Nonlinear relationships between vital rates and state variables in demographic models

JOHAN P. DAHLGREN,1,3 MARÍA B. GARCÍA,2 AND JOHAN EHRLEN1

1Department of Botany, Stockholm University, Stockholm SE 10691 Sweden
2Instituto Pirenaico de Ecología (CSIC), Apartado 13034, Zaragoza 50080 Spain

Abstract. To accurately estimate population dynamics and viability, structured population models account for among-individual differences in demographic parameters that are related to individual state. In the widely used matrix models, such differences are incorporated in terms of discrete state categories, whereas integral projection models (IPMs) use continuous state variables to avoid artificial classes. In IPMs, and sometimes also in matrix models, parameterization is based on regressions that do not always model nonlinear relationships between demographic parameters and state variables. We stress the importance of testing for nonlinearity and propose using restricted cubic splines in order to allow for a wide variety of relationships in regressions and demographic models. For the plant *Borderea pyrenaica*, we found that vital rate relationships with size and age were nonlinear and that the parameterization method had large effects on predicted population growth rates, \( \lambda \) (linear IPM, 0.95; nonlinear IPMs, 1.00; matrix model, 0.96). Our results suggest that restricted cubic spline models are more reliable than linear or polynomial models. Because even weak nonlinearity in relationships between vital rates and state variables can have large effects on model predictions, we suggest that restricted cubic regression splines should be considered for parameterizing models of population dynamics whenever linearity cannot be assumed.

Key words: *Borderea pyrenaica*; demography; integral projection models; linearity assumption; matrix models; natural splines; nonlinear regression; plant population dynamics; restricted cubic splines.

INTRODUCTION

Individuals of different age or size often differ in survival probability and fecundity, and thus in their contribution to population growth. Due to such differences, it is important to include a state or stage structure in models of population dynamics. In matrix models, populations are divided into discrete states and the transition rates between states over a time step are modeled using a transition matrix (Leslie 1945, Lefkovitch 1965, Caswell 2001). The Integral Projection Model (IPM; Easterling et al. 2000) is an extension of the matrix model, allowing state variables to be continuous. In both matrix models and IPMs, effects of state variables on vital rates can be incorporated as regression models (Morris and Doak 2002, Ellner and Rees 2006). Parameterization using continuous functions is more statistically efficient than the traditional method of parameterizing classes separately in matrix models, and allows more reliable estimates for sparsely sampled groups, such as old individuals. However, one potential problem with using regression analyses for model parameterization is that relationships between vital rates and state variables may not be adequately described by simple or polynomial generalized linear models.

Nonlinear relationships are sometimes accounted for in demographic models parameterized using regression analyses. For example, exponential relationships between state variables and vital rates are accounted for by log-transformations of state variables, or by the use of Poisson regression. In addition, probabilities of survival and flowering are typically modeled with logistic regressions, implicitly accounting for asymptotic relationships. Also, incorporation of polynomial terms in regression models is a common practice to account for nonlinear relationships. However, using polynomial regressions to describe vital rates in demographic models may be risky because the shape of one tail of the resulting curve may be unduly influenced by observations at the other tail, in particular if the observations are not evenly spaced over the range of the predictor variable (Harrell 2001). Furthermore, neither linearizations nor the inclusion of polynomial terms may adequately describe nonlinear state variable–vital rate dependencies.
relationships if functional relationships differ between different regions of state variable values. Changes in functional relationships between state variables and vital rates may not be uncommon in organisms with marked transitions between life cycle stages. For example, the relationship between age and survival in birds has been shown to sometimes be “saw-toothed” (Low and Pärt 2009).

Incorporating nonlinear regression models into demographic models is straightforward in many situations, and this possibility was highlighted in the paper introducing the IPM (Easterling et al. 2000). However, more complex nonlinear vital rate–state variable relationships have rarely been included in demographic models. Here we illustrate a practical and reliable method of flexibly incorporating nonlinear relationships into IPMs using restricted cubic regression splines. We present results suggesting that model output may be very sensitive to assumptions of linearity, and that allowing nonlinear relationships in demographic models should be considered whenever there is no a priori reason to assume either linear or other simple function forms.

We used four different methods of parameterizing structured population models of the long-lived plant Borderea pyrenaica, in which vital rates may depend on both age and size. First, we used continuous classes parameterized using generalized linear models (GLMs), where linearity is assumed for continuous vital rate rates; logistic linear for proportions; and log linear for counts (“linear IPMs”). Second, we used continuous classes parameterized using GLMs, where the linearity assumption is relaxed by also including quadratic and cubic age and size effects on vital rates (“polynomial IPMs”). Third, we used continuous classes parameterized using GLMs where more complex effects of age and size on vital rates were allowed by using restricted cubic regression splines (“restricted cubic spline IPMs”). Fourth, we used a few discrete classes parameterized based on observed class transitions (“matrix models”), representing the most commonly used type of demographic model. We used models based on these four parameterization methods and asked (1) if relationships of vital rates with size and age were significantly nonlinear, (2) if incorporating statistically significant nonlinear relationships into demographic models influenced projections, (3) if modeling vital rates with restricted cubic splines yielded more plausible results than modeling vital rates using polynomial terms, and (4) if predictions of the matrix model, where states are discrete but relationships with vital rates may be nonlinear, more closely resemble the nonlinear models with continuous state variables than the linear model. Finally (5) we simulated data based on linear and nonlinear relationships between state variables and vital rates to investigate the consequences of modeling linear relationships as nonlinear and vice versa.

**METHODS**

**The data set**

*Borderea pyrenaica* Miègeville (Dioscoreaceae) is a small dioecious herb occurring in the central Pyrenees. The plant has morphological features that enable easy and precise aging by inspection of the underground storage organ (García and Antor 1995), making it possible to include age as a state variable in the population models, in addition to size. Individuals sometimes reach ages >300 years. The study system is described in detail in García and Antor (1995). To parameterize the structured population models, we used records of survival, size, and fecundity over four years (1995–1998) on a total of 519 female individuals in permanent plots in one population (“Pineta”). One more census in 1999, unearthing individuals, was used to determine if missing plants were dead or dormant (individuals that were missing in two consecutive censuses were assumed to be dead), and to record the age of plants by inspecting buried tubers. When calculating survival probabilities, we also included male plants because no differences between sexes have been found in previous analyses and very few observations of death were available, despite the sample size (M. B. García, J. P. Dahlgren, and J. Ehrlé un unpublished manuscript). Size was calculated as ln(number of leaves × shoot length^2), which has been found to more closely correlate with total plant biomass than other measures. Age estimates for individuals of unknown age but known size were imputed based on restricted cubic spline regressions of age on size. In order to avoid having zero mortality for the oldest individuals in the population models, an additional death of an individual of the median age of individuals in the oldest age class in the matrix model (145 years) was added to the data. This addition did not change whether relationships between vital rates and state variables were significantly nonlinear or not.

**Relaxing the linearity assumption in generalized linear models of vital rates**

There exist several methods for relaxing the linearity assumption in GLMs. Apart from simply adding polynomial terms, such methods include GAM (generalized additive models; Hastie and Tibshirani 1990), fractional polynomials (Royston and Altman 1994), and B-spline regression (e.g., Easterling et al. 2000). In this study we used restricted cubic spline regression (Stone and Koo 1985, Harrell 2001) to test nonlinearity and model vital rates because of the ease of incorporating complex nonlinearities in relationships between vital rates and one or several variables. Regression spline models are constituted by piecewise polynomial regressions that are fitted between join points (knots). Restricted cubic splines are cubic polynomial splines.
that are forced to have similar curvature at the join points, and with the tails of the curve restricted to be linear so as to make them more reliable for predictions than in other regression spline techniques. These restrictions mean that the number of parameters needed to describe the fitted curve equals the number of knots minus one. The exact location of knots has been found to be of little importance and 4- or 5-knot restricted cubic splines with evenly spaced knots satisfactorily model many kinds of nonlinear relationships, assuming a sufficient number of observations to avoid overfitting (Harrell 2001). More specifically, we used “5-knot restricted cubic splines” fitted using the “Design” package for R version 2.9.1 (Harrell 2001, R Development Core Team 2009). The “rcs” function in the Design package was used to automatically divide predictor variables with evenly spaced knots, and models were fitted using the generic “glm” function. The function named “Function” in the Design package was used obtain an expanded function describing the fitted regression model in standard S (R) language, for incorporation into the IPM code.

**Demographic models**

One linear, one polynomial, and one restricted cubic spline IPM, as well as a matrix model, were specified and population growth rates obtained from the four models were compared. For parameterization of the IPMs, generalized linear models were used to fit relationships between the state variables (age and size) and the vital rates survival, mean and variance in growth, flowering, and seed number. The data were pooled over years. Random individual effects were not significant and were omitted in final models (M. B. García, J. P. Dahlgren, and J. Ehrlén, unpublished manuscript). The integral projection models were specified as

\[
    n(y, a + 1, t + 1) = \int K(y, x, a)n(x, a, t) \, dx
\]

where \(a\) is age, \(x\) is size in year \(t\), and \(y\) is size in year \(t + 1\), and integration is performed over all possible sizes. Age was assumed to be continuous in the regressions determining relationships with vital rates, but we did not integrate over age in the population model because age of an individual is increased by one for each time step (modeled year). In the IPM equation, \(n\) is a density function describing the distribution of individuals among ages and sizes, and \(K\) is the projection kernel given by functions of survival \((s)\), growth \((g)\), flowering probability \((f_p)\), seed number \((f_s)\), and the size distribution of establishing seedlings \((p_d)\):

\[
    K(y, x, a) = s(x, a)g(y, x, a) + f_p(x, a)f_s(x, a)p_d(y).
\]

Growth was modeled as a normal distribution with the mean from a linear regression of \(y\) on \(x\) and \(a\) and the variance as the variance around this regression line. Survival was modeled as a logistic regression of presence or absence on \(x\) and \(a\), flowering as a logistic regression on flowering or not flowering on \(x\) and \(a\), and seed number as a Poisson regression on \(x\) and \(a\). In the implementation of the IPM, the integration over size was performed using the midpoint rule and a \(100 \times 100\) sized matrix (cf. Ellner and Rees 2006) for each of the 300 possible ages. Changes in survival probability were assumed to cease when individuals reached age 200, as further projections depended on few data. At age 300, individuals were entered into an absorbing “300+” class (few individuals reached this class, which is in accordance with the field data).

For each vital rate in the IPMs, a linear, a polynomial, and a restricted cubic spline regression model on age and size were fitted separately. Nonsignificant effects \((P > 0.05)\), based on likelihood ratio tests, were removed. Significance tests of nonlinear spline components were based on the null hypothesis that all nonlinear spline components equaled zero (Harrell 2001).

In order to compare estimated population growth rates between the IPMs and a matrix model, we also specified a model with three size classes and five age classes. This model was parameterized based on observed transitions between classes (cf. Morris and Doak 2002). Size classes were distributed approximately evenly over the observed size ranges and were separated at sizes 6.5 and 8.5 \((\ln(\text{number of leaves} \times \text{shoot length}^3))\). Age classes were wider for more advanced ages and separated at ages 20, 40, 80, and 150 years.

Population growth rates \((\lambda)\) were estimated by iterating the models until stable distributions were reached. Elasticity values showing the proportional consequences of small changes in either reproduction or state transition probabilities for population growth rate were calculated based on the stable distributions and reproductive value distributions as in Ellner and Rees (2006). Confidence intervals of \(\lambda\) were calculated for each model from 1000 bootstrap replicates using the package “boot” in R. Observations (individuals) in the same number as recorded were resampled with replacement from the data matrix and the models were re-parameterized and \(\lambda\) was calculated for each replicate.

**Simulations**

To investigate the consequences of choosing the “wrong” parameterization method and to decide which regression modeling approach may be the most reliable, we performed simulations where true vital rate curves were known. We modeled three different scenarios corresponding to the true relationships, being the linear, polynomial, and spline models presented in Table 1, respectively. For each scenario, 1500 individuals of random age and size were drawn based on the observed distributions, taking the covariance into account by using Cholesky decomposition in the “mvtnorm” R package. In a second step, vital rates were assigned to
individuals using the parameters from the “true” regression models. In accordance with the regression analyses, random error distributions were assumed to be binomial for survival and flowering, normal for growth, and Poisson for seed number. To the simulated data, we fitted linear, polynomial, and restricted cubic spline models, resulting in nine combinations of true and modeled relationships. Quadratic and cubic age and size terms were included in all polynomial models, and 5 knots were used in the restricted cubic spline models. Because we used a simulated data set with many observations, we included all terms regardless of their statistical significance, because full models have been found in other studies (Harrell 2001) to produce more reliable predictions than reduced models. The vital rate functions fitted to the simulated data were used as components of IPMs and population growth rate was calculated as described previously. All steps were repeated 200 times for each of the three scenarios. In addition, we used 500 observations per vital rate to examine if relative differences between methods remained the same.

RESULTS AND DISCUSSION

Relationships between states and vital rates in Borderea pyrenaica were nonlinear in several cases (Table 1), most markedly so for survival (logit-transformed in the logistic regressions; Fig. 1). The nonlinear relationship between age and survival entailed that no effect of age was found when assuming linearity. Age also had a significant nonlinear effect on seed number in both nonlinear models (log-transformed in the Poisson regressions; Fig. 2), Size effects on growth, flowering probability, and seed number were significantly, but only slightly, curved after transformations. Nonlinear relationships between state variables and vital rates also have been recorded in other systems (e.g., Zuidema et al. 2010; but see, e.g., Easterling et al. 2000). Indeed, it seems unlikely that perfectly linear relationships are common even after various statistical transformations, as it would imply that the total effect of all traits affecting vital rates is linearly related to the state variable used.

Long-term population growth rate (λ) was estimated to be 1.00 in both the polynomial and restricted cubic spline nonlinear IPMs and 0.95 in the linear IPM (Fig. 3). The true population growth rate is not known, but given that Borderea pyrenaica is very long-lived and population size has remained stable for many years, a mean population growth rate of 1.00 seems more reasonable than a growth rate of 0.95. Fecundity elasticity constituted 0.08% of total elasticity in the polynomial model and 0.01% in the spline model. Both differed markedly from the linear model, in which fecundity was found to be much more important for population growth rate and constituted 4.59% of total elasticity. This shows that assumptions of functional relationships can be important in demographic models (cf. Pfister and Stevens 2003). The substantial difference in λ resulted mainly because positive age effects on survival were nonlinear, and no effects were found when assuming linearity (results not shown). However, nonlinearities in size effects also changed predicted growth rates by several percentage points.

Polynomial and restricted cubic spline vital rate models explained a similar amount of variation (Table 1), and population growth rates in the polynomial and restricted cubic spline IPMs were similar (Fig. 3). However, there were important differences in both the vital rate and population models. Most notably, seed number and flowering probability were predicted to increase sharply for plants of very high ages in the polynomial models, but not in restricted cubic spline models (Fig. 2). As a result, fecundity elasticity increased with age for very old ages (250–300 years) in the polynomial model, suggesting that changes in the fecundity of old individuals would affect population growth rate more than changes in middle-aged individuals. This elasticity pattern appears highly unlikely and illustrates the problems with using polynomial regression for parameterization of demographic models.

---

Table 1. Significant relationships in generalized linear models of Borderea pyrenaica vital rates over linear and nonlinear (polynomial regression or restricted cubic splines) age and size predictors.

<table>
<thead>
<tr>
<th>Vital rate</th>
<th>Predictor explained (%)</th>
<th>AIC</th>
<th>Linear model</th>
<th>Deviance</th>
<th>Predictor explained (%)</th>
<th>AIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Survival</td>
<td>Size*** 11</td>
<td>707.6</td>
<td>Age2*** + Size2** 14</td>
<td>692.4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Growth</td>
<td>Size*** 88</td>
<td>1115</td>
<td>Age2*** 88</td>
<td>1104</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Flowering</td>
<td>Size*** 32</td>
<td>574.0</td>
<td>Age2*** 33</td>
<td>567.7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>No. seeds</td>
<td>Age*** + Size*** 27</td>
<td>4401</td>
<td>Age*** + Size*** 32</td>
<td>4206</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Notes: In the predictor columns, a “+” symbol indicates that statistics are from models including both age and size. Key to abbreviations: AIC, Akaike information criterion; df, degrees of freedom in the null model. For polynomial models, superscripts refer to whether models are quadratic or cubic; quadratic terms are also included in cubic models. For spline models, all significant parameters were also significantly nonlinear and 5-knot restricted cubic spline functions were used for each included variable.

* P < 0.05; ** P < 0.01; *** P < 0.001.
In the (matrix) population model with a few discrete classes, \( \lambda \) was estimated to be 0.96. Contrary to our expectations, predicted \( \lambda \) differed significantly from predictions of the nonlinear IPMs but not of the linear IPM. The substantial difference in \( \lambda \) compared to the nonlinear IPM may possibly have been caused by the fact that a matrix model with few classes was not as flexible in capturing nonlinear relationships in this case.

Fitting linear and nonlinear models to simulated data where the true functional relationships were known showed that restricted cubic spline models were clearly more reliable than linear models, and slightly more reliable than using polynomial models. Simulation results suggested that modeling vital rates using restricted cubic splines when they are, in fact, linear would result in a slight overestimation of population growth rate (\( \Delta \lambda = 0.010 \)) and that using polynomial regression would lead to a little larger overestimation (\( \Delta \lambda = 0.013 \)). Variation in predicted \( \lambda \) was similar when using restricted cubic splines (SD = 0.0267) and polynomials (SD = 0.0263), and only a little larger than when using linear regression (SD = 0.0173). On the other hand, modeling relationships as linear when they were nonlinear (either based on polynomials or restricted cubic splines) would lead to a larger deviation from the true value (\( \Delta \lambda = 0.045 \)) than would applying nonlinear models to linear relationships. When the underlying relationship was nonlinear, standard deviations were

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Deviance explained (%)</th>
<th>AIC</th>
<th>df</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age*** + Size*</td>
<td>15</td>
<td>692.3</td>
<td>1287</td>
</tr>
<tr>
<td>Size***</td>
<td>88</td>
<td>1103</td>
<td>719</td>
</tr>
<tr>
<td>Size*</td>
<td>33</td>
<td>572.4</td>
<td>659</td>
</tr>
<tr>
<td>Age** + Size*</td>
<td>32</td>
<td>4213</td>
<td>368</td>
</tr>
</tbody>
</table>

### Fig. 1. Survival probability and growth of the long-lived Borderea pyrenaica (Dioscoreaceae; a small dioecious herb of the central Pyrenees) as functions of size (at mean age) and age (at mean size) based on restricted cubic spline (solid gray lines), polynomial (dashed gray lines), and linear (solid black lines) generalized linear models (GLMs). Relationships with survival probability were modeled with logistic regression, and growth relationships were modeled with linear regression. Size was measured as \( \text{ln(number of leaves} \times \text{shoot length})^2 \); age was measured in years. Values are back-transformed in the figure. Age did not significantly affect growth in any of the models.
Fig. 2. Flowering probability and number of seeds per plant as functions of size (at mean age) and age (at mean size) based on restricted cubic spline (solid gray lines), polynomial (dashed gray lines), and linear (solid black lines) GLMs. Relationships with flowering probability were modeled with logistic regression, and seed number with Poisson regression. Values are back-transformed in the figure. Age did not significantly affect flowering probability in the linear and restricted cubic spline models.

Fig. 3. Deterministic population growth rates, λ (estimated mean with bootstrapped 95% confidence intervals), predicted from four structured population models, including IPM, integral projection models.
similar using all three modeling techniques (0.0087–0.0102). Thus, taken together, the simulation results suggest that spline models are stable as well as more accurate, on average. Sample size did not seem to affect these relationships (i.e., results were similar when using 500 instead of 1500 observations in the simulations), and polynomial and restricted cubic regression spline models both predicted true relationships of the other model type well ($\Delta \lambda < 0.003$).

In many cases of applied regression, assuming linearity, or strict adherence to, e.g., exponential or logistic curves, may be unwarranted (Harrell 2001, Sauerbrei et al. 2007). Even though assuming linear relationships may be motivated in demographic models for many species (e.g., Morris and Doak 2002), and taking into account the fact that nonlinear models may not provide better fits in all species (e.g., Easterling et al. 2000), the diverging predictions of the continuous models presented here show that this assumption is critical and that routine-like use of generalized linear models may hamper our understanding of population dynamics and of population viability. Restricted cubic regression splines are both relatively simple and sufficient to model continuous nonlinear relationships (Harrell 2001), making the method ideal for modeling vital rates when there is no mechanistic knowledge of the effects of state variables. If there is knowledge about mechanisms, models such as the growth model used by Zuidema and colleagues (2010) in IPMs of tropical trees are important options. In the present study, join points between splines (knots) were evenly spaced among observations. If prior information regarding “class divisions” exists, users may also choose to specify knot locations in spline models.

In conclusion, our results suggest that even weak nonlinearities may have important effects in IPMs. For our data set, polynomial and restricted cubic spline regressions yielded similar population growth rates, but both based on statistical theory and as shown by the unrealistic vital rate relationships and the simulations, restricted cubic spline regression is likely to be a more reliable method. We thus recommend careful tests for nonlinear relationships and the use of a method such as restricted cubic splines to parameterize models of population dynamics in cases where linearity cannot be assumed. Inclusion of nonlinear relationships is likely to make demographic models more realistic without large increases in complexity. Such models should therefore, in many cases, be an optimal option for modeling population dynamics and examining population viability.

Acknowledgments

This study was financially supported by the Strategic Research Programme EkoKlim at Stockholm University (J. P. Dahlgren), the National Parks project (ref 018/2008; M. B. García), and the Swedish Research Council for Environment, Agricultural Sciences and Spatial Planning (FORMAS; J. Ehrlén).

Literature Cited


