Estimation of Plant Demographic Parameters from Stage-Structured Censuses

Olivier David,1,* Aurélie Garnier,2 Catherine Larédo,1,3 and Jane Lecomte2

1UR 341 Mathématiques et Informatique Appliquées, INRA, F-78350 Jouy-en-Josas, France
2Université Paris-Sud 11, Laboratoire Ecologie Systématique et Evolution, UMR 8079, 91405 Orsay, France; CNRS, 91405 Orsay, France; AgroParisTech, 75231 Paris, France
3Laboratoire de Probabilités et Modèles Aléatoires, CNRS–UMR 7599, Paris, France
*email: Olivier.David@jouy.inra.fr

Summary. This article presents some statistical methods for estimating the parameters of a population dynamics model for annual plants. The model takes into account reproduction, immigration, seed survival in a seed bank, and plant growth. The data include the number of plants in several developmental stages that were measured in a number of populations for a few consecutive years; they are incomplete since seeds could not be counted. It is assumed that there are no measurement errors or that measurement errors are binomial and not frequent. Some statistical methods are developed within the framework of estimating equations or Bayesian inference. These methods are applied to oilseed rape data.

Key words: Bayesian inference; Branching processes; Ecology; Estimating equations; Importance sampling; Matrix population models; Oilseed rape; Population dynamics; State-space models.

1. Introduction

The statistical analysis of ecological data raises many problems, in particular when dealing with demographic field data. These data are generally complex, with temporal and spatial dimensions, and highly variable due to demographic and environmental stochasticity. Moreover, the observations are often incomplete (because of practical or biological constraints) or subject to measurement errors. Moreover, environmental conditions are generally not (or little) known or controlled (Clark and Bjornstad, 2004). Specific statistical methods are therefore necessary to correctly estimate the demographic parameters of the population dynamics models developed for risk assessment and population management (Caswell, 2001; Haccou, Jagers, and Vatutin, 2005).

Three main types of statistical methods have been developed to deal with these methodological difficulties. In the first type of method, population dynamics are modeled using the formalism of (multitype) branching processes, possibly including immigration, and measurement errors are neglected. The parameters of these models are then estimated by maximum likelihood, CLS, estimating equations (EE) or Bayesian inference (Bhat and Adke, 1981; Wei and Winnicki, 1990; Bai and Durairajjan, 1996; Lalam and Jacob, 2004; González et al., 2008). This method of EE only requires the knowledge of the expectation and variance of observations and has also been referred to as quasi-likelihood (McCullagh and Nelder, 1989), M- or Z-estimators (Chapter 5, Van der Vaart, 1998) and mean and variance models. The second type of method is based on a deterministic matrix model with stochastic observation errors (Gross, Craig, and Hutchison, 2002; Hooten et al., 2007). The third type of method is based on state-space models that couple a state process (i.e., population dynamics) and an observation process (i.e., observation errors) (Durbin and Koopman, 2001; Buckland et al., 2007). Both measurement errors and missing data (i.e., the absence of observations for some developmental stages in the population) can be specified in the observation process. The computation of the likelihood requires integrations with respect to the unobserved states of the state process. When some conditions are satisfied (Gaussian errors, constant variance, linear and Markovian state process), the likelihood can be computed using the Kalman filter (Durbin and Koopman, 2001). More complicated state-space models can be analyzed thanks to computer-intensive methods, for example, Monte Carlo likelihood methods (de Valpine, 2004), sequential importance sampling (IS), and Markov Chain Monte Carlo (MCMC) (Clark and Bjornstad, 2004; Buckland et al., 2007). The reliability of the results obtained with such computer-intensive methods is not always sufficient due to issues of slow algorithm convergence or particle depletion. When there are no observations for some stages (or types) of the state process, model parameters may not all be identifiable (Chapter 8, Seber and Wild, 1989). The Bayesian state-space methods in Buckland et al. (2007) and Clark and Bjornstad (2004) can take account of some hidden stages.

The objective of this article is to present some new statistical methods developed to analyze demographic data from plant populations. We consider populations of annual plants with a soil seed bank and that experience seed immigration. The data include the number of plants in several developmental stages that were observed in a number of populations for a few years. The data are incomplete since seeds were not counted. Within a stepwise approach, we first developed a maximum likelihood method based on the assumptions that...
the offspring and immigration distributions are Poisson distributions and that there are no measurement errors (see the companion paper of Larédou, David, and Garnier, 2009). The present article considers situations where the offspring and immigration distributions are more dispersed than Poisson distributions and where population size may be measured with errors for a few large populations. Reproduction and immigration are modeled using Poisson distributions with means following gamma distributions (Melbourne and Hastings, 2008). The resulting state-space model is not standard (Section 2). First, the case when there are no measurement errors is considered, and statistical methods based on EE or the Bayesian framework are developed using explicit results on the distribution of observations (Section 3). These methods are then adapted to take account of measurement errors due to sampling (Section 4).

The developed methods are applied to demographic data on feral oilseed rape (Section 5). Oilseed rape is a plant model used to evaluate the impact of the persistence and expansion of feral populations (i.e., populations escaped from crops) on transgene escape in the context of introducing transgenic crops into agroecosystems (Hails, 2002). However, the life cycle of feral oilseed rape is complex (Garnier, Deville, and Lecomte, 2006), especially due to the possible ability of seeds to survive in a soil seed bank and the possible existence of different sources of seed immigration into the populations (adjacent oilseed rape crops (Pivard et al., 2008) and grain trucks). The analyzed data were collected within about 200 feral populations over three years. These data present three main difficulties: seeds were not observed, data are very dispersed, and counts were not exhaustive for large and dense populations.

2. State-Space Model

2.1 State Process

Plants are classified into rosettes (R) and mature plants (F). An age structure is taken into account for seeds because seed parameters may depend on age. Seeds are classified into new seeds (S) and age i seeds (S_i), 1 ≤ i ≤ l. The seeds that are at least one year old are referred to as old seeds and are located in a soil seed bank.

According to Buckland et al. (2007), biological processes are divided into a reproduction process, an immigration process, a seed process, and a growth process (Figure 1). These processes are presented below for a given population. The term year corresponds to a life cycle; it starts at the birth of new seeds and ends just before the birth of the new seeds of the next generation. All the state variables of the model vary in N.

**Reproduction process.** Mature plants release seeds at the end of summer. Since our data are very dispersed (see Web Appendix F), fecundity is modeled using a Poisson distribution with a mean following a gamma distribution (Melbourne and Hastings, 2008). Thus, the number FS_{t}^{(j)} of new seeds that are released during year t by a plant j that was mature at the end of the year t – 1 (see Figure 1) satisfies the following equation:

\[ FS_{t}^{(j)} | M_{t}^{(j)} \sim \mathcal{P}(M_{t}^{(j)}), \quad M_{t}^{(j)} \sim \mathcal{G}(\alpha_{m}, \theta), \]

where \( \mathcal{P}(\lambda) \) denotes a Poisson distribution with mean \( \lambda \), \( \mathcal{G}(\alpha, \theta, x) \) denotes a gamma distribution with probability density function \( \theta^{\alpha}e^{-\theta x}/\Gamma(\alpha) \), \( \alpha_{m} > 0 \) and \( \theta > 0 \). The distribution of \( FS_{t}^{(j)} \) is also equal to \( NW(\alpha_{m}, \theta)(\theta + 1) \), where \( NW(\alpha, \theta) \) denotes a negative binomial distribution with parameters \( \alpha \) and \( p \) (see Web Appendix A). Let the mean per-capita fecundity of a mature plant be \( m = E(FS_{t}^{(j)}) = \alpha_{m}/\theta \). It can be shown that

\[ var(FS_{t}^{(j)}) = \left(1 + \frac{1}{\theta}\right)m. \]  

Let \( F_{t} \) denote the number of plants that are mature at the end of year \( t \) and \( FS_{t} \) denote the number of new seeds that are released year \( t \) by all the plants mature at the end of year \( t – 1 \) (see Figure 1). Plants are assumed to reproduce independently. Thus, \( FS_{t} \) is the sum of \( F_{t-1} \) independent negative binomial distributions and follows the distribution \( NW(\alpha_{m}, \theta; (\theta + 1)) \) or the Poisson–gamma distribution:

\[ FS_{t} \sim NW(\alpha_{m}, \theta; (\theta + 1)) \text{ or } FS_{t} \sim \mathcal{PG}(\alpha_{m}, \theta). \]

**Immigration process.** Some new seeds can immigrate into the population at the beginning of a life cycle. Since our data are very dispersed, the number IS_{t} of new seeds that immigrate into the population in year \( t \) is modeled using a Poisson–gamma distribution

\[ IS_{t} \sim \mathcal{PG}(\alpha_{i}, \theta), \]

where \( \alpha_{i} > 0 \). The distribution of \( IS_{t} \) is also equal to \( NW(\alpha_{i}, \theta)(\theta + 1) \). Let the mean number of seeds immigrating into the population each year be \( u = E(IS_{t}) = \alpha_{i}/\theta \). It can be shown that

\[ var(IS_{t}) = \left(1 + \frac{1}{\theta}\right)u. \]

The variable IS_{t} is assumed to be independent of the variables of previous years, of FS_{t} and of the current number of old seeds. In order to simplify the model, the offspring and
immigration distributions are assumed to have the same parameter \( \theta \), that is, they have the same ratio of expectation over variance.

**Seed process.** A seed can enter/remain in a seed bank, germinate in autumn or die. More precisely, an age \( i \) seed (\( 0 \leq i \leq l \)) at the beginning of year \( t \) can: (1) survive and become an age \( i + 1 \) seed at the beginning of year \( t + 1 \) with probability \( a_i \) (with the convention that \( l + 1 = l \) for age \( l \) seeds); (2) emerge as a rosette with probability \( e_i \); or (3) die with probability \( 1 - a_i - e_i \) (Figure 1). Let \( S_d \) denote the number of age \( i \) seeds present in the population at the beginning of year \( t \). Note that the number \( S_{d0} \) of new seeds is the sum of the number of seeds produced within the population and the number of seeds resulting from immigration:

\[
S_{d0} = FS_t + IS_t.
\]

Let \( R_d \) denote the number of rosettes that emerged from age \( i \) seeds at the beginning of year \( t \) and \( D_d \) denote the number of age \( i \) seeds that died during year \( t \). Seeds are assumed to evolve independently. Thus, the resulting number of old seeds, rosettes, and dead seeds follow a multinomial distribution

\[
(S_{d1+t+1}, R_d, D_d) \mid S_d \sim \mathcal{M}(S_d, a_i, e_i, 1 - a_i - e_i),
\]

where \( \mathcal{M}(N, p_1, p_2, 1 - p_1 - p_2) \) denotes a three-dimensional multinomial distribution with parameters \( N, p_1, p_2 \).

**Growth process.** An emerged rosette can either: (1) verbalize in winter, bolt in spring and finally become a mature plant, with probability \( c \); or (2) die. Annual plants are considered in this article, in other words, plants that die after seed production. Thus, the mature plants of a given year all originate directly from the rosettes of the same year. Rosettes are assumed to evolve independently. Thus, the number of mature plants follows a binomial distribution

\[
F_t \mid R_t \sim \mathcal{B}(R_t, c),
\]

where \( R_t \) is the number of rosettes present during year \( t \) and \( \mathcal{B}(N, p) \) denotes a binomial distribution with parameters \( N, p \). Note that \( R_t \) is the sum of the numbers of rosettes emerged from seeds of all ages: \( R_t = \sum R_{di} \).

Let \( S_{1:t:t} \) denote the vector \( (S_{1t}, \ldots, S_{10}) \) (vectors here are considered as rows). It can be shown that the conditional expectation of the vector \( (F_t, S_{1:t:t}) \) follows a matrix population model with immigration (Caswell, 2001; Buckland et al., 2007):

\[
E((F_t, S_{1:t:t})' \mid (F_{t-1}, S_{1:t-1})') = C \cdot B \cdot (A(F_{t-1}, S_{1:t-1})' + (u, 0, \ldots, 0)'),
\]

where prime denotes transposition. In the latter equation, the matrix \( A \) is a reproduction projection matrix, that is, the diagonal matrix of size \( l + 1 \) with the first diagonal element equal to \( m \) and the other diagonal elements equal to 1. The matrix \( B \) is a seed projection matrix, that is, the square matrix of size \( l + 1 \) with the first row equal to \( (e_0, 0, \ldots, 0) \), diagonal elements equal to \( a_0, a_1, \ldots, a_{l-1} \), and the other elements equal to zero. The matrix \( C \) is a growth projection matrix, that is, the diagonal matrix of size \( l + 1 \) with the first diagonal element equal to \( c \) and the other diagonal elements equal to 1.

Random variables from different populations are assumed to be independent. This state process is denoted by Model 1. This model can be refined to account for the fixed effects of factors or covariates. For example, the fecundity \( m \) or the survival probability \( c \) could be modeled as functions of climatic or mowing covariates; the rate of immigration could be modeled as a function the area of the population patch (see Section 5).

**2.2 Observation Process**

**Missing data.** Observations are available for rosettes and mature plants for the successive years \( t = 0, \ldots, T(T > 1) \) and for \( K \) populations. Thus, the observed variables are \( R_{it} \) and \( F_{ik} \) for \( t = 0, \ldots, T \) and \( k = 1, \ldots, K \), where the population index \( k \) is introduced in notations when useful. However, the numbers of seeds, both new and old, are not observed.

We assume that populations are established by the immigration of new seeds (i.e., there is no pre-existent seed bank) and that immigration can start at the beginning of the year \( T' < 0 \). This assumption allows us to consider populations as replications (i.e., without different seed bank histories) and to simplify the model (i.e., we do not have to consider parameters for the initial seed banks).

**Measurement errors.** In this article, we consider both cases when \( R_t \) and \( F_t \) are observed with/without measurement errors. Measurement errors are due to sampling and occur in a few large populations only. When \( R_t \) is large and plants are dense, it is difficult to count plants exhaustively, so that rosettes and mature plants are counted in a fraction of the population only, for example, in regularly spaced quadrats. Otherwise, rosettes and mature plants are counted in the whole population and it is assumed that there are no measurement errors. The observed fractions of a population are located at the same position for rosettes and mature plants but may be located at different positions in different years.

This state-space model is not standard. Some variables of the state process are not observed, process and observation errors are not Gaussian and do not have constant variance, and the occurrence of observation errors depends on population size.

**3. Analysis with Missing Data without Measurement Errors**

**3.1 Likelihood**

We first consider the case when there are no measurement errors. Let \( Y_t = (R_t, F_t) \) denote the vector of observations for one population and year \( t \), and let \( Y_{0t} = (Y_0, \ldots, Y_t) \) denote the vector of observations for one population until time \( t \). The likelihood for one population can be written as

\[
\mathcal{L} = P(Y_{0T}) = P(Y_0) \prod_{t=1}^{T} P(Y_t \mid Y_{0t-1})
\]

\[
= P(R_0) \left[ \prod_{t=1}^{T} P(R_t \mid Y_{0t-1}) \right] \left[ \prod_{t=0}^{T} P(F_t \mid R_t) \right].
\]

The likelihood for all the populations is the product of the likelihoods of populations since populations are assumed to be independent.
In equations (5), the distribution $P(R_t | Y_{0:t-1})$ depends on all the vectors $Y_0, \ldots, Y_{t-1}$ because the two-dimensional observed process $(R_t, F_t)$ is not Markovian. When the offspring and immigration distributions are Poisson distributions, the distribution of $R_t | Y_{0:t-1}$ is a Poisson distribution with a mean that has a simple explicit expression (Larédò et al., 2009). It is more difficult to obtain explicit results on this distribution when the offspring and immigration distributions are negative binomial distributions. Thus, to simplify mathematical calculations, the state process is approximated by a short-term memory model, denoted by Model 2: the distribution of $R_t$ is modeled using the conditional distributions associated with the solid arrows of Figure 1 only. In other words, the effects on $R_t$ of reproduction and immigration events occurring before $t-1$ are neglected. Note that the parameters $a_1$, $i \geq 1$ and $e_i$, $i \geq 2$ are not involved in Model 2.

**Proposition 1.** Under Model 2, for $t \geq 2$,

$$R_t | Y_{0:t-1} = NB_{1t} + NB_{2t},$$

where $NB_{1t}$ and $NB_{2t}$ are two conditionally independent random variables with distribution $NB_{1t} | Y_{0:t-1} \sim\ NB(a_{0m}F_{t-1} + a_0 R_{t-1}, \frac{\theta}{\theta + e_0})$ and $NB_{2t} | Y_{0:t-1} \sim\ NB(a_{0e}F_{t-1} + a_0, \frac{\theta}{\theta + e_0})$.

See Web Appendix B for a proof. The fact that the Poisson and gamma distributions are conjugate is used in this proof. In equation (6), the variables $NB_{1t}$ and $NB_{2t}$ are due to the emergence of the old seeds and new seeds, respectively, present at the beginning of year $t$.

Equation (6) depends on model parameters through four quantities only, for example, $e_0$, $\theta$, $a_{0m}$, $a_{0e}$, $a_{0e}/e_0$, and $e_0/\theta$. This is due to the absence of observations for seeds, making it impossible to identify all of the parameters (see Larédò et al., 2009, for a study of identifiability when the immigration and offspring distributions are Poisson distributions). The parameters $e_0$, $\theta$, $a_{0m}$, and $a_{0e}$ can be interpreted as the mean number of rosettes resulting from the reproduction of one plant and from immigration, respectively.

The survival probability $c$ appears in equations (5) in the term $\prod_{t=0}^{T-1} P(F_t | R_t)$ only. It can be estimated using a binomial analysis and, in particular, the maximum likelihood estimate of $c$ is equal to $\hat{c} = (\sum_{t=k}^{T} F_{tk})/(\sum_{t=k}^{T} R_{tk})$. The rest of this section explains how the equations (5) and (6) can be used to estimate the other parameters.

### 3.2 Estimating Equations

The method of EE is based on the expectation and variance of $R_t | Y_{0:t-1}$ only. Thus, it is applicable under second moment assumptions on observations rather than full distributional assumptions. For dispersed data, the EE method can be implemented using the corollary of Proposition 1 given below that gives simple expressions of the expectation and variance of $R_t | Y_{0:t-1}$ (see Web Appendix B for a proof). The symbol $\ll$ means “small with respect to.”

**Corollary 1.** Under Model 2, if $\theta/e_0 \ll 1$ and $a_{0m}F_{t-1} + a_0 R_{t-1}$, then for $t \geq 1$,

$$E(R_t | Y_{0:t-1}) = (a_{0e}/e_0)R_{t-1} + e_0mF_{t-1} + e_0u,$$

$$var(R_t | Y_{0:t-1}) = (1 + a_{0e}/e_0)(a_{0e}/e_0)R_{t-1}$$

$$+ (e_0/\theta)(e_0mF_{t-1} + e_0u).$$

The condition $\theta/e_0 \ll 1$ in Corollary 1 implies that the offspring and immigration distributions have large variances (see equations (1) and (4)). The condition $a_{0m}F_{t-2} + a_0 \ll R_{t-1}$ tends to be satisfied when $\theta$ and $c$ are small since $a_{0m} = m/\theta$ and $a_0 = u/\theta$ then tend to be small and $F_{t-2}$ then rarely takes large values. The quantity $e_0/\theta$ only appears in the conditional variance of observations in equations (7). The EE estimators of $\tau = (e_{0m}, e_{0u}, a_{0e}/e_0)$ and $e_0/\theta$ are then the solution to the following system of equations, see McCullagh and Nelder Section 9.4, 1989, for $K = 1$ or Hu et al. (pages 67 and 68, 2004, for independent observations):

$$\sum_{k=1}^{K} \sum_{t=1}^{T} \frac{\partial E(R_{tk} | Y_{0:t-1,k})}{\partial \tau_j} \times \frac{R_{tk} - E(R_{tk} | Y_{0:t-1,k})}{var(R_{tk} | Y_{0:t-1,k})} = 0,$$

$$\sum_{k=1}^{K} \sum_{t=1}^{T} \frac{\partial var(R_{tk} | Y_{0:t-1,k})}{\partial \theta} \times \frac{e_{tk}^2 - var(R_{tk} | Y_{0:t-1,k})^2}{var(R_{tk} | Y_{0:t-1,k})^2} = 0,$$

where $j = 1, \ldots, 3$, $E(R_{tk} | Y_{0:t-1,k})$ and $var(R_{tk} | Y_{0:t-1,k})$ are given by equations (7). This system of equations is solved numerically.

The standard errors of estimators can be estimated using asymptotic arguments (see Chapter 5, Van der Vaart, 1998), the asymptotic framework here being $K \rightarrow \infty$:

$$\hat{var}(\tau) = \left( \sum_{k=1}^{K} X_k^T \hat{V}_k^{-1} X_k \right)^{-1},$$

where $X_k$ is the $T \times 3$ matrix with rows equal to $(F_{t-1,k,1}, R_{t-1,k})$, $t = 1, \ldots, T$, $V_k$ is the $T \times T$ diagonal matrix with diagonal elements equal to $var(R_{tk} | Y_{0:t-1,k})$, $t = 1, \ldots, T$, and $\hat{V}_k$ is the estimate of $V_k$ obtained using equations (7) and parameter estimates. Alternatively, robust estimates of standard errors can be obtained using the non-parametric bootstrap (Chapter 21, Fox, 2008). In this method, bootstrap samples of populations are drawn and parameters are estimated by EE for each bootstrap sample. The resulting sample of estimates provides robust estimates of standard errors and confidence intervals.

The EE estimator of $\tau$ may be biased since equations (7) are approximations. The bias of estimators was quantified using simulations. To limit numberical problems, the conditional least squares (CLS) estimator of $\tau$ was used in these simulations. It is the solution of the first three equations of System 8, in which the denominator $var(R_{tk} | Y_{0:t-1,k})$ is replaced by 1. The biases of the CLS estimators of the components of $\tau$ were smaller than 10% of the parameter values (see Web Appendix D). Moreover, the correlation between the CLS estimators of $e_{0m}$ and $a_{0e}/e_0$ was high when the plant survival probability $c$ was assumed to be constant among populations and years, but decreased when $c$ varied randomly between populations and years. The bootstrap provides estimates of the expectations and biases of estimators, but bias corrections should be used with caution (see Chapter 10, Efron and Tibshirani, 1993).

### 3.3 Bayesian Inference

Methods based on the likelihood, such as maximum likelihood and Bayesian methods, generally provide better estimators than EE when all the assumptions of the model are satisfied. In the present article, only Bayesian methods are presented.
Bayesian inference is based on the posterior distribution of model parameters (Robert, 2001). According to Bayes’ theorem, this distribution is proportional to the product of the prior distribution and the likelihood

\[ P(\beta | Y_{0:T}) = P(\beta)P(Y_{0:T} | \beta) / P(Y_{0:T}), \]

where \( \beta \) is the vector of model parameters, \( P(\beta) \) is the prior distribution of \( \beta \) and \( P(Y_{0:T} | \beta) \) is the likelihood. Since parameters are not all identifiable here, the analysis is parameterized with \( \beta = (\epsilon_0m, \epsilon_0a, a_0e_1/e_0, \theta, c) \) or \( \beta = (\alpha_m, \alpha_u, a_0e_1/e_0, c_0/\theta, c). \)

The posterior distribution cannot be explicitly calculated but the analysis can be implemented using Monte Carlo methods. Some dispersed simulated data were analyzed using OpenBUGS, a computer software for Bayesian inference using MCMC (Thomas et al., 2006). This MCMC algorithm performed well for models without seed banks but did not converge for models with seed banks. For the latter models, the analysis can be carried out using IS (Chapter 6, Robert, 2001). In this method, a large number of values of \( \beta \), denoted by \( \beta_i \), are simulated independently according to a proposal distribution \( Q \). The posterior expectation of some function \( f \) of \( \beta \), \( E(f(\beta) | Y_{0:T}) \), is then estimated by

\[ \hat{E}(f(\beta) | Y_{0:T}) = \sum_i w_i f(\beta_i) / \sum_i w_i, \]

where \( w_i = P(\beta_i)P(Y_{0:T} | \beta_i)/Q(\beta_i) \). Thus, this method provides estimates of the posterior expectations and standard deviations of the components of \( \beta \). The importance weights \( w_i \) are used to take account of the fact that the proposal distribution is different from the posterior distribution. These weights involve the likelihood that can be approximated by equations (5) and (6). The proposal distribution should have heavy tails; otherwise, the estimate of \( E(f(\beta) | Y_{0:T}) \) tends to be mainly based on a small number of \( \beta_i \)’s with large weights (Chapter 6, Robert, 2001). Note that there is no assumption on parameter values in this method, e.g., \( \epsilon_0/\theta \) is not assumed to be large.

4. Analysis with Missing Data and Measurement Errors

4.1 Measurement Errors

We have restricted ourselves to estimating methods based on the total number of rosettes and mature plants observed per population; for example, if plants are counted in several quadrats within a population, only the sum of the quadrat counts is used. For a given population, let \( Z_t \) denote the random variable equal to 1 if a sampling is carried out during year \( t \), and equal to 0 otherwise. Let \( OR_t \) and \( OF_t \) denote the random variables equal to the number of rosettes and mature plants, respectively, counted in the observed part of the population if \( Z_t = 1 \), and equal to \( R_t \) and \( F_t \), respectively, otherwise. In this section, given \( R_t \), the positions of rosettes are assumed to be drawn independently according to a uniform distribution; this assumption is referred to as Assumption 1.

Under this assumption:

\[ OR_t | R_t \sim B(R_t, d_t), \]

where \( d_t \in [0, 1] \) is the sampling fraction, i.e., the ratio of the area of the observed part over the total area of the population for year \( t \). The latter is smaller than 1 when a sampling is carried out and is equal to 1 otherwise. Rosettes survive and become mature plants with probability \( c \), thus, \( OF_t | OR_t \sim B(OR_t, c) \). In this section, the vector of observations for one population and year \( t \) is \( Y_t = (Z_t, OR_t, OF_t). \)

4.2 Likelihood and Parameter Estimation

The likelihood for one population can be written as:

\[ L = P(Z_0)P(OR_0 | Z_0) \prod_{t=1}^{T} P(Z_t | Y_{0:t-1})P(OR_t | Y_{0:t-1}, Z_t) \times \prod_{t=0}^{T} P(OF_t | OR_t). \]

(11)

Although the occurrence of sampling depends on population dynamics, \( Z_t \) is assumed here to be independent of population dynamics in order to simplify mathematical calculations; this assumption is referred to as Assumption 2. Under this assumption, \( P(Z_t | Y_{0:t-1}) \) in equation (11) does not depend on model parameters. An estimator of \( c \) is \( \hat{c} = (\sum_{t} OF_t)/ (\sum_{t} OR_t) \). The rest of this section deals with the estimation of the other parameters.

Since the distribution of \( OR_t | R_t \) is binomial, Proposition 1 leads to the following corollary.

**COROLLARY 2.** Under Model 2 and Assumptions 1 and 2, if there is no sampling in years \( t-1 \) and \( t-2 \), then for \( t \geq 2 \):

\[ OR_t | (Y_{0:t-1}, Z_t) = NB_{11} + NB_{21}, \]

where \( NB_{11} \) and \( NB_{21} \) are two conditionally independent random variables with distribution \( NB_{11} | (Y_{0:t-1}, Z_t) \sim NB(\alpha_{11}F_{t-1} + \alpha_{u} + R_{t-1} - \epsilon_0 e_1/e_0, \theta + \epsilon_0 e_1/e_0)\) and \( NB_{21} | (Y_{0:t-1}, Z_t) \sim NB(\alpha_{21}F_{t-1} + \alpha_u + \theta e_0, \theta + \epsilon_0 e_1/e_0). \)

The following proposition gives an expression of the conditional expectation of \( OR_t \) for dispersed data that is valid even when a sampling is carried out in \( t-1 \) or \( t-2 \) (see Web Appendix C for a proof).

**PROPOSITION 2.** Under Model 2 and Assumptions 1 and 2, if \( \theta/ (\epsilon_0 e_1/e_0) \ll 1 \) and \( \alpha_u F_{t-1} + \alpha_u \ll R_{t-1} \), then for \( t \geq 1 \):

\[ E(OR_t | (Y_{0:t-1}, Z_t)) = d_t (a_0 e_1/e_0) \hat{R}_{t-1} + \epsilon_0 m \hat{F}_{t-1} + \epsilon_0 a_u, \]

(13)

where \( \hat{R}_{t} = OR_t/d_t \) and \( \hat{F}_{t} = OF_t + c(1 - d_t)OR_t/d_t \).

In Proposition 2, the assumption \( \theta/ (\epsilon_0 e_1/e_0) \ll 1 \) implies that the process errors tend to be larger than the observation errors. The prediction \( \hat{F}_{t} \) of \( F_{t} \) in Proposition 2 has a simple interpretation: when there is a sampling in the population, \( \hat{F}_{t} \) is equal to the sum of: (1) the number of mature plants in the observed part of the population; and (2) a prediction of the number of mature plants in the unobserved part of the population based on the observed number of rosettes.

For dispersed data, the EE method can be implemented using equation (13) and variance formulas given in Web Appendix C. These formulas for the conditional expectation
and variance of $OR_t$ depend on $c$. This problem may be solved by using a plug-in approach, i.e., by using these formulas assuming that the value of $c$ is known and equal to $\hat{c}$. Alternatively, the system of estimating equations may be extended so that all parameters are estimated simultaneously. In a simulation study, the biases of the CLS estimators of the components of $\tau$ were smaller than 10% of the parameter values (see Web Appendix D).

The conditional distribution of observations is needed to calculate the likelihood in the IS method. This distribution for populations with no sampling in years $t - 1$ and $t - 2$ can be approximated by equation (12). Web Appendix C gives a method for calculating an approximation of this distribution for populations with a sampling in $t - 1$ and no sampling in $t - 2$. Under the assumptions of Proposition 2, it is suggested in equation (13) to approximate the conditional distribution of observations with a sampling in $t - 1$ and $t - 2$ by equation (12) with $F_{t-2}$, $R_{t-1}$ and $F_{t-1}$ predicted by $\hat{F}_{t-2}$, $\hat{R}_{t-1}$ and $\hat{F}_{t-1}$.

5. Application to Feral Oilseed Rape Data

Data. In order to estimate the demographic parameters of feral oilseed rape, a pluriannual survey of 192 feral populations was carried out on the sides of three roads and three paths in a French open-field farming landscape (Solommes, Loir-et-Cher). Rosettes and mature plants were counted in October and June, respectively, during three consecutive life cycles at the beginning of the 2000, with the exception that rosettes were not counted in the first life cycle. These life cycles are denoted by $t = 0, 1, 2$. Additional sessions of observations in February for $t = 0, 1, 2$ provided counts for vermalized rosettes, the intermediate developmental stage between rosettes and mature plants. However, when population size was large and plants particularly dense, it was not possible to count all plants. Thus, plants were counted within regularly spaced quadrats with a sampling fraction of $d_t = 1/11$. This corresponded to five populations for $t = 0$, eight for $t = 1$, and ten for $t = 2$. It was not possible to count seeds in the seed banks and seeds resulting from immigration. Seeds produced by plants were not counted so as to limit disturbances of reproduction and dispersal. The length $L$ of each population patch was recorded (patches were approximately one meter wide). Descriptive statistics of the data are given in Web Appendix F.

Model. The model in Section 2 was used with a few changes. Immigration was divided into an immigration due to truck losses and an immigration due to the harvest of an adjacent crop. The rates of immigration were assumed to be proportional to the length of the population patch. Immigration from truck losses and fields was modeled using a Poisson-gamma distribution with a mean of $uL$ and $vL$, respectively, and a variance of $(1 + 1/\theta)uL$ and $(1 + 1/\theta)vL$, respectively. The terms of the model corresponding to the emergence of rosettes from seed banks were neglected for $t = 1$ because the rosettes present at $t = 0$ were not counted and the estimate of $a_0e_1/e_0$ was small. The survival probability of a plant was divided into the survival probability of a rosette in winter ($e_u$) and the survival probability of a vernalized rosette in spring ($e_v$), and these probabilities were made year-dependent. The identifiable quantities were $e_um$, $e_u$, $e_v$, $a_0e_1/e_0$, $e_0/\theta$, $e_v$, $t = 1, 2$, $e_u$, $t = 0, 1, 2$. The data were analyzed using the EE method together with a nonparametric bootstrap, and the IS method. The details of the implementation of these methods are given in Web Appendix E.

Estimates. The EE and IS analyses give different estimates but similar conclusions (Table 1). They suggest that plant fecundity is low. The rate of immigration from trucks appears significant and substantial for long populations. The rate of immigration from crops is more difficult to estimate, presumably because few populations were adjacent to an oilseed rape crop (30 populations for $t = 1, 15$ for $t = 2$). The quantity $a_0e_1/e_0$ seems to be low. Although the EE and IS estimates of $e_0/\theta$ are quite different, they are both large. The bootstrap estimates of standard errors suggest that the model-based estimates may be too low. The bootstrap estimates of the expectations of estimators are consistent with the EE estimates. It is difficult to know which of the EE or the IS estimates are the best ones since criteria calculated to assess these estimates gave conflicting results. For example, the IS estimates had a larger (approximate) log-likelihood but a larger sum of the absolute values of residuals. These estimates should be interpreted with caution because the model

<table>
<thead>
<tr>
<th>Parameter</th>
<th>EE</th>
<th>Bootstrap</th>
<th>IS</th>
</tr>
</thead>
<tbody>
<tr>
<td>$e_0m$</td>
<td>0.34</td>
<td>0.30</td>
<td>0.51</td>
</tr>
<tr>
<td>$e_u$</td>
<td>0.97</td>
<td>0.11</td>
<td>0.93</td>
</tr>
<tr>
<td>$e_v$</td>
<td>3.71</td>
<td>0.72</td>
<td>3.61</td>
</tr>
<tr>
<td>$a_0e_1/e_0 \times 10^4$</td>
<td>-</td>
<td>-</td>
<td>3.56</td>
</tr>
<tr>
<td>$e_0/\theta$</td>
<td>131.80</td>
<td>9.71</td>
<td>124.20</td>
</tr>
</tbody>
</table>
is based on many assumptions and the statistical procedures
involve approximations (see Section 6). A more detailed pre-
sentation of these data and their analysis (including estimates
of plant survival probabilities) can be found in the compan-
ion paper of Garnier, David, Lecomte, Deville, and Larédé (in
prep.).

6. Discussion

Model. Our model involves various assumptions. Density
dependence is assumed to be negligible, i.e., model parameters
do not depend on population size. In our application, this
assumption was justified by the fact that plant density was
low in most populations. However, rosette density was high in
a few populations, implying that rosette emergence or survival
might have been affected by density in these populations. The
model could be refined to account for density dependence.

In our model, emigration is neglected and the rate of im-
migration does not depend on the size of neighboring popula-
tions. In our application, this assumption was justified by the
low estimate of fecundity. A spatial version of the model could
developed in which populations form a metapopulation
and exchange seeds. Genetic data could be useful for such
models (Burczyk et al., 2006).

Our model takes account of demographic stochasticity (the
probabilistic nature of events at the level of individuals), dem-
ographic heterogeneity in fecundity (the variation in vital
rates among individuals within a population), and environ-
mental stochasticity in immigration (the random variation
of parameters in space and time) (Melbourne and Hastings,
2008). Additional random effects could be introduced into
the model. For example, random spatial or population effects
could be used to model the survival probabilities of plants in
our application.

It would be interesting to relax the assumptions that the
offspring and the immigration distributions have a common
parameter, and that measurement errors are not frequent with
a binomial distribution. In our application, measurement er-
ors appeared more dispersed than binomial errors.

More work is needed to develop the extensions mentioned
above. When such statistical analyses become possible, it will
be more expedient to fit several models to the data and to
use methods of model selection/averaging (Chapter 7, Robert,
2001).

It is recommended to check the consistency between the
model and the data using methods of model validation
(Chapter 6, Gelman et al., 2004). For example, in our appli-
cation, some initial results suggested that only three rosette
numbers (out of 384) were not in agreement with the model
because they were far from their 95% credible intervals gen-
erated from the model. Moreover, some posterior predictive
p-values calculated for rosette numbers for t = 1, 2 were close
to 50% (Chapter 6, Gelman et al., 2004).

The EE and bootstrap methods are supposed to be less de-
pendent on model assumptions than the methods based on the
likelihood. It would be interesting to assess the robustness of
these methods quantitatively using simulations. Moreover, it
would also be interesting to develop statistical methods based
on semi-parametric models so as to make fewer assumptions
(Wood, 2001).

Statistical methods. The methods presented in this article
are based on some approximations/simplifications. It is rec-
ommended to make sure that the assumption that the quantity
\( \alpha_m F_{t-2} + \alpha_n \) is small, which is involved in the EE method,
is satisfied. In our application, the latter quantity (extended to
take account of both types of immigration and patch length)
was often low: its median, mean, and max. were equal to 0.07,
0.41, and 5.45, respectively. If this assumption is not satisfied,
the EE system could be modified (using Web Appendices B
or C), or the IS method could be implemented since it is less
dependent on assumptions related to parameter values. The
developed methods are easy to implement but they lead to
biased estimators (however, our simulations show that this
bias can be moderate) and they do not provide information
on some seed parameters. Thus, it would be interesting to fit
our model using other methods, e.g., sequential Monte Carlo
methods (Buckland et al., 2007).

Because of missing data, the methods considered here pro-
vide estimates of only some parameter products. Likewise,
some parameters were confounded in the animal censuses of
Besbeas, Freeman, and Morgan (2005) and Véran and
Lebreton (2008). To solve this problem, these authors im-
plemented an integrated population modeling approach by
simultaneously analyzing the census data and some capture-
recapture data. It would be interesting to apply such meth-
ods to plant surveys. The present article indicates which pa-
rameters are confounded and can help to specify which addi-
tional data could be collected. For example, the parameters
\( e_o, m, a, \) and \( \theta \) are identifiable if additional data on \( e_0 \) or \( m \)
are available.

Applications. Our work was motivated by a survey of feral
oilseed rape populations with missing data. The developed
methods could be applied/adapted to other plant populations
that may have a soil seed bank and that are subject to a
potential immigration of seeds, for example, feral sunflower
(annual), feral rye (annual), or feral alfalfa (perennial). They
could also have applications in animal surveys with miss-
ing data. For example, the animal census data analyzed by
Besbeas et al. (2005) and Véran and Lebreton (2008) in the
context of conservation biology are similar to the oilseed rape
data; they include the number of breeding females but the
number of juveniles is missing.

7. Supplementary Materials

Web Appendices A, B, C, D, E, and F, referenced in Sec-
tions 2, 3, 4, 5 and 6, are available under the Paper Informa-
tion link at the Biometrics website http://www.biometrics.
tibs.org.

Acknowledgements

We would like to thank Alexandra Deville, who took part in
the collection of the data used in Section 5. This study was
funded by the European project, “Sustainable Introduction of
GMOs into European Agriculture”, and by the Agence Na-
tionale de la Recherche (project: “Flux des (trans-)gènes et
impact sur la biodiversité”). We thank two referees and an
associate editor for their helpful suggestions. We thank Gail
Wagman for helping us to write the paper.
REFERENCES


Received January 2009. Revised September 2009. Accepted September 2009.
Queries

Q1  Author: Please provide the publisher’s location in Durbin and Koopman (2001).
Q2  Author: Please provide the publisher’s location in Gelman et al. (2004).
Q3  Author: Please provide the publisher’s location in Haccou et al. (2005).
Q4  Author: Please provide the publisher’s location in McCullagh and Nelder (1989).
Q5  Author: Please provide the publisher’s location in Robert (2001).
Q6  Author: Please provide the publisher’s location in Seber and Wild (1989).
Q7  Author: Please provide the publisher’s location in Van der Vaart (1998).