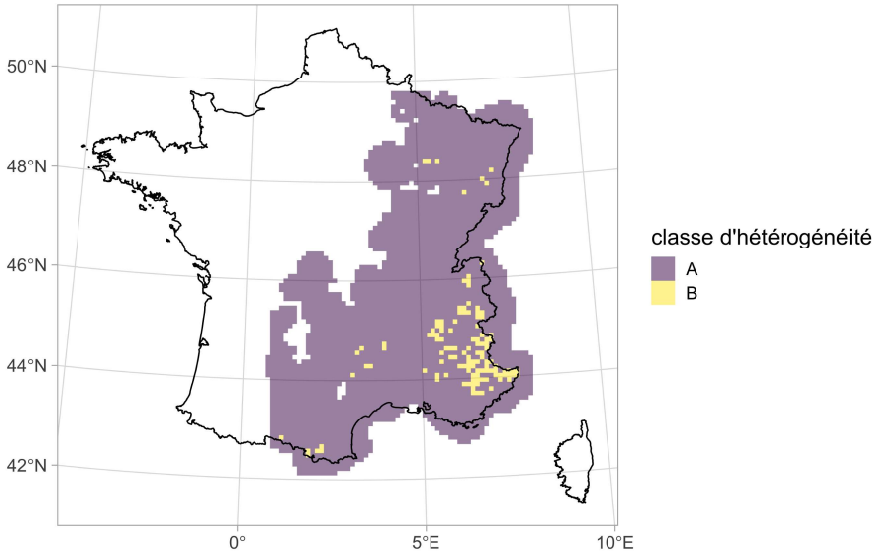

Figure 3.6 – Visualization of heterogeneity : map of the heterogeneity class to which each site in the study area is assigned using the Viterbi algorithm.

3.5. Discussion

In this chapter, we have seen how hidden Markov models (HMM) can be used in ecology to respond to questions concerning the demographics and distribution of species in their natural environment. The flexibility and ecological relevance of the HMM modeling framework have contributed to its increasing popularity in ecology, where it is used in relation to a wide range of questions (McClintock *et al.* 2020). The main advantage of the HMM approach lies in the ability to infer the ecological state of individuals and species which are, at best, partially observable : these are hidden variables. In addition to the ability to explicitly distinguish between observation processes and states, it is possible to decompose potentially complex processes into several simpler steps (Choquet 2008), facilitating model construction (Santostasi *et al.* 2019 ; Louvrier, Chambert, Marboutin and Gimenez 2018). Finally, HMMs make it possible to infer state dynamics in time and space. Note that model selection and approaches to testing the quality of adjustment of models to data are not covered here ; for a detailed discussion of these issues, see (Zucchini *et al.* 2016) and (McClintock *et al.* 2020).

Nevertheless, HMMs do have limits, three of which will be discussed here. The first is numerical in nature. As we saw in our case study concerning occupancy models, the likelihood function may present local maxima, which makes global maximization complex. The solution to this problem generally involves testing several sets of initial values for numerical optimization, via random drawings, as in the case described above ; another option is to use estimated parameters for a simplified model less subject to local maxima as the initial values. Other approaches may aslo be used (Brooks and Morgan 1994). A further problem is linked to the non-identifiability of models for which the likelihood is uniform in areas, for example in the case of redundant parameters ; this problem can be diagnosed (Cole 2019).

The second limitation concerns the Markovian hypothesis itself. This hypothesis implies that the time taken to move from one state to another follows a geometric distribution, and this is not always verified in practice. One solution to this problem is to consider Markov chains with an order greater than 1, or, in other words, to to assign memory to HMMs. In terms of demographics, this consists, for example, in admitting that the probability of movement between geographical sites depends not only on the current site, but also on previously-visited sites (Cole *et al.* 2014 ; Rouan, Choquet and Pradel 2009). Another solution is to model the time spent in a state directly, in the form of a semi-Markov model (Choquet *et al.* 2011 ; King and Langrock 2016).

The third limitation relates to the discrete nature of states in HMMs. In cases where a finite number of states are used to approximate the distribution of a continuous variable, such as the mass of an individual or the geographical range of a species, the question of discretization must be adressed. Evidently, the number of states may be increased to make the discretization finer, but at the cost of increased complexity, via an increase in the number of parameters and/or states to estimate. The problems

relating to high-dimensional space states can be mitigated by exploiting the fact that only certain transitions are possible, increasing calculation efficiency (Glennie *et al.* 2019) ; another option is to group states (Besbeas and Morgan 2019).

In this chapter, we have demonstrated the adjustment of HMMs in a frequentist setting, combining an efficient expression of likelihood using the forward algorithm with numerical optimization in order to obtain estimators of the maximum likelihood of parameters, then using the Viterbi algorithm to reconstruct the most likely sequence of states (hidden variables) in a process known as decoding. Our approach can be implemented in R and is reproducible (the code is available to download from GitHub : `https://oliviergimenez.github.io/code_livre_variables_cachees/gimenez.html`). There are several computer-based solutions for implementing a frequentist approach and for using HMMs to analyze capture-recapture or occupancy data(Choquet *et al.* 2009 ; Laake 2013) ;(Gimenez, Blanc, Besnard, Pradel, Doherty Jr, Marboutin and Choquet 2014 ; Fiske and Chandler 2011). Other tools which may be used in this context include the EM algorithm (see Chapter 4) or the Bayesian approach, implemented via Markov chain Monte Carlo methods (MCMC). The Bayesian approach is enjoying increasing popularity for adjusting statistical models in the field of ecology, notably due to the availability of flexible, powerful programs (de Valpine *et al.* 2017 ; Plummer *et al.* 2003). A major advantage of the Bayesian approach is that hidden variables are treated as parameters to estimate, making it easy to take account of a measure of uncertainty with regard to these variables. However, the drawback is that standard MCMC samplers do not perform particularly well in cases where both parameters and hidden variables must be determined. One solution is to apply sampling to the parameters alone, marginalizing states via the forward algorithm (Turek *et al.* 2016 ; Yackulic *et al.* 2020), but this has a negative effect on the estimation of hidden variables. Research into the use of the Viterbi algorithm within a Bayesian framework is currently ongoing (Lember *et al.* 2019).

3.6. Thanks

This work was partly financed in the context of the ANR DEMOCOM project (ANR-16-CE02-0007). The authors would like to thank the participants in the “wolf” network under the supervision of the French Office for Biodiversity (Office Français de la Biodiversité) ; C. Duchamp for assistance and provision of wolf data (France) ; M. Canestrini and F. Moretti for wolf sample collection (Italy) ; M. Galaverni for assistance with genetic analysis. Julie Louvrier wishes to thank the Université de Montpellier and the OFB for her thesis grant. Nina Santostasi wishes to thank the Sapienza University of Rome for her thesis grant.