HSMM driven by the observations for estimating weed dynamics

Hanna Bacave[1], Pierre-Olivier Cheptou[2], Nikolaos Limnios[3], Nathalie Peyrard[1]

[1] INRAE, UR MIAT, Université de Toulouse, Castanet-Tolosan, France.

[2] CEFE-CNRS, Université de Montpellier, France.

[3] Sorbonne University Alliance, Université de Technologie de Compiègne, LMAC, France.

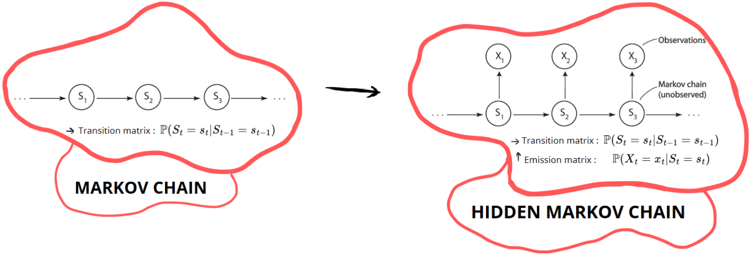
Pré-journée des doctorants, Journées des Statistiques du Sud, June 18, 2024

First step : What are Hidden Markov Models (HMM) ?

Hidden Markov Models (HMM) are used to study time series.

Example (HMM uses cases)

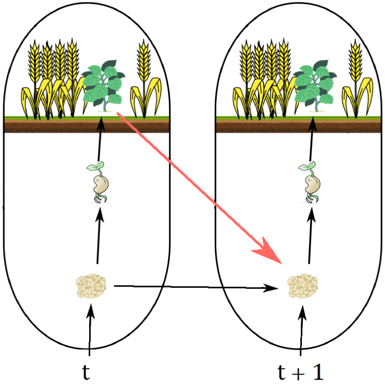
Field of study	**Uses**
Medicine	Analyze epidemilogic surveillance data (Le Strat and Carrat 1999)
Ecology	Reconstruct hidden or partially observed ecological dynamics (McClintock et al. 2020)
Finance	Predict the regime of a monetary system thanks to the exchange rate (Engel and Hamilton 1990)


S_1
S_2
S_3
→ Transition matrix : $\mathbb{P}(S_t = s_t | S_{t-1} = s_{t-1})$
MARKOV CHAIN
X_1
X_2
X_3
Observations
Markov chain (unobserved)
S_1
S_2
S_3
→ Transition matrix : $\mathbb{P}(S_t = s_t | S_{t-1} = s_{t-1})$
↑ Emission matrix : $\mathbb{P}(X_t = x_t | S_t = s_t)$
HIDDEN MARKOV CHAIN

Improve the understanding of weed dynamics

Aim of the work

Estimate key parameters of the dynamics based only on observation of standing plants.



Parameters involved in the dynamics :

- colonization
- germination
- survival
- seed production

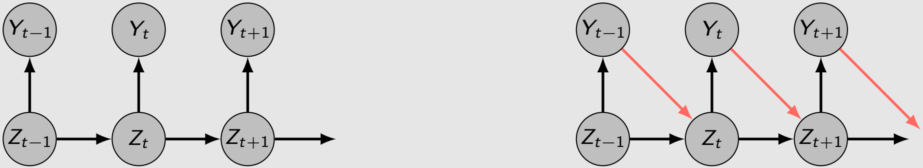
History of weed dynamics modeling

First models (Hanski and Gaggiotti 2004)

The initial models of weed dynamics assumed that the state of the seed bank was known.

Observation-Driven Hidden Markov Model (OD-HMM, Pluntz et al. 2018)

The OD-HMM is an extension of the HMM to the case where there is a **dependence** between **observations** $Y = (Y_t)_{t \in \mathbb{N}}$ and **hidden states** $Z = (Z_t)_{t \in \mathbb{N}}$.



(a) HMM.

(b) OD-HMM.

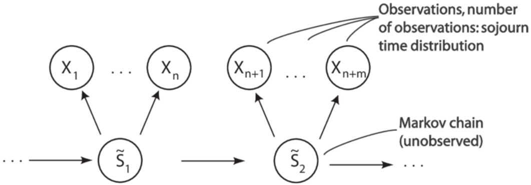
Figure: Graphical representation of conditional independencies in the chain (Z_t, Y_t).

OD-HMM vs OD-HSMM?

Limits of the OD-HMM

Weed seeds can remain in the soil for many years, which is not modeled with HMM.

⇒ **Solution :** Hidden semi-Markov Models (HSMM).


Observations, number of observations: sojourn time distribution
X_1
X_n
X_{n+1}
X_{n+m}
$\tilde{S}_1$
$\tilde{S}_2$
Markov chain (unobserved)

Main difficulty of extending HSMM to the case where there is a dependence between observations and hidden states

The dependence between observations and hidden states influences the sojourn time distribution at each point in time.

⇒ **Solution :** Use the couple (state, time since entry) as the new state of a hidden Markov chain (Barbu and Limnios 2008).

Distributions defining a OD-HSMM for weed dynamics

Discrete time. State spaces : $\Omega_Z = \Omega_Y = \{0, 1\}$.
Time elapsed since the entry into the current state : $U = (U_t)_{t \in \mathbb{N}}$, with $\Omega_U = \mathbb{N}^*$.

- **Initial probability :**

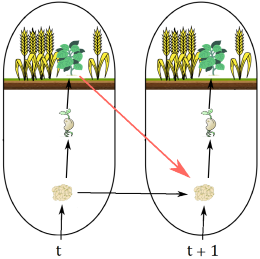
$$\pi(z_0) = \mathbb{P}(Z_0 = z_0, U_0 = 1),$$

where $p_0 = \mathbb{P}(Z_0 = 1)$;

- **Emission probability :**

$$R(z_t, y_t) = \mathbb{P}(Y_t = y_t | Z_t = z_t),$$

representing the germination probability g ;



- **Transition probability :**

$$P_{y_{t-1}}(z_{t-1}, u_{t-1}, z_t, u_t) = \mathbb{P}(Z_t = z_t, U_t = u_t | Z_{t-1} = z_{t-1}, U_{t-1} = u_{t-1}, Y_{t-1} = y_{t-1}),$$

parameterized by probabilities of colonization c, dispersion d and survival : optimal survival s_0 + survival degeneration λ.

Example

$$P_1((1, 3), (0, 1)) = (1 - s_0 e^{-\lambda \times 2})(1 - c)(1 - d), \text{ and } R(1, 0) = 1 - g.$$

How to estimate the parameters involved in weed dynamics?

We want to estimate the parameters $\theta = (c, d, g, s_0, \lambda)$.

Approximate Bayesian Computation (ABC) algorithm - rejection method

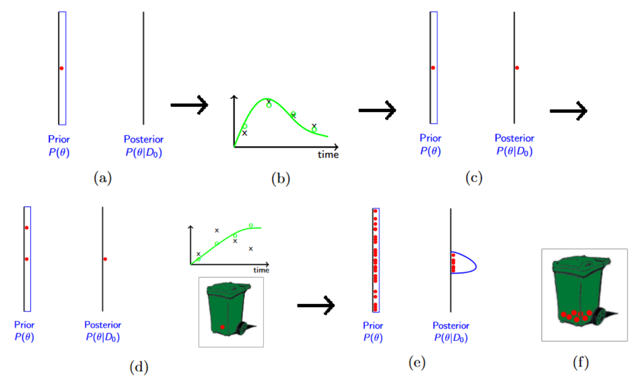


Figure: ABC tutorial taken from Toni and Stumpf 2009

We use the sequential ABC (ABC SMC) algorithm from *EasyABC* R package (Jabot et al. 2013), with the Lenormand method (Lenormand et al. 2013).

Selection of summary statistics

Example (Typical sequence of observations with $M = 10$)

$$Y^*_{0:M} = (0\ 0\ 0\ 0\ 1\ 0\ 0\ 1\ 1\ 1).$$

We consider the following summary statistics :

1. number of 1 ;
2. number of transitions from 0 to 1 ;
3. information about the lengths of consecutive 0's and 1's (Minimum, Maximum, Mean, Quantile at 25%, Median, Quantile at 75%) ;
4. mode and mode value in consecutive 0 and 1 lengths.

We consider three ways to group them :

- Group 1 : summary statistics 1, 2 and 3
- Group 2 : summary statistics 1, 2 and 4
- Group 3 : all summary statistics

Experiments on simulated data - Protocol

Protocol (repeated 30 times)

$$\begin{aligned}\theta^* &= (c^*,\ d^*,\ g^*,\ \lambda^*,\ s_0^*)\\ &= (0.5,\ 0.5,\ 0.5,\ 0.5,\ 0.5).\end{aligned}$$

1. Simulate a true sequence of hidden states and observations $(Z^*_{0:M}, Y^*_{0:M})$.
2. For each group of summary statistics :
   - Run the ABC algorithm to estimate the distribution $\hat{\theta}$.

For each parameter, we plot the box-plot associated to the mode of the 30 estimated distributions.

Experiments on simulated data - Results for the colonization parameter c and the dispersal parameter d

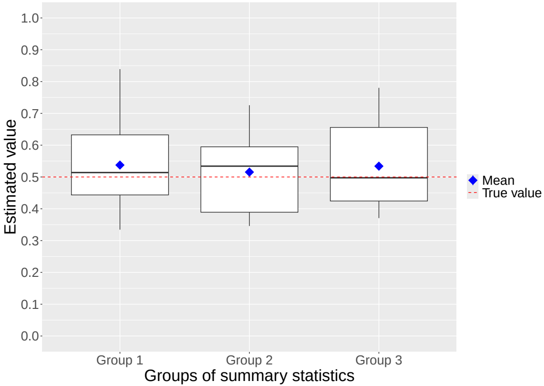


Figure: Estimation of the parameter c

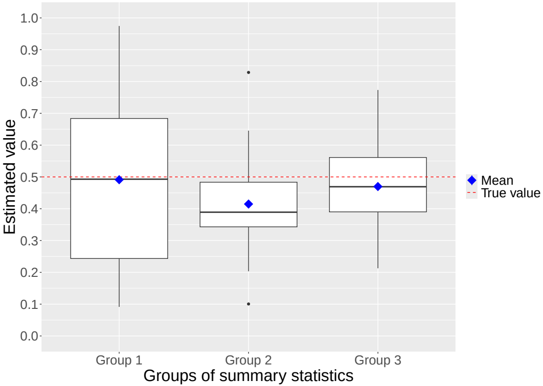


Figure: Estimation of the parameter d

Experiments on simulated data - Results for the germination parameter g and the survival parameter s_0

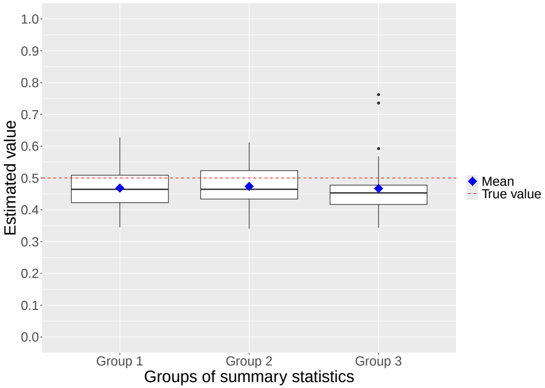


Figure: Estimation of the parameter g

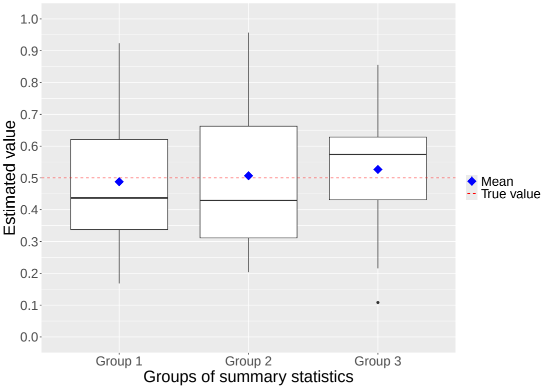


Figure: Estimation of the parameter s_0

Experiments on simulated data - Results for the survival degeneration parameter λ

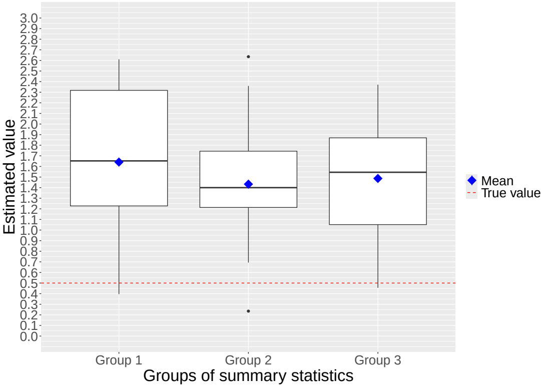


Figure: Estimation of the parameter λ

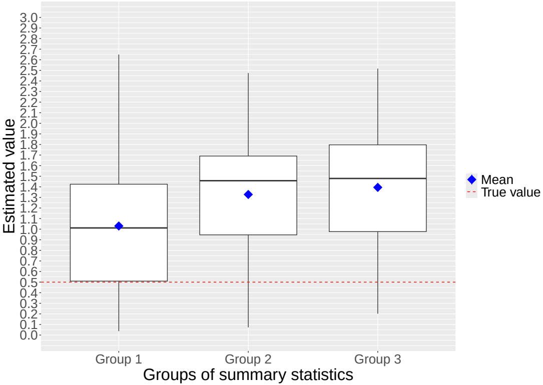


Figure: Estimation of λ with $g^* = 1$

$\Rightarrow$ Difficulty to estimate λ.

Conclusions

Advantage of the OD-HSMM : Realistic model taking into account the fact that the seed bank is hidden and that seeds can remain in soil for several years.

Estimation with ABC algorithm :

- Good estimation quality (except for λ).
- Difficulty to estimate λ $\Rightarrow$ to be explored by studying the correlations between parameters.
- Hard to choose the best group of summary statistics $\Rightarrow$ to be chosen with weight analysis.

Perspectives

Extend OD-HSMM to the case where multiple OD-HSMMs interact in order to model colonization between multiple fields (Le Coz et al. 2019).

Main difficulties :

- How to properly define the model?
- How to aggregate data from other fields?
- ...?

Thank you for your attention !

References I

Barbu, V-S., and N. Limnios. 2008. *Semi-Markov Chains and Hidden Semi-Markov Models towards Applications.* Springer.

Engel, C., and J.-D. Hamilton. 1990. “Long Swings in the Dollar: Are They in the Data and Do Markets Know It?” *The American Economic Review* 80 (4): 689–713.

Hanski, I. A., and O. E. Gaggiotti. 2004. “5 - Application of Stochastic Patch Occupancy Models to Real Metapopulations.” In *Ecology, Genetics and Evolution of Metapopulations,* 105–132. Burlington: Academic Press.

Jabot, F., T. Faure, and N. Dumoulin. 2013. “EasyABC: performing efficient approximate Bayesian computation sampling schemes using R.” *Methods in Ecology and Evolution* 4 (7): 684–687.

Le Coz, S., P. O. Cheptou, and N. Peyrard. 2019. “A spatial Markovian framework for estimating regional and local dynamics of annual plants with dormancy.” *Theoretical Population Biology* 127:120–132.

References II

Le Strat, Y., and F. Carrat. 1999. “Monitoring epidemiologic surveillance data using hidden Markov models.” *Statistics in Medicine* 18 (24): 3377–3513.

Lenormand, M., F. Jabot, and G. Deffuant. 2013. “Adaptative approximate Bayesian computation for complex models.” *Computational Statistics.*

McClintock, B.-T., R. Langrock, O. Gimenez, E. Cam, D.-L. Borchers, R. Glennie, and T.-A. Patterson. 2020. “Uncovering ecological state dynamics with hidden Markov models.” *Ecology Letters* 23 (12): 1878–1903.

Pluntz, M., S. Le Coz, N. Peyrard, R. Pradel, R. Coquet, and P. O. Cheptou. 2018. “A general method for estimating seed dormancy and colonisation in annual plants from the observation of existing flora.” *Ecology Letters* 21:1311–1318.

Toni, T., and M. Stumpf. 2009. “Tutorial on ABC rejection and ABC SMC for parameter estimation and model selection.”