

Survival and Aging in the Wild via Residual Demography

Hans-Georg Müller¹, Jane-Ling Wang¹, Wei Yu², Aurore Delaigle³ and James R. Carey^{4,5}

¹Department of Statistics, University of California, Davis, One Shields Ave., Davis, CA 95616, USA

²Department of Mathematics, University of California, Davis, One Shields Ave., Davis, CA 95616, USA

³Department of Mathematics, University of California, San Diego, La Jolla, CA 92093, USA

⁴Center for the Economics and Demography of Aging, University of California, Berkeley, CA 94720, USA

⁵Department of Entomology, University of California, Davis, One Shields Ave., Davis, CA 95616, USA

3 July 2007

Correspondence:

Hans-Georg Müller, Department of Statistics, University of California, Davis, One Shields Ave., Davis, CA 95616, USA. Tel: +1 530 752 6217; fax: +1 530 752 1537; e-mail: mueller@wald.ucdavis.edu

Summary

Information about the age distribution and survival of wild populations is of much interest in ecology and biodemography, but is hard to obtain. Established schemes such as capture-recapture often are not feasible. In the proposed residual demography paradigm, individuals are randomly sampled from the wild population at unknown ages and the resulting captive cohort is reared out in the laboratory until death. Under some basic assumptions one obtains a demographic convolution equation that involves the unknown age distribution of the wild population, the observed survival function of the captive cohort, and the observed survival function of a reference cohort that is independently raised in the laboratory from birth. We adopt a statistical penalized least squares method for the deconvolution of this equation, aiming at extracting the age distribution of the wild population under suitable constraints. Under stationarity of the population, the age density is proportional to the survival function of the wild population and can thus be inferred. Several extensions are discussed. Residual demography is demonstrated for data on fruit flies *Bactrocera oleae*.

Keywords: Bactrocera oleae; Captive cohort; Convolution equation; Deconvolution; Inverse problem; Nonparametric estimation; Penalized least squares

1. Introduction

Understanding aging in the wild is one of the most important problems in biodemography, yet available methodology is limited. With appropriate demographic tools, empirical data derived from field studies are expected to be used to frame and test theories of aging, inform research concerned with aging mechanisms, and establish baselines for the natural history of aging (Finch 1990, 2001; Promislow 1991; Gaillard *et al.* 1994; Reznick *et al.* 2001; Tatar and Yin 2001). We introduce the new concept of residual demography with the aim to deduce the age distribution of individuals in the wild. Residual demography utilizes data on laboratory survival from both wild-caught individuals of unknown age and wild-type individuals of known age. We demonstrate that the age structure of the wild population can be deduced from the survival of these individuals in the laboratory, in the general situation where survival in the wild and in the laboratory might differ. This study is motivated by the need to develop a method to estimate age structure in field populations of insects.

Residual demography includes two experimental steps, namely (1) the capture and transfer of wild individuals from the field to the laboratory where they are reared out under defined conditions and their date of death is recorded (captive cohort); (2) Creation of a reference cohort by obtaining individuals at birth from the wild and rearing them under the same defined laboratory conditions and recording death rates. The analysis of these data is then based on statistical deconvolution methodology. The aim is to construct an age distribution of the wild population that, when combined with the survival schedule of the reference cohort, yields the distribution of deaths observed in the captive cohort. If the population is stationary and a stable equilibrium of the age distribution has been reached, the density of the age distribution in the wild corresponds to the survival function of the population, a fact that is known from renewal theory (Feller, 1950).

A demographic identity to infer aging and survival in the wild has been discussed previously under the assumption that survival in the wild and after marking individuals is subject to the same force of mortality (Müller *et al.* 2004). This major restriction is dropped in the proposed residual demography approach, where the force of mortality is allowed to change upon transferring an individual of unknown age from the wild to the laboratory. This reflects underlying biology more closely and makes it possible to obtain inference for the age distribution in the wild in the face of differing mortality rates between field and laboratory. We also briefly discuss extensions of this new approach to the case of non-stationary birth rates and unequal sampling probabilities

of capture, and to the closely related problems of estimating age-at-capture of individuals, and of estimating the force of mortality (hazard function) in the wild.

We show that the assumption that the force of mortality acting on an individual depends solely on the individual's age and on whether the individual is in the wild or in the laboratory implies a demographic convolution equation. Consequently, the recovery of the unknown age distribution requires to solve a statistical deconvolution problem (see Madden *et al.* 1996 for an overview). An implementation by penalized least squares is shown to provide a feasible solution to this inverse problem. Crucial for this specific deconvolution problem are functional constraints such as smoothness, non-negativity and monotonicity which can be conveniently incorporated via suitable penalty terms.

Assumptions made for the application of the proposed approach in biodemographic studies are that the survival of sampled individuals in captivity is determined by the corresponding age-specific mortality in captivity, irrespective of when the capture occurs, and that each individual present in the wild population has the same chance of being sampled, irrespective of age. If additional information is available, some of these assumptions can be relaxed. For inference about survival in the wild, going beyond inferring merely the age distribution in the wild, another needed assumption is stationarity of the population. This assumption also may be relaxed if additional information is available. We demonstrate the proposed deconvolution methodology with cohorts sampled from the fruit flies *Bactrocera oleae*.

In section 2, we derive the novel demographic convolution equation that underlies the concept of residual demography and discuss various extensions. In section 3, we discuss the deconvolution and its actual implementation. A biodemographic data set is used in section 4 to demonstrate the methodology in action. Further discussion follows in section 5.

2. A demographic convolution equation

In residual demography, subjects from a wild population are randomly caught and placed in a cohort of wild-caught flies of unknown age, the *captive cohort*. It is assumed that the unknown age A of a captured individual is distributed according to the age distribution of the population, i.e., all individuals in the wild are equally likely to be sampled. The captive cohort is reared under well-defined laboratory conditions until all subjects in this cohort are dead. In addition to the captive cohort, a *reference cohort* is assembled which consists of newborn subjects of the same type as the wild population. The reference cohort is raised under identical conditions as the

captive cohort until age-at-death is recorded for all subjects. If this cohort is sufficiently large, reasonably accurate estimates of the survival schedule of subjects under laboratory conditions can be constructed. For an individual that is captured at unknown age A , then enters the captive cohort and dies after an observed residual life time in captivity C , age-at-death T is

$$T = A + C. \quad (1)$$

Since A is unknown, so is T and therefore common methods of survival or life table analysis do not apply.

The following notation will be used throughout where the existence of the underlying quantities is assumed: For a non-negative random variable X , let $F_X(t) = P(X \leq t)$ denote the distribution function, $\bar{F}_X(t) = 1 - F_X(t)$ the survival function, $f_X(t) = \frac{d}{dt}F_X(t)$ the probability density function, $\lambda_X(t) = f_X(t)/\bar{F}_X(t)$ the hazard rate, and $\Lambda_X(t) = \int_0^t \lambda_X(s) ds$ the cumulative hazard rate. Here the choice $X = W$ will label these quantities for subjects in the wild, $X = C$ for subjects in the captive cohort, and $X = R$ for subjects in the reference cohort. Furthermore, the density of the distribution of age-at-capture is denoted by f_A .

In the remainder of this section, we derive the convolution equation which is basic for residual demography, list the needed assumptions in section 2.2 and then proceed to discuss several specific features and extensions in sections 2.3 to 2.7.

2.1 The basic equation

The hazard rate of a subject at age t is then

$$\lambda(t) = \begin{cases} \lambda_W(t) & \text{if } t \leq A \\ \lambda_R(t) & \text{if } t > A. \end{cases} \quad (2)$$

Observing $\bar{F}_C(t) = E\{P(C > t|A)\}$, we obtain from $\bar{F}_{C|A}(t|A) = e^{-(\Lambda_R(t+A)-\Lambda_R(A))}$ that

$$\bar{F}_C(t) = \int_0^\infty e^{-(\Lambda_R(t+a)-\Lambda_R(a))} f_A(a) da = \int_0^\infty \bar{F}_R(t+a) \frac{f_A(a)}{\bar{F}_R(a)} da. \quad (3)$$

From the reference cohort, one easily obtains an estimate of \bar{F}_R and from the captive cohort an estimate of \bar{F}_C , so that the only unknown function in this *convolution equation* is f_A .

By differentiating both sides of (3) w.r. to t , one obtains

$$f_C(y) = \int_0^\infty f_R(x+y) \frac{f_A(x)}{\bar{F}_R(x)} dx. \quad (4)$$

Upon integrating both sides of (3), R denoting the lifetime random variable in the reference cohort,

$$EC = \int_0^\infty \bar{F}_C(t) dt = \int_0^\infty E(R - a | R > a) f_A(a) da, \quad (5)$$

so that residual life expectancy in the captive cohort is seen to be a mixture of expected residual lifetimes in the reference cohort.

If we make the additional assumption of a stationary and stable wild population, the relationship between f_A and survival function \bar{F}_W of the wild population is given by

$$\bar{F}_W(t) = \frac{f_A(t)}{f_A(0)}. \quad (6)$$

Deconvolving equation (4) then leads to estimates of the wild survival schedule \bar{F}_W . We note that by integrating (6) on both sides, we find the relationship $EW = 1/f_A(0)$, i.e., the age density at age 0 is equal to the reciprocal life expectancy in the wild.

The convolution equations (3), (4) can be extended to more general situations. These include the case of sampling with unequal probabilities, for example if older individuals of the wild population are more likely to be sampled than younger individuals, and the case where the population is not stationary and has time-varying birth rates. This is of interest when seasonal variations in birth rates in the wild would be expected. Further details on these extensions are given in the following section 2.3, where we also discuss the situation of a nonstationary survival schedule. The special cases $F_R = F_W$, predicted age at capture and the estimation of the force of mortality for the wild population are discussed in sections 2.4, 2.5 and 2.6, respectively, while issues of identifiability are the topic of section 2.7.

2.2 Assumptions

A basic assumption underlying all elements of our approach is that the force of mortality depends only on age of an individual and the current environment, and does not depend on past environmental exposure. This assumption implies that an individual in captivity has the same mortality at a given age whether it has spent its prior life in captivity or in the wild.

In the biodemographic context, this assumption is supported by findings of Carey *et al.* (1998) and Sgo and Partridge (1999), who report the fast adaptation of both Mediterranean fruit flies and *Drosophila* to new survival schedules corresponding to changed dietary environments, irrespective of past exposures.

A second assumption is that age-bias of captures in the wild is minimal. If age-bias cannot be neglected, a correction factor can be introduced as described in section or, alternatively, the wild age density f_A can be taken to refer to the age structure of wild-captured flies rather than wild flies. This alternative interpretation continues to lead to useful information, for example regarding the presence of old flies in the wild.

A third assumption for the basic convolution equation in section 2.2 is that the population is stable, and birth rates are stationary throughout the observation period. This assumption may be violated in wild populations and it can be discarded under a slightly modified scheme that is discussed in section 2.3.

These three assumptions suffice to derive the convolution equation, which determines the age distribution in the population. In order to solve the equation, we assume that the age density is smooth, monotone decreasing and that the wild survival schedule is such that the solution of the equation is identifiable. None of these assumptions imposes a serious restriction, for identifiability this is discussed in section 2.7.

In order to use the solution of the basic convolution equation to derive the wild survival schedule, we need to make the additional assumption that population survival is stationary. This means that population (cohort) hazard rates depend only on age of an individual but do not depend on calendar time. As survival may depend on available resources or other changing environmental conditions, hazard rates in some cases may vary with calendar time. Extensions of the basic model to cover this case are discussed in section 2.3.

2.3 Extensions of the convolution equation

We first consider the extension to the case with unequal sampling probabilities. In this situation availability of a sampling function $\beta(a) = P(\text{Individual is sampled}|\text{Individual is of age } a)$ is useful; this function may be known or may be determined from additional experiments. We then define a selected age density function $\check{f}_A(a) = \beta(a)f_A(a) / \int \beta(a)f_A(a) da$ and replace (3) with a generalized version,

$$\bar{F}_C(t) = \int_0^\infty \bar{F}_R(t+a) \frac{\check{f}_A(a)}{\bar{F}_R(a)} da.$$

A second extension addresses the non-stationary case where the survival schedule of the wild population is stationary, but the age distribution of a population is changing over time, due to a birth rate that is itself changing over time, so that the wild age distribution is not stationary. A

modified sampling plan, tailored to this situation, is to sample individuals from the wild not just at one but at each of various calendar times z , and then to rear several captive cohorts captured at these times z . Using the same arguments as before, the ensemble of the corresponding captive cohorts then gives rise to a family of convolution equations

$$\bar{F}_C(t, z) = \int_0^\infty \bar{F}_R(t + a) \frac{f_A(a, z)}{\bar{F}_R(a)} da, \quad (7)$$

where $f_A(a, z)$ is the probability density of the age distribution for the population at calendar time z , which can be recovered by deconvolution.

With birth rate at calendar time y quantified by a function $\gamma(y)$, the assumption of stationary survival probabilities (i.e., the probability for an individual to survive from one age to the next does not depend on calendar time) leads to $f_A(a, z) = \gamma(z - a) \bar{F}_W(a) / \int \gamma(z - t) \bar{F}_W(t) dt$. Setting $a = 0$, one finds $\int \gamma(z - t) \bar{F}_W(t) dt = \gamma(z) / f_A(0, z)$ and therefore

$$\bar{F}_W(t) = \frac{f_A(t, z)}{f_A(0, z)} \frac{\gamma(z)}{\gamma(z - t)}, \quad (8)$$

Hence the target survival schedule \bar{F}_W can be recovered in the same way by deconvolution as in the stationary case, given the birth rates $\gamma(z)$, which need to be known or determined from field studies. In the stationary case, $\gamma(z) = \gamma(z - a)$ and the original sampling schedule with a captive cohort sampled at just one calendar time z will suffice, as (8) is then equivalent to (6). For stationary survival in the wild it is advantageous to pool estimates obtained for the r.h.s of equation (8) for various values of z , as the l.h.s. does not depend on z . If estimates $\hat{f}_A(a, z)$ of age densities $f_A(a, z)$ are obtained from samples collected at $K \geq 1$ calendar times z_1, \dots, z_K , such a pooled estimate is given by

$$\hat{\bar{F}}_W(t) = \frac{1}{K} \sum_{j=1}^K \frac{\hat{f}_A(t, z_j)}{\hat{f}_A(0, z_j)} \frac{\gamma(z_j)}{\gamma(z_j - t)}. \quad (9)$$

A third extension of the basic convolution equation to the case of non-stationary survival in the wild is a consequence of the above considerations. The nonstationary case arises for example if hazard rates are season-dependent. In this case “cross-sectional” survival functions $\bar{F}_W(t, z) = P(W > t | z) = \exp(-\int_0^t \lambda_W(s | z - t + s) ds)$ at calendar time z are still within reach. Adapting (8), we obtain

$$\bar{F}_W(t, z) = \frac{f_A(t, z)}{f_A(0, z)} \frac{\gamma(z)}{\gamma(z - t)}. \quad (10)$$

This suggests to sample subjects from the wild on a sufficiently dense grid of calendar times z over a domain D and then to construct a survival surface $\bar{F}_W(t, z)$ for all relevant t and $z \in D$.

In some cases an additional local smoothing step across the values of z in local neighborhoods may improve efficiency of such surface estimates, analogously to (9). Pooling the deconvolved wild survival estimates according to (9) under the nonstationary model (10) has the effect to reduce variability, thus counteracting the smaller cohort sizes which are a consequence of the more frequent sampling. Such a pooling method will target $\bar{F}_{W \text{ ave}}(t) = \int_D \bar{F}_W(t, z) dz$, an “overall average” survival function which may serve as a useful summary measure. A similar pooling effect is achieved if various captive samples obtained at various calendar dates are pooled together into one larger sample before deconvolution.

2.4 The case of identical survival schedules

In some situations it is of interest to consider the case where survival in the wild (characterized by the survival function F_W) and survival in the laboratory reference cohort (characterized by the survival function F_R) are identical, i.e., where one has $F_R = F_W$. In this case,

$$\bar{F}_W(a) = f_A(a)/f_A(0)$$

and this implies that (4) simplifies to

$$f_C(y) = f_A(0) \int_0^\infty f_W(y+x) dx = f_A(0) \bar{F}_W(y),$$

so that

$$\bar{F}_W(y) = f_C(y)/f_C(0),$$

obtained by plugging in and observing that $\bar{F}_W(0) = 1$.

This corresponds to an identity from renewal theory (Feller 1950) that has been revisited in a demographic context in Müller *et al.* (2004). An example from anthropological research was given there. Since an explicit solution can be found in this special case, a more complex deconvolution method is not needed.

2.5 Predicted age at capture

It may be sometimes of interest to predict age-at-capture A for an individual for which a remaining lifetime after capture $C = c$ has been observed, but nothing else is known. In this case, the best prediction is the conditional expectation $E(A|C = c)$.

Straightforward calculations lead to

$$E(A|C = c) = \int a f_{A|C}(a, c) da = \frac{1}{f_C(c)} \int a \frac{f_R(a + c) f_A(a)}{\bar{F}_R(a)} da. \quad (11)$$

Estimates for age-at-capture can then be obtained by substituting smooth estimates such as kernel density estimators (e.g. Müller 1997) for f_C , f_R and \bar{F}_R , and smooth deconvolution estimates (18) for f_A , followed by numerical integration.

2.6 Force of mortality in the wild

Under stationarity assumptions, (6) implies for the density f_W of the lifetime distribution in the wild that

$$f_W(t) = -\frac{f'_A(t)}{f_A(0)}.$$

The force of mortality λ_W is thus found to be

$$\lambda_W(t) = \frac{f_W(t)}{\bar{F}_W(t)} = -\frac{f'_A(t)}{f_A(t)}. \quad (12)$$

Given estimates $\hat{f}_A(t)$ from (18), one possibility to obtain the needed derivative $f'_A(t)$ by forming difference quotients,

$$\hat{f}'_A(t) = \frac{1}{\Delta} [\hat{f}_A(t + \Delta/2) - \hat{f}_A(t - \Delta/2)]$$

and then plugging $\hat{f}'_A(t)$, $\hat{f}_A(t)$ into (12).

2.7 Identifiability issues

We provide here a discussion of the problem under which assumptions a (unique) solution of the convolution equation (4) exists. The mathematical arguments provided below imply that deconvolution will not work if wild survival is a mixture of exponential distributions. As these distributions are associated with constant (in case the mixture has just one component) or decreasing hazard rates as age increases, they are not likely to be encountered as survival distributions in wild aging applications.

More specifically, define the deconvolution problem as finding the solution of

$$\operatorname{argmin}_{f_A \in \mathcal{F}} \|f_C(y) - \int_0^\infty K(x, y) g(x) dx\|,$$

where we redefine the target as $g(x) = f_A(x)/\bar{F}_R(x)$ and view $K(x, y) = f_R(x + y)$ as the (symmetric) kernel of a linear operator Ω_K in Hilbert space, so that $f_C = \Omega_K(g)$. Here \mathcal{F} is the space of all smooth density functions with non-negative and bounded support, and $\|\cdot\|$ is a suitable norm, usually chosen as the L^2 norm for the case of square integrable functions. In general, Ω_K is a compact operator and as such is not globally invertible.

If the kernel is degenerate, i.e., there exist finitely many linearly independent functions α_j , $j = 1, \dots, p$, such that

$$K(x, y) = \sum_{j=1}^p \alpha_j(x)\alpha_j(y),$$

then the convolution equation at best determines the integrals $\int \alpha_j(x)g(x) dx$, $j = 1, \dots, p$, and therefore in general the function g is not determined. An extreme case is $p = 1$ where only $\int \alpha_1(x)g(x) dx$ is determined by the convolution equation for some function α_1 , which clearly shows that deconvolution is not identifiable in this case.

The property $K(x, y) = \alpha_1(x)\alpha_1(y)$ is characteristic for the exponential density for which $f_R(x + y) = \lambda \exp(-\lambda x) \exp(-\lambda y)$ for a parameter $\lambda > 0$. This density therefore provides an example where deconvolution is not feasible. The case where f_R is a mixture of exponential distributions also leads to a degenerate kernel with $p > 1$. If on the other hand the eigenfunctions of the operator Ω_K form a basis of the underlying function space, a solution can be found in the image space $\mathcal{I} = \Omega_K\mathcal{F}$ under certain assumptions (e.g., He *et al.* 2002).

3. Deconvolution for recovering age distribution and survival in the wild

The basic demography concept requires is to solve the convolution equation (4). Deconvolution generally is a difficult task and falls into the class of inverse problems (Nowak 1998; Carroll and Hall 2004). The proposed approach to solve equation (4) for f_A is based on the idea to approximate pertinent density functions, including f_A , by step functions. The step functions approximating densities f_R and f_C can then be estimated from the data. Plugging these estimates into convolution equation (4) leads to a linear system for the coefficients of the step function approximations whose solution provides an estimate for f_A . This linear system is generally ill-conditioned and must be regularized. For the necessary regularization we propose to use penalized least squares, as detailed in the following.

3.1 Deconvolution by regularization

Regularization is partially achieved by the step function approximation which can be coarsened to any desired degree in order to reduce the dimension of the system, and to a larger extent by introducing a penalized least squares algorithm. The solution must satisfy certain properties such as being a density and being smooth. When estimating a survival schedule via (6) it also must be monotone decreasing. Directly solving the linear system under hard constraints was found to be inferior to the proposed deconvolution via penalized least squares, with penalties ρ_1 for violating smoothness, ρ_2 for violating non-negativity, and ρ_3 for violating that the integral under the curve is 1 (as required for a density). When estimating the survival function in the wild, we also introduce a penalty ρ_4 for violating monotonicity. Details about the implementation are given in subsection 3.2, with the target criterion (18).

From Eq. (17) below, the linearized and discretized version of the convolution equation (4) can be written as

$$\sum_{j \geq 0} \left[\frac{f_{Rk+j}}{\bar{F}_{Rj}} \right] f_{Aj} = f_{Ck}, \quad (13)$$

where f_{Xj} denotes the fraction of a cohort that is at cohort age j , and \bar{F}_{Xj} denotes the probability to survive to age j , for subjects in the reference cohort ($X = R$), the captive cohort ($X = C$) or in the wild cohort ($X = A$), where A stands for chronological individual age. We thus see that the feasibility of deconvolution hinges on the properties of the design matrix B defined by (4) with elements $b_{jk} = \frac{f_{Rk+j}}{\bar{F}_{Rj}}$.

A simple discrete example to illustrate Eq. (13) is as follows. Assume there are only three age groups, such that aging in the wild corresponds to $f_{A0} = 1/3, f_{A1} = 1/3, f_{A2} = 1/3$, while aging in captivity is accelerated and one observes the values $f_{R1} = 1/2, f_{R2} = 1/3, f_{R3} = 1/6$, with the associated values $\bar{F}_{R0} = 1, \bar{F}_{R1} = 1/2, \bar{F}_{R2} = 1/6, \bar{F}_{R3} = 0$ for the reference cohort. Plugging into the above equation, a straightforward calculation shows that one will observe for the captive cohort $f_{C1} = 13/18, f_{C2} = 4/18, f_{C3} = 1/18$. In this simple toy example the deconvolution problem can be directly solved as the 3×3 matrix corresponding to (13) has determinant -1 and is thus directly invertible. In general, when the number of bins is large, this matrix is ill-conditioned. Direct inversion is then not possible and the estimate one would obtain by non-penalized least squares would be neither smooth nor a density; regularization is crucial.

3.2 Regularized deconvolution via penalized least squares

Assume the captive cohort consists of n_C subjects with observed residual lifetimes C_1, \dots, C_{n_C} and the reference cohort consists of n_R subjects with observed residual lifetimes R_1, \dots, R_{n_R} . To approximate the densities appearing in (4) by step functions, we define a suitable equidistant grid of M points x_1, \dots, x_M , such that $x_j = (j-1)\Delta$ for a constant $\Delta > 0$. By choosing a small value for Δ , this approximation can be made as precise as desired. Choosing an integer M such that $x_M = \max_{1 \leq i \leq n_R} \{R_i\} - \Delta$, densities f_A and f_C will be approximated on intervals $[0, x_M]$ and density f_R on interval $[0, x_{2M}]$. Defining right-open intervals $I_j = [x_j, x_{j+1})$, $j \geq 1$, the approximated densities are

$$\tilde{f}_A(x) = \sum_{j=1}^{M-1} f_{Aj} 1_{I_j}(x), \quad \tilde{f}_C(x) = \sum_{j=1}^{M-1} f_{Cj} 1_{I_j}(x), \quad \tilde{f}_R(x) = \sum_{j=1}^{2M-1} f_{Rj} 1_{I_j}(x), \quad (14)$$

where $1_S(x) = 1$ if $x \in S$, and $1_S(x) = 0$ otherwise, for any set S .

In a first step, we estimate densities f_R of the survival time distribution of the reference cohort and f_C of the residual lifetime of the captive cohort by estimating the respective coefficients in (14),

$$\hat{f}_{Rj} = \frac{1}{\Delta \sum_{i=1}^{n_R} 1_{\{R_i \in [0, x_{2M}]\}}} \sum_{i=1}^{n_R} 1_{\{R_i \in I_j\}},$$

$$\hat{f}_{Cj} = \frac{1}{\Delta \sum_{i=1}^{n_C} 1_{\{C_i \in [0, x_M]\}}} \sum_{i=1}^{n_C} 1_{\{C_i \in I_j\}},$$

which leads to the density estimates

$$\hat{f}_R(x) = \sum_{j=1}^{2M-1} \hat{f}_{Rj} 1_{I_j}(x), \quad \hat{f}_C(x) = \sum_{j=1}^{M-1} \hat{f}_{Cj} 1_{I_j}(x). \quad (15)$$

Defining $\hat{\hat{F}}_R(x) = 1 - \int_0^x \hat{f}_R(z) dz$, substituting these estimates then transforms convolution equation (4) into

$$\begin{aligned} \hat{f}_C(x_k) &= \int_0^\infty \hat{f}_R(x_k + x) \frac{\tilde{f}_A(x)}{\hat{\hat{F}}_R(x)} dx \\ &= \sum_{j=1}^{M-1} \tilde{f}_A(x_j) \int_{I_j} \hat{f}_R(x_k + x) / \hat{\hat{F}}_R(x) dx \\ &= \sum_{j=1}^{M-1} \tilde{f}_A(x_j) \hat{f}_R(x_k + x_j) \int_{I_j} \frac{1}{\hat{\hat{F}}_R(x)} dx, \end{aligned} \quad (16)$$

for $1 \leq k \leq M - 1$. This corresponds to a linear system

$$\hat{f}_C = B\tilde{f}_A, \quad (17)$$

for the unknown coefficients f_{Aj} of the target function \tilde{f}_A , where the $(M - 1) \times (M - 1)$ matrix B has elements

$$b_{kj} = \hat{f}_R(x_k + x_j) \int_{I_j} \frac{1}{\hat{F}_R(x)} dx, \quad 1 \leq k, j \leq M - 1,$$

and

$$\hat{f}_C = (\hat{f}_C(x_1), \dots, \hat{f}_C(x_{M-1}))', \quad \tilde{f}_A = (\tilde{f}_A(x_1), \dots, \tilde{f}_A(x_{M-1}))'.$$

We aim at solving this system for the coefficients f_{Aj} . Constructing the estimate \hat{f}_A as minimizer of a penalized least squares criterion leads to a solution that is a density and also a smooth function.

Denote by $\|\cdot\|$ the Euclidean norm, and by H a $(M - 1) \times (M - 1)$ matrix with diagonal elements -2 and both side diagonal elements 1, suitably modified near the end points of the diagonal, so that for a $(M - 1)$ -vector v we have $(Hv)_k = v_{k+1} - 2v_k + v_{k-1}$, $k = 2, \dots, M - 2$ and $\hat{f}'_A H' H \hat{f}_A$ corresponds to the vector of squared second order difference quotients of \hat{f}_A , apart from a normalization factor. Furthermore, denote by $\rho_1, \rho_2, \rho_3, \rho_4 \geq 0$ four nonnegative penalty parameters, to be chosen carefully. The target function is a density, therefore it is nonnegative and integrates to 1. It is also assumed to be a smooth and monotone falling function. Accordingly, we include penalty terms that penalize against violations of these four properties.

The penalized least squares criterion is then

$$\begin{aligned} \mathcal{P}(\tilde{f}_A) = & \|B\tilde{f}_A - \hat{f}_C\|^2 + \rho_1 \sum_{j=1}^{M-1} [\tilde{f}_A(x_j)]^2 1_{\{\tilde{f}_A(x_j) < 0\}} + \rho_2 [\Delta \sum_{j=1}^{M-1} \tilde{f}_A(x_j) - 1]^2 \\ & + \rho_3 \tilde{f}'_A H' H \tilde{f}_A + \rho_4 \sum_{j=1}^{M-1} (\tilde{f}_A(x_{j+1}) - \tilde{f}_A(x_j))^2 1_{\{\tilde{f}_A(x_{j+1}) - \tilde{f}_A(x_j) > 0\}}. \end{aligned} \quad (18)$$

Here the first penalty term penalizes against negative estimates and the second against the estimated function not integrating to 1. The third penalty term promotes smoothness of estimates, by penalizing against the sum of squared second order difference quotients, while the fourth penalty term penalizes against the solution not being monotone falling. Then the minimizer of $\mathcal{P}(\tilde{f}_A)$ with regard to \tilde{f}_A is the desired estimate \hat{f}_A of the density of the age distribution. We note that the least squares approach can be easily extended to weighted least squares by adding weights for each histogram bin of \tilde{f}_C , using for example the Poisson approximation to the histogram bin counts

(which are binomially distributed), or alternatively to a penalized maximum likelihood. Once \hat{f}_A has been obtained, the estimate of the wild survival function $\hat{\hat{F}}_W$ is an immediate consequence according to (6).

The choice of the grid x_k is often tied to the implicit scaling of the lifetables of captive and reference cohorts. In studies of flies or nematodes, it is customary to assess age-at-death in days, so that the natural choice is $\Delta = 1$ day (see Müller *et al.* 1997 and Wang *et al.* 1998, regarding issues of discretization in biodemographic analysis). In human studies, the natural unit might rather be a year. The penalty parameters can be chosen by simulations that mimic the observed data. Optimization routine `fminimax` (Matlab) was used in the implementation of the penalized least squares solutions \hat{f}_A .

4. Residual demography for fruit flies

To assess wild aging and survival schedule of the tephritid fruit fly (*Bactrocera oleae*) in the wild, flies were sampled from a wild population in olive groves near Davis/California. A captive cohort consisting of 457 olive flies and a reference cohort of 82 flies were assembled and reared under controlled conditions until all flies were dead and the proposed residual demographic method was applied to these data. The fitted survival schedule was then used as basis for a simulation study to determine the variability of the deconvolution algorithm.

4.1 Wild survival for *Bactrocera oleae*

Flies were collected from infested olives in Davis, California from June through August, 2004. Newly-eclosed individuals (82 flies) were collected for the reference cohort and placed individually in 1-oz clear plastic containers (condiment cups), provided with adult food (mixture of 3 parts sugar and 1 part yeast hydrolysate) and water, maintained at 25C (± 3), 65% (± 10) relative humidity; 12:12 light:dark cycle, and monitored each day to record age of death. Captive cohort information was collected on 457 adult olive flies captured in McPhail traps during this same period and in the same orchard. These flies were pooled into one large captive cohort (see the discussion in section 2.3). Individuals were removed from the traps in the laboratory using an aspirator, placed in individual 1-oz cages with food and water, maintained under the same conditions as in the baseline life tables studies, and monitored daily for survival.

The proposed deconvolution algorithm was applied and yielded reasonable results for the estimate of the density of the age distribution in the wild, as shown in Fig. 1 (with penalty parameters defined in (18) chosen as $(\rho_1, \rho_2, \rho_3, \rho_4) = (30, 20, 400, 0.5)$), along with 95% bootstrap confidence bands. The construction of these bands is described in the following subsection.

Simulated age densities, assuming that the age density shown in Fig. 1 corresponds to the true underlying distribution, are shown in Fig. 2. Further details on these simulations can be found in the following section. Fig. 3 displays the estimate of the wild survival function, obtained via (6), and the corresponding 95% bootstrap confidence bands. Some characteristics of this estimate of the wild survival function are first quartile Q1, at 15d (d=days), second quartile Q2 (median), at 31d, third quartile Q3, at 53d, and mean age at death, at 35d.

4.2 Confidence bands and simulations

In Fig. 1, pointwise 95% bootstrap confidence bands are shown along with the estimate of the density f_A . These confidence bands were obtained by creating bootstrap samples of captive cohorts of the same size as the actual observed captive cohort, by sampling 457 ages-at-death with replacement from the captive cohort data. Then for each bootstrap sample the deconvolution procedure was carried out, using the observed lifetime data for the reference cohort. In this way, 1000 bootstrap samples were created. The empirical 2.5% and 97.5% quantiles of the resulting sample of bootstrap density estimates were calculated at each age t and are shown as lower and upper confidence bands in the figure. Bootstrap confidence bands can be analogously constructed for the estimated wild survival function as shown in Fig. 3.

To simulate the behavior of the proposed deconvolution method, we proceeded as follows: For each simulation, we fixed the reference cohort data at the actually observed values, taking the estimate of the reference age-at-death distribution as the true distribution. We assume that the true underlying age density is the estimate f_A , as shown in Fig. 1. For each simulation run, we construct a cohort of 457 captive flies by generating their lifetimes. This step is described next.

We first obtain simulated ages-at-capture A by sampling from the assumed density f_A , using the graph of the estimated density in Fig. 1 and rejection sampling. Then we generate a simulated residual lifetime after capture C for each age-at-capture as follows: Independently from simulated age-at-capture A , we generate a random survival time R , distributed according to f_R , again by rejection sampling. If $R < A$, another pair (A, R) is generated. If $R \geq A$, we record $C = R - A$.

Then C is distributed as the residual lifetime after capture, with density f_C . For each simulation run, we generate in this way a captive cohort consisting of 457 residual lifetimes. These residual lifetimes are then entered into the deconvolution algorithm, along with the estimated reference density and distribution function. The result is a simulated estimate of the wild age density f_A .

These simulated captive lifetimes are then entered into the deconvolution algorithm. The resulting wild age estimate is plotted as one single estimated wild age density (grey line) in Fig. 2. The described procedure is carried out 100 times, yielding 100 wild estimated age densities, all of which are shown in Fig. 2. As the deviation between the simulated densities and the target density is relatively small, given the complexity of the deconvolution task, we conclude from this simulation that the method performs relatively well in a real-life setting.

5. Discussion and concluding remarks

The convolution equation (4) forms the basis to determine wild age and survival schedules from data obtained from both a captive and a reference cohort. A methodological difficulty is in devising a workable deconvolution scheme. We found that penalized least squares provides a viable method for biodemographic deconvolution. If one aims at recovering the wild survival schedule, the age distribution in the wild is obtained first and the assumption of a stationary population is needed to determine the survival schedule from the age distribution. The stationarity assumption can be relaxed by invoking a slightly more complicated sampling design, combined with knowledge about relative number of births, to be obtained from additional field studies.

We note that this methodology is of interest beyond demography: It provides a novel instance of a deconvolution problem under constraints, motivating further statistical research. In survival analysis, information about the onset of a condition may be missing for a group of subjects who enter a study (this would correspond to the unknown age of a subject in the biodemographic framework), while the distribution of the onset times is of great interest; an example is infection with HIV (Bacchetti and Jewell 1991). If the cohort sampled from the “wild” has an unknown onset time (captive cohort) and a second cohort of subjects is observed for which onset occurs during the time subjects are being studied (reference sample), survival information from the subjects in both groups can be combined to obtain inference about the timing of the onset. This can be done by setting up an analogue of equation (3) and then solving the deconvolution problem. Similarly, in reliability applications the convolution model will be useful in situations where one

wishes to infer the longevity of manufactured items under realistic everyday use conditions. The reference cohort in this case corresponds to a sample of items tested under laboratory conditions, while the captive cohort consists of items that are randomly sampled from the population of items in use and subsequently tested under laboratory conditions.

Studies concerned with aging in the wild have traditionally focused on either survival estimates using mark-recapture techniques (Caughley 1977; Austad 1993; Krebs 1999) or relatively rough life table differences between cohorts that were subjected to different selection pressures in the field (Reznick *et al.* 2004; Tatar *et al.* 1997). The applicability of these tools for demographic analysis is limited, as age-distribution information is not provided, in contrast to residual demography.

On one hand, residual demography complements mark-recapture methods which require the capture, marking, and re-capture of large numbers of individuals of known age (Buckland, 1982, Lebreton *et al.* 1992, Pradel 1996, Williams 2002, Moorhouse and MacDonald 2005), by relaxing the requirement of known age at capture. On the other hand, it provides an alternative methodology that requires capture only once and is amenable to parameter-free models. This is not the case for mark-recapture, where the sparseness of available information usually necessitates parametric model specifications (with associated maximum likelihood or Bayesian statistical methodology). However, one persistent finding in biodemographic studies has been the enormous plasticity of hazard rates found under various experimental conditions (Carey *et al.* 2002); this puts a premium on flexible nonparametric approaches. Whereas capture-mark-recapture methods can be used for large insects such as butterflies (Boggs *et al.* 2004), these methods are of limited usefulness for very invertebrates because of both low recapture rates and the likelihood of injury. Thus residual demography might be the only available methodology to study age structure and longevity in wild populations of certain organisms.

Further analytical and algorithmic developments will be needed to more directly address the various non-stationarities that likely exist in real populations, as discussed in sections 2.2 and 2.3. It will also be of interest to combine the proposed method in a suitable way with mark-recapture methodology, enhancing both residual demography and mark-recapture methods in the process. Further discussion of such extensions can be found in Carey *et al.* (2007).

While the deconvolution step requires careful implementation, as it involves an ill-posed inverse problem, we have demonstrated the feasibility to obtain and use information contained in captured individuals of unknown age. Residual demography provides a concept to assess age distribution in the wild, which is of interest in its own right, and under additional assumptions

allows to draw inference about wild survival. Setting up the prerequisite captive and reference cohorts is particularly feasible for species such as flies or nematodes and other easy-to-raise and easy-to-sample short-lived species. The combination of both laboratory and field studies, extracting information from both the captive and the reference cohort, is poised to shed further light on aging and survival in the wild.

Acknowledgments

This research was supported by NIH grant P01-AG08761 and NSF grants DMS03-054448, DMS04-04630 and DMS05-05537. We are obliged to Ken Wachter and Steven Orzack for extremely helpful feedback on an earlier draft of this paper, and to James Vaupel, Linda Partridge, Lawrence Harshman and Anatoli Yashin for encouragement regarding the concept of residual demography.

References

- Abrams PA (1993) Does increased mortality favor the evolution of more rapid senescence? *Evolution* 47, 877-887.
- Austad SN (1993) Retarded senescence in an insular population of Virginia opossums (*Didelphis virginiana*). *Journal of Zoology* 229, 695-708.
- Bacchetti P, Jewell NP (1991) Nonparametric estimation of the incubation period of AIDS based on a prevalent cohort with unknown infection times. *Biometrics* 47, 947-960.
- Boggs CL, Watt WB, Ehrlich PR (eds) (2004) *Butterflies: Ecology and Evolution Taking Flight*. Chicago, University of Chicago Press.
- Buckland ST (1982) A mark-recapture survival analysis. *Journal of Animal Ecology* 51, 833-847.
- Carey JR, Liedo P, Müller HG, Wang JL, Vaupel JW (1998). Dual modes of aging in Mediterranean fruit fly females. *Science* 281, 996-998.
- Carey JR, Liedo P, Harshman L, Zhang Y, Müller HG, Partridge L, Wang JL (2002) Life history response of Mediterranean fruit flies to dietary restriction. *Aging Cell* 1, 140-148.

- Carey JR, Papadopoulos N, Müller HG, Katsoyannos B, Kouloussis N, Wang JL, Wachter K, Yu W, Liedo P (2007). Population aging and extraordinary life span in wild medflies. *Aging Cell* (submitted).
- Carroll RJ, Hall P (2004) Low-order approximations in deconvolution and regression with errors in variables. *Journal of the Royal Statistical Society B66*, 31-46.
- Caughley G (1977) *Analysis of Vertebrate Populations*. Chichester, Wiley.
- Feller W (1950). *An Introduction to Probability Theory and its Applications*. John Wiley, New York.
- Finch CE (1990) *Longevity, Senescence, and the Genome*. Chicago: The University of Chicago Press.
- Finch CE (2001) History and prospects: symposium on organisms with slow aging. *Experimental Gerontology* 36, 593-597.
- Gaillard JM, Allaine D, Pontier D, Yoccoz NG, Promislow DEL (1994) Senescence in natural populations of mammals: A reanalysis. *Evolution* 48, 509-516.
- Good PI (2004) *Permutation, Parametric, and Bootstrap Tests of Hypotheses*. Springer New York
- Hawkes K, O'Connell JF, Jones NGB, Alvarez H, Charnov EL (1998) Grandmothering, menopause, and the evolution of human life histories. *Proceedings of the National Academy of Sciences of the United States of America* 95, 1336-1339.
- He G, Müller HG, Wang JL (2003) Functional canonical analysis for square integrable stochastic processes. *J. Multivariate Analysis* 85, 54-77.
- Jones NGB, Hawkes K, O'Connell JF (2002) Antiquity of postreproductive life: Are there modern impacts on hunter-gatherer postreproductive life spans? *American Journal of Human Biology* 14, 184-205.
- Krebs CJ (1999) *Ecological Methodology*. Second Ed., Menlo Park, Benjamin Cummings
- Lebreton JD, Burnham KP, Clobert J, Anderson DR (1992) Modelling survival and testing biological hypotheses using marked animals: A unified approach with case studies. *Ecological Monographs* 62, 67-118.

- Madden FN, Godfrey KR, Chappell MJ, Hovorka R, Bates RA (1996) Deconvolution techniques. *Journal of Pharmacokinetics and Biopharmaceutics* 24, 283-299.
- Moorhouse TP, MacDonald DW (2005) Indirect negative impacts of radio-collaring: sex ratio variation in water voles. *Journal of Applied Ecology* 42, 91-98.
- Müller HG (1997) Density estimation. In: *Encyclopedia of Statistical Science*, Ed. Kotz S, Read CB, Banks, DL, Wiley, New York, 185-200.
- Müller HG, Wang JL, Capra WB (1997) From lifetables to hazard rates: The transformation approach. *Biometrika* 84, 881-892.
- Müller HG, Wang JL, Carey, JR, Caswell-Chen EP, Chen C, Papadopoulos N, Yao F (2004) Demographic window to aging in the wild: Constructing life tables and estimating survival functions from marked individuals of unknown age. *Aging Cell* 3, 125-131.
- Nowak RD (1998) Penalized least squares estimation of Volterra filters and higher order statistics. *IEEE Trans. Signal Proc.* 46, 419-428.
- Pradel R (1996) Utilization of capture-mark-recapture for the study of recruitment and population growth rate. *Biometrics* 52, 703-709.
- Promislow DEL (1991) Senescence in natural populations of mammals: a comparative study. *Evolution* 45, 1869-1887.
- Reznick D, Buckwalter G, Groff J, Elder D (2001) The evolution of senescence in natural populations of guppies (*Poecilia reticulata*): a comparative approach. *Experimental Gerontology* 36, 791-812.
- Reznick DN, Bryant MJ, Roff D, Ghalambor CK, Ghalambor DE (2004) Effect of extrinsic mortality on evolution of senescence in guppies. *Nature* 431, 1095-1099.
- Sgro, M, Partridge, L (1999) A delayed wave of death from reproduction in *Drosophila*. *Science* 286, 2521-2524.
- Tatar M, Yin CM (2001) Slow aging during insect reproductive diapause: why butterflies, grasshoppers and flies are like worms. *Experimental Gerontology* 36, 723-738.

Tatar M, Grey DW, Carey JR (1997) Altitudinal variation for senescence in *Melanoplus* grasshoppers. *Oecologia* 111, 357-364.

Wang JL, Müller HG, Capra WB (1998) Analysis of oldest-old mortality: Lifetables revisited. *Annals of Statistics* 26, 126-163.

Williams BK, Nichols JD, Conroy MJ (2002) *Analysis and Management of Animal Populations*. Academic Press, London.

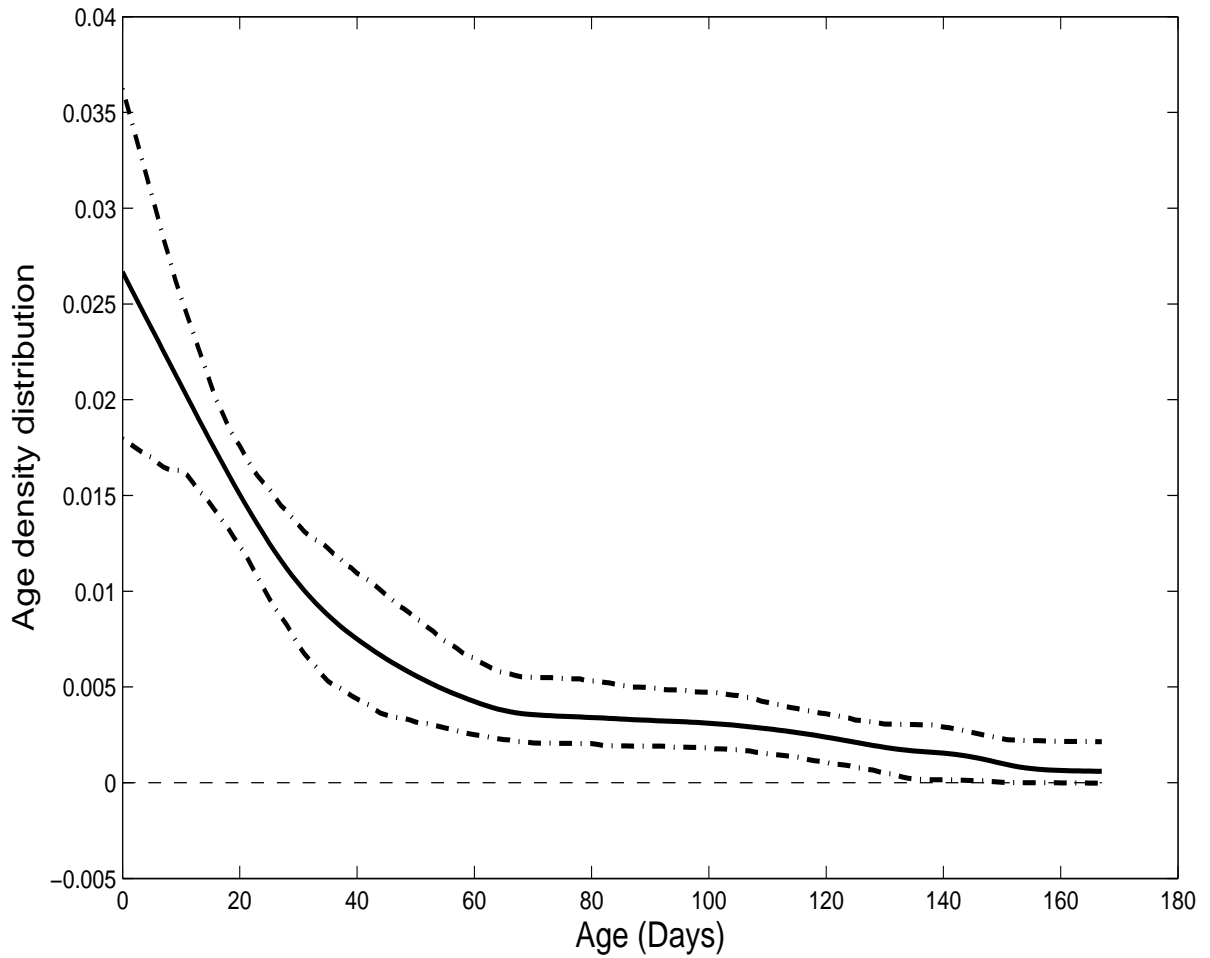


Figure 1: Upper panel: Monotone deconvolution estimate of the age density f_A in the wild for *Bactrocera oleae*, with 95% bootstrap confidence intervals. Middle panel: Estimated density f_A (thick solid) and 100 simulated deconvolution density estimates (grey lines), created as in Fig. 1; the dotted curve denotes the mean of the simulated age densities. Lower panel: Corresponding estimate of the survival function in the wild, with 95% bootstrap confidence intervals.

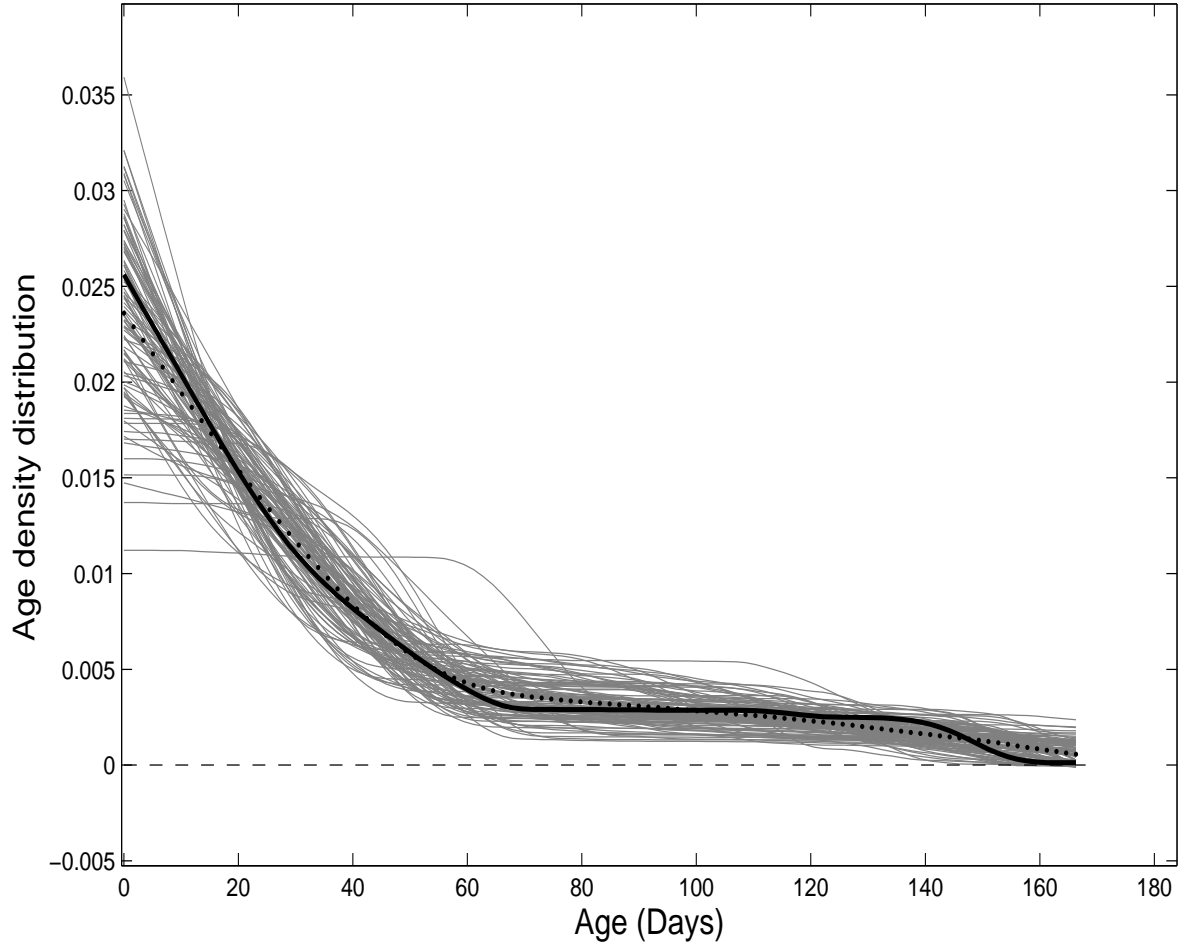


Figure 2: Estimated density f_A (thick solid) and 100 simulated deconvolution density estimates (grey lines), each one obtained by sampling age-at-capture from the original wild age density estimate shown in Fig. 1 and then creating an artificial captive cohort which serves as input for the deconvolution algorithm. The cross-sectional average of the deconvolution estimates (dotted line) is seen to be close to the target function.

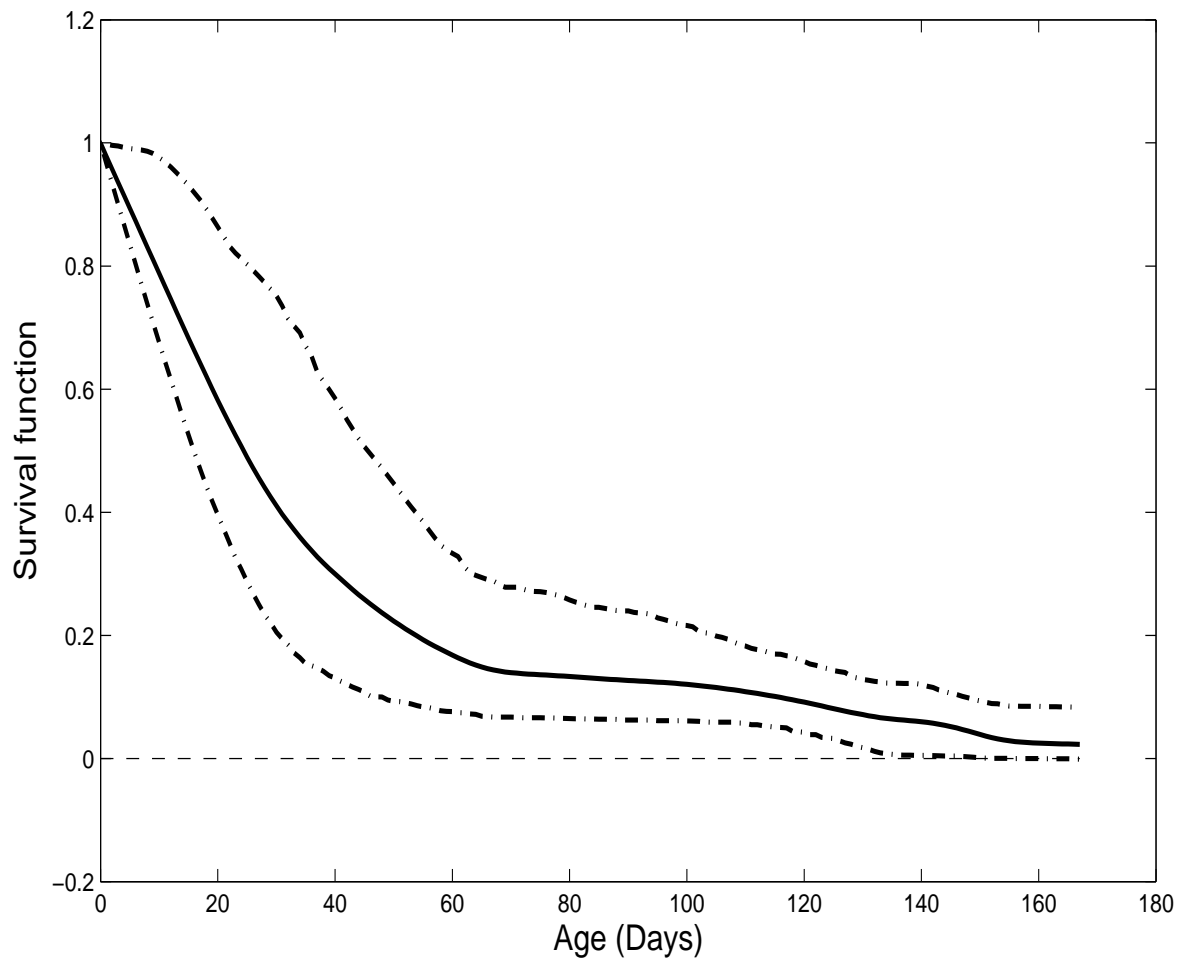


Figure 3: Estimate of the survival function in the wild *Bactrocera oleae*, with 95% bootstrap confidence intervals.