Inference for Incompletely Observed Branching Processes

Catherine Larédo

M.I.A. INRA, Jouy-en-Josas et LPMA Paris 6-Paris 7, UMR7599

joint work with

O. David Mathématiques et Informatique Appliquées (M.I.A.), INRA, Jouy-en-Josas

A. Garnier PhD student Laboratoire d'Ecologie, Systématique et Evolution (E.S.E.) Paris 11, Orsay.

Dynstoch-Amsterdam-2007

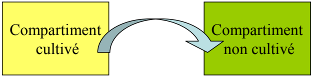
(1) Biological Problem

Context

- Many species can escape from fields and survive outside fields
- Raises numerous questions concerning
  ⇒ their foundation and origine,
  ⇒ their ability to persist,
  ⇒ their dispersal vectors.
- problems linked to Theoretical Ecology: populations dynamics in a pertubated habitat.
- problems linked to Applied Ecology: environmental risks
  ⇒ release of Genetically Modified Plants,
  ⇒ escape of transgenes in the landscape.

(2) Escape of a cultivated species

Echappement d'une espèce cultivée

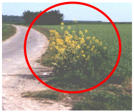


Champ, jachères

Bords de champs,
Bords de routes/chemins…

Conséquences…
- modification des communautés des bordures (invasibilité, adaptation locale, compétitivité)
- flux de gènes (pollen et graines)

Concerne de nombreuses espèces…



Colza

Tournesol

Sorgho

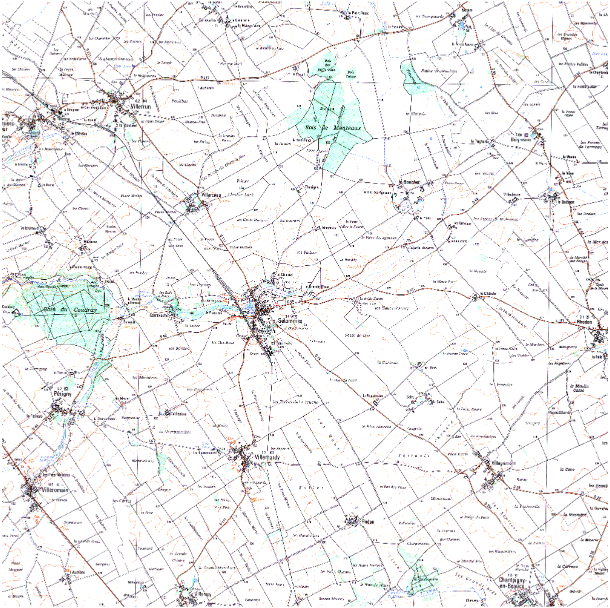
Blé

Luzerne

(3) Reasons for choosing oilseed rape (*Brassica napus .L*)

- Environmental risks associated with cultivating transgenic oilseed rape (herbicide resistant).
- Abundant populations outside fields.
- Existence of wild species able to hybridize with feral plants.
- Persistence of seeds in the soil for several years: presence of a seed bank.
- Populations might be maintained by immigration: neighbouring fields or seeds released by trucks.

(4) Region of Sélommes



(5) Production basin: Sélommes

- Ground survey of 500 feral populations on three roads and three paths
- Monthly observations from January 2001 to June 2003 et localization with G.P.S.
- Counts of the number of plants in each developmental stage within each population .
- Observations of possible covariates: presence/absence of cultivated oilseedrape, same year, herbicide treatments, favourable Winter,..

(6) Sélommes, Loir-et-Cher: Production basin for oilseed rape.
January 2001- June 2003: suivi of cultivated fields and feral populations.
Map of the experiment with the three roads and three paths.

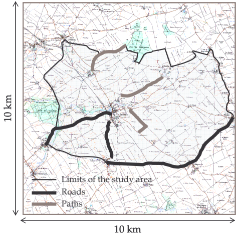


Fig. 2-I.2. : Map of the study area representing the three paths and three roads where crops and feral populations of oilseed rape were surveyed from January 2001 to June 2003.

(7) Aims

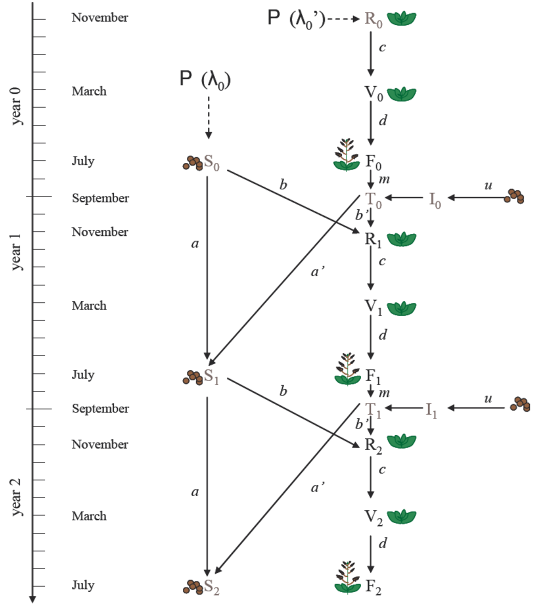
- Explore the processes involved in the dynamics of these populations.
- Link individual scale and population scale.
- Intrinsic randomness of survival and populations offsprings.
- Ecology $\Rightarrow$ Stage structured models (Leslie matrices, Caswell 2001).

- Stochastic modelling for the dynamics of these populations
- Framework: Multitype branching processes with immigration in one of the types
- Parametric inference for the demographic parameters of the laws ruling the dynamics of these populations.
- Using all the data collected in Sélommes
- Problem: one type is never observed: the seeds
  $\Rightarrow$ New problem in Statistical Inference

Inference for Incompletely Observed Branching Processes
Part 2

June 6, 2007

(8) Life cycle graph



(9) Annual plant model structured in 5 stages: $X_i = (S_i, T_i, R_i, V_i, F_i)$.

- Seeds buried in the soil i.e. in the seed bank: S_i
- Seeds on the soil: T_i
- Rosettes before Winter or non vernalised rosettes: R_i
- Rosettes after Winter or vernalised rosettes: V_i
- Mature plants carrying pods: F_i.

Model parameters

- $P(\text{seed in } S_i \rightarrow \text{non-vernalised rosette in } R_i) = b$
- $P((\text{seed in } T_i \rightarrow \text{non-vernalised rosette in } R_i) \ = b'$
- $P(\text{seed in } S_i \rightarrow \text{seed in } S_{i+1}) = a$
- $P(\text{seed in } T_i \rightarrow \text{seed in } S_{i+1}) \ = \ a'$
- $P(\text{non-vernalised rosette in } R_i \rightarrow \text{vernalised rosette in } V_i) = c$
- $P(\text{vernalised rosette in } V_i \rightarrow \text{mature plant in } F_i) = d$
- $G(.)$: Offspring distribution of plants in F_i ($\Rightarrow$ seeds in T_{i+1})
- I_{i+1}: Immigration r.v. distribution μ ($\Rightarrow$ seeds in T_{i+1}).

(10) Proposition:

- $X_i = (S_i, T_i, R_i, V_i, F_i)$ multitype branching process
- Initial distribution $\pi_0(x) = \pi_0(s, r, v, f, t)$

$$P(S_0 = s, T_0 = t)\ p_3(r/s, t)\ p_4(v/r)\ p_5(f/v)$$

with $p_4(v/r) = \mathcal{B}(r; c)(v),\ p_5(f/v) = \mathcal{B}(v; d)(f)$

$$p_3(r/s, t) = (\mathcal{B}(s; b) \star \mathcal{B}(t; b'))(r)$$

- Transition kernel $p(x; x')$

$$p(x, x') = p_1(s'/s, t, r)\ p_2(t'/f)\ p_3(r'/s', t')\ p_4(v'/r')\ p_5(f'/v')$$

$$p_2(t'/f) = \ (G^{\star f} \star \mu)(t')$$

$$p_1(s'/s, t, r) = \frac{\mathcal{M}(s; a, b) \star \mathcal{M}(t; a', b'))(s', r')}{\mathcal{B}(s; b) \star \mathcal{B}(t; b'))(r)}$$

Notation: $\mathcal{M}(N; a, b, c)(i, j, k) = \mathcal{M}(N; a, b)(i, j)$ for $i + j \leq N$.

(11) Notations

Parameters: $\theta = (\theta^1, \theta^2, \theta^3, c, d, a, b, a', b')$

- θ^1 distribution of (S_0, T_0),
- $\theta^2 \rightarrow$ offspring distribution $G(\theta^2, .)$,
- $\theta^3 \rightarrow$ immigration distribution $\mu(\theta^3, .)$.

Complete observations

- $K = 300$ independent populations during n years.
- Observations in population k at generation i: $x_i^k = (s_i^k, t_i^k, r_i^k, v_i^k, f_i^k)$.
- Observations up to generation n: $O_{0:n}^k = (x_0^k, \dots, x_n^k)$.
- Whole observations up to time n: $O_{0:n} = (O_{0:n}^1, \dots, O_{0:n}^K)$.

True value of the parameter: θ_0

(12) Asymptotics

Three possible asymptotics:

1. K prescribed and $n \to \infty$
2. n prescribed and $K \to \infty$
3. $K \to \infty$ and $n \to \infty$

Here: $K = 500$ and $n = 3$.

- reasonable to choose (2)
- Other studies often belong to case (1)
- statistical inference also investigated in case (1)
(here X_i subcritical branching with immigration $\Rightarrow$ positive recurrent)

(13) Likelihood

$$\log L(\theta; O_{0:n}) = l(\theta; O_{0:n}) = \sum_{i=0}^{5} l_i(\theta; O_{0:n}) \text{ with}$$

- $l_0(\theta; O_{0:n}) = l_0(\theta^1; O_{0:n}) = \sum_{k=1}^{K} \log p_{\theta^1}(s_0^k, t_0^k)$,
- $l_1(\theta; O_{0:n}) = \sum_{k=1}^{K} \sum_{i=0}^{n} \log(\mathcal{B}(s_i^k; b) \star \mathcal{B}(t_i^k; b'))(r_i^k)$
- $l_2(\theta; O_{0:n}) = l_2(c; O_{0:n}) = \sum_{k=1}^{K} \sum_{i=0}^{n} \log \mathcal{B}(r_i^k; c)(v_i^k)$
- $l_3(\theta; O_{0:n}) = l_3(d; O_{0:n}) = \sum_{k=1}^{K} \sum_{i=0}^{n} \log \mathcal{B}(v_i^k; d)(f_i^k)$
- $l_4(\theta; O_{0:n}) = l_5(\theta^2, \theta^3; O_{0:n}) = \sum_{k=1}^{K} \sum_{i=0}^{n} \log((G^{\star f_i^k} \star \mu\ )(t_{i+1}^k))$.
- $l_5(\theta; O_{0:n}) = l_5(a, b, a', b'; O_{0:n}) =$

$$\sum_{k=1}^{K} \sum_{i=0}^{n} \log\left(\frac{\mathcal{M}(s_i^k; a, b) \star \mathcal{M}(t_i^k; a', b')(s_{i+1}^k, r_i^k)}{\mathcal{B}(s_i^k; b) \star \mathcal{B}(t_i^k; b'))(r_i^k)}\right)$$

,

(14)Maximum likelihood estimates

$$\hat{c} = \frac{\sum_{k=1}^{K} \sum_{i=0}^{n} \ v_i^k}{\sum_{k=1}^{K} \sum_{i=0}^{n} \ r_i^k} \ ; \ \hat{d} = \frac{\sum_{k=1}^{K} \sum_{i=0}^{n} \ f_i^k}{\sum_{k=1}^{K} \sum_{i=0}^{n} \ v_i^k}. \qquad (1)$$

Under P_{θ_0} as $K \to \infty$,

- $(\hat{c}, \hat{d})$ strongly consistent, asymptotically Gaussian at rate $\sqrt{K}$.
- $l_1 + l_5 = l_1' \to$ quasilikelihood $\tilde{l}_1'$: same results for (a, b, a', b').
- $l_4 \to$ branching part:
  -loglikelihood: $\sum_{k=1}^{K} \sum_{i=0}^{n} (Log\ (G_{\theta^2}^{\star f_i^k} \star \mu_{\theta^3})(t_{i+1}^k)\ )$
  - conditional least squares or variants: (Wei & Winnicki 1990)
  $\sum_{k=1}^{K} \sum_{i=0}^{n} (t_{i+1}^k - m_{\theta^2} f_i^k - u_{\theta^3})^2$,
  (m_{θ^2}: mean of G_{θ^2} and u_{θ^3}: mean of μ_{θ^3}).
- Consistent and asymptotically Gaussian estimators of $(m_{\theta^2}, u_{\theta^3})$

Conclusions: Standard study, estimation at rate $\sqrt{K}$.
Remark: Asymptotics for Markov chains $K = 1; n \to \infty$:
$\Rightarrow$ would lead to similar results.

(15) Framework

Incomplete Observations

- Impossible in practice to observe S_i (nb of seeds in the seed bank) and T_i (nb of seeds on the soil).
- Requires to study the process $\{Y_i = (R_i, V_i, F_i); i = 1, \ldots, n\}$.
- (Y_i) is no longer Markov.
- (Y_i) is not linked to a Hidden Markov Model since (S_i, T_i) does not evolve independently

New statistical problem

- What parameters are identifiable when only (Y_i) is observed? (i.e. $(y_i^k); i = 1 \ldots n; k = 1 \ldots K)$)
- How to estimate these parameters?
- Properties of these estimators?
- Non standard inference pb $\Rightarrow$ requires a specific study.

(16) Poisson case model

* Informative example leading to explicit computations
* Analogy with the Kalman filter

Assumptions

- Offspring distribution G: Poisson law $\mathcal{P}(m)$
- Immigration distribution μ in type T_i: Poisson $\mathcal{P}(u)$
- Initial distribution of S_0 (seeds in the seed bank): Poisson $\mathcal{P}(\sigma)$
- Initial distribution of T_0 (non-vernalised rosettes): Poisson $\mathcal{P}(\tau)$
- S_0 and T_0 are independent r. v.

(17) Probabilistic properties

Notations

- Recap $\mathcal{F}_i = \sigma((S_k, T_k, R_k, V_k, F_k); k = 0, \dots, i)$.
- Define $\mathcal{G}_i = \sigma(R_k, V_k, F_k); k = 0, \dots, i)$.
- Set $Y_k = (R_k, V_k, F_k)$.
- Set $\Lambda_0 = a\sigma + a'\tau$ and for $i \geq 1$,
- $\Lambda_i = a^i \Lambda_0 + a' u \frac{1-a^i}{1-a} + a' m(F_{i-1} + aF_{i-2} + a^2 F_{i-3} + \dots a^{i-1} F_0)$
- $\Lambda'_i = mF_i + u$ for $i \geq 0$,

Theorem

Under Assumptions (A1)-(A2), $Y_i = (R_i, V_i, F_i)$ satisfies

- initial distribution is $\tilde{\pi}_0(y) = \mathcal{P}(b\sigma + b'\tau)\ p_4(v/r)\ p_5(f/v)$
- conditional distribution $\mathcal{L}(Y_{i+1}/\mathcal{G}_i)$,
  $P(Y_{i+1} = (r', v', f')/\mathcal{G}_i) = \mathcal{P}(b\Lambda_i + b'\Lambda'_i)(r')\ p_4(v'/r')\ p_5(f'/v')$

Explicit dependence on the past up to time 0 through the r.v. F_i

Rk: Conditionally on $\mathcal{G}_i$, S_{i+1} and T_{i+1} independent $\mathcal{P}(\Lambda_i)$, $\mathcal{P}(\Lambda'_i)$.

(18) Incomplete Model Likelihood

Notations

Observations: $\tilde{O}_{0:n} = (y_i^k; i = 1 \dots n; k = 1 \dots K)$
Parameter: $\theta = (\sigma, \tau, m, u, c, d, a, b, a', b')$
Define: $\lambda_i^k(\theta)$ and $\lambda_i'^k(\theta)$ realizations of $\Lambda_i(\theta), \Lambda_i'(\theta)$ in population k.
Define $\Phi_i = b\Lambda_i + b'\Lambda_i'$ and $\varphi_i^k(\theta) = b\lambda_i^k(\theta) + b'\lambda_i'^k(\theta)$

Likelihood for one population

- $L(\theta; y_0^k, \dots, y_n^k) = \tilde{\pi}_0(\theta; y_0^k) \prod_{i=0}^{n-1} P_\theta(Y_{i+1} = y_{i+1}^k / y_i^k, \dots, y_0^k)$.
- $P_\theta(Y_{i+1} = y_{i+1}^k / y_i^k, \dots, y_0^k) =$
  $\mathcal{P}(\varphi_i^k(\theta))(r_{i+1}^k)\ p_4(\theta; v_{i+1}^k / r_{i+1}^k)\ p_5(\theta; f_{i+1}^k / r_{i+1}^k)$

Loglikelihood $\tilde{l}(\theta, \tilde{O}_{0:n})$ associated with $\tilde{O}_{0:n}$

- $\tilde{l}(\theta, \tilde{O}_{0:n}) = \sum_{i=0}^{4} \tilde{l}_i(\theta, \tilde{O}_{0:n})$ with
- $\tilde{l}_0(\theta, \tilde{O}_{0:n}) = \sum_{k=1}^{K} \log \mathcal{P}(b\sigma + b'\tau)(r_0^k)$.
- $\tilde{l}_2(\theta, \tilde{O}_{0:n}) = l_2(c, O_{0:n})$; $\tilde{l}_3(\theta, \tilde{O}_{0:n}) = l_3(d, O_{0:n})$.
- It remains to study $\tilde{l}_0(\theta, \tilde{O}_{0:n})$ and $\tilde{l}_4(\theta, \tilde{O}_{0:n})$.

(19)Study of $\tilde{l}_0(\theta, \tilde{O}_{0:n})$,$\tilde{l}_4(\theta, \tilde{O}_{0:n})$

Preliminaries

- Set $\mu = \Lambda_0 = a\sigma + a'\tau$ and $\nu = b\sigma + b'\tau$
- $\lambda_i^k = a^i\mu + a'u\frac{1-a^i}{1-a} + a'm(f_{i-1}^k + af_{i-2}^k + a^2 f_{i-3}^k + \dots a^{i-1} f_0^k)$
- $\lambda_i'^k = (mf_i^k + u)$ and $\varphi_i^k = b\lambda_i^k + b'\lambda_i'^k$
- $\tilde{l}_0(\theta, \tilde{O}_{0:n}) = \tilde{l}_0(\nu; r_0^1, \dots r_0^k) \Rightarrow \nu$ identifiable
- MLE: $\widehat{\nu} = \frac{\sum_{k=1}^K r_0^k}{K}$ consistent asympt. Gaussian at rate $\sqrt{K}$.

Estimation of $\theta = (\mu, m, u, a, b, a', b')$

(c,d) omitted now.

- All the difficulties are in the study of this last term
- $\tilde{l}_4(\theta, \tilde{O}_{0:n}) = \tilde{l}_4(\phi_i^k, \tilde{O}_{0:n}) = \sum_{i=1}^K(-\varphi_i^k + r_i^k \log \varphi_i^k)$
- What parameters are identifiable given that all the available information is contained in the φ_i^k?

(20)Estimating θ from $\tilde{l}_4(\theta, \tilde{O}_{0:n})$

Define $\mathcal{K}(P, Q)$ as the Kullback-Leibler information of Q w.r.t. P
Recap If $P \sim \mathcal{P}(\lambda_0)$, $Q \sim \mathcal{P}(\lambda)$,
$\mathcal{K}(P, Q) = \lambda - \lambda_0 - \lambda_0(\log \lambda - \log \lambda_0)$.

Theorem

Let θ_0 be the true parameter value. Then, almost surely under P_{θ_0}, as $K \to +\infty$

$$\frac{1}{K}\tilde{l}_4(\theta, \tilde{O}_{0:n}) \to -E_{\theta_0} \sum_{i=0}^{n-1} \mathcal{K}(\mathcal{P}(\Phi_i(\theta_0), \Phi_i(\theta)).$$

Φ_i : random variables depending on θ and on the r.v. $F_0, \ldots, F_i$.

(21) Identifiability

Corollary

1. If $n = 0$, only $\nu = b\sigma + b'\tau$ is identifiable
2. If $n = 1$, the identifiable parameters are: $\nu, b\mu + b'u, b'm$
3. If $n = 2$, the identifiable parameters are: $\nu, b\mu + b'u, b'm, ab\mu + b'u + a'bu, \frac{a'b}{b'}$
4. If $n \geq 3$, the identifiable parameters are: $\nu, b\mu, b'u, b'm, a, \frac{a'b}{b'}$.

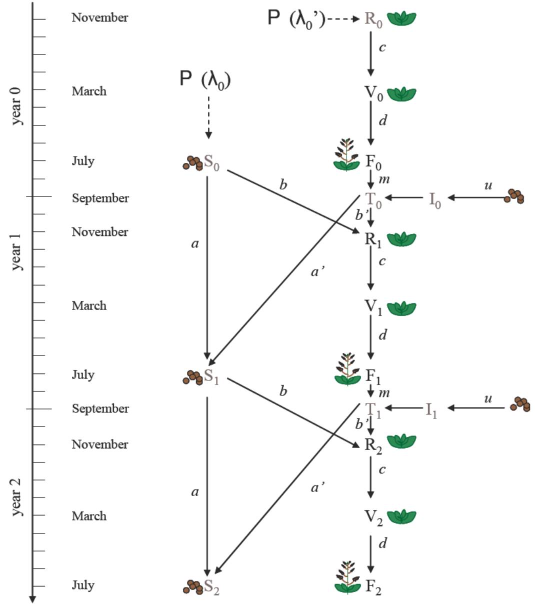
Consequences

- Only the combinations appearing in (3) can be estimated.
- Natural here to use E.M. or Bayesian approaches: ongoing work.
- Very important to ecologists: how parameters are linked using the available information $\Rightarrow$ impossible with E.M
- Observations collected in the ground survey: $n = 2$.
- Many observed populations $K = 500 \Rightarrow$ rate $\sqrt{K}$.
- K large $\Rightarrow$ Ability to introduce covariates in the estimation.

Inference for incompletely observed branching processes

Dynstoch 2007

Life cycle graph



Estimation of the parameters

	Estimates of the x_i's			Li
	"All" $n = 595$	"With crop" $n = 55$	"Without crop" $n = 540$	
$\hat{x}_1$	2.8 [0.9, 4.8]	7.4 [1.6, 13.3]	1.8 [0.3, 3.4]	x_1
$\hat{x}_2$	1.5 [-0.1, 3.1]	-2.4 [-22.6, 17.8]	1.5 [0.17, 2.9]	x_2
$\hat{x}_3$	16.4 [5.7, 27.1]	7.8 [-8.3, 23.8]	18.2 [7.7, 28.6]	x_3
$\hat{x}_4$	12.5 [3.1, 21.8]	119.0 [12.4, 225.7]	6.4 [0.3, 12.4]	x_4
$\hat{\tau}$	22.7	17.9	19.8	

Link with the model parameters

- $x_1 = b'm$: "Efficient fecundity"
- $x_2 = a'bm$:"Efficient delayed fecundity "
- $x_3 = ub' + b\lambda_0$: seeds in the seed bank + immigrating seeds.
- $x_4 = ub' + a'um + a\lambda_0$

Estimated values for the model parameters

Known values from the bibliography (Claessen, data from U.K.)

- Incorporation in the seed bank: $\hat{a'} = 0.006$
- Annual survival in the seed bank: $\hat{a} = 0.15$
- Emergence rate from the seed bank: $\hat{b'} = 0.0043$

Derived estimated values for the other model parameters

- R $\rightarrow$ V: $\hat{c} = 0.31$ (favourable Winter); $\hat{c} = 0.14$ (hiver non favorable);
- V $\rightarrow$ F : $\hat{d} = 0.05$
- Offspring distribution G : mean $\hat{m} = 700$
- Immigration: $\hat{u} = 110$ seeds/m (with crop); $\hat{u} = 25$ seeds/m (without crop)
- S $\rightarrow$ R: $\hat{b} = 0.36$
- Seeds in the seed bank at time 0: $\hat{\lambda}_0 = 25$.