Ecological Monographs 71(1), 2001, pp. 1-25

PARTIALLY SPECIFIED ECOLOGICAL MODELS

Simon N. Wood

Statistical Ecology Group, The Mathematical Institute, University of St Andrews
North Haugh, St Andrews KY16 9SS, U.K. snw@st-and.ac.uk

Abstract. Models are useful when they are compared with data. Whether this comparison should be qualitative or quantitative depends on circumstances, but in many cases some statistical comparison of model and data is useful and enhances objectivity. Unfortunately, ecological dynamic models tend to contain assumptions and simplifications which enhance tractability, promote insight, but spoil model fit and this can cause difficulties when adopting a statistical approach. Furthermore, the arcane numerical analysis required to fit dynamic models reliably presents an impediment to objective model testing by fitting. This paper presents methods for formulating and fitting *partially specified* models, which aim to achieve a measure of generality by avoiding some of the irrelevant incidental assumptions that are inevitable in more traditional approaches. This is done by allowing delay differential equation models, difference equation models and differential equation models to be constructed with part of their structure represented by unknown functions, while part of the structure may contain conventional model elements that contain only unknown parameters. An integrated practical methodology for using such models is presented along with several examples, which include use of models formulated using delay differential equations, discrete difference equations/matrix models, ordinary differential equations and partial differential equations. The methods also allow better estimation from ecological data by model fitting, since models can be formulated to include fewer unjustified assumptions than is usually the case if more traditional models are used, while still including as much structure as the modeller believes can be justified by biological knowledge: model structure improves precision, while fewer extraneous assumptions reduce unquantifiable bias.

Key words: population dynamic model fitting; partially specified model; semi-parametric model; semi-mechanistic model; multiple smoothing parameter; non-linear spline model; delay differential equation; ecological dynamic model fitting.

Introduction

"The population went up and down, and so did the model" is not totally unjustified as a caricature of the way in which ecological models are often compared to data. There are good reasons for this. Most formal theory for statistical data modelling is based on linear and close to linear models. There is not an equivalent body of theory for statistical analysis using the kind of non-linear models on which the theory of ecological dynamics is based. Furthermore, many models contain assumptions that have been introduced for reasons of tractability, rather than biology. These can be expected to result in mismatches between model and data, even when the biology underpinning a model is correct, and this undermines the utility of formal comparison of models with data. This paper aims to reduce these problems for an extensive family of non-stochastic models.

Why fit models? The most obvious and practical motivation is to infer something about the system to which the model is being fitted. For example, mortality rates are very difficult to measure directly in the field, but population densities are easier: by fitting an appropriate model to the latter it may be possible to infer the former. Falsification and validation of models is another reason. Formal fitting can show whether a model is really capable of producing observed dynamics or not, as well as pinpointing the features of data that are not explained

by the theory embodied in a badly fitting model. Of course many models produce such ridiculous dynamics that they have fulfilled their purpose and can be discarded long before fitting is necessary, but for many others fitting is helpful. Some models are produced for prediction and here it is particularly important to calibrate models by fitting to data. Finally there is the comparison of models (hypothesis testing): one way of distinguishing between competing hypotheses about how a system works is to formulate these hypotheses as models and compare their ability to fit data from the system.

Although it may be desirable to fit models to data, it also turns out to be difficult. Most dynamic ecological models are non-linear, so even if a sensible measure of fit can be defined, methods that are guaranteed to find the *best* fit do not generally exist. There are therefore a large number of alternative non-linear optimization methods to choose from. Treating these methods as black boxes tends to give variable results: for some models fitting seems straightforward, while for others the fitting method makes very slow or no progress, terminates for no apparent reason or 'converges' to parameter values that depend strongly on the initial parameter values used. Such practical difficulties substantially undermine the usefulness of model fitting as a scientific tool.

There are at least three reasons for these difficulties. Firstly, it is necessary to choose some measure of goodness of fit in order to fit models, and it is very easy to come up with choices that are very difficult to minimise. Secondly, even for well behaved measures of model fit, the numerical analysis involved in solving models and calculating the quantities required by fitting methods must be performed carefully: otherwise methods will be slow and convergence unreliable. Thirdly, efficiency and reliability are undermined if methods fail to make good use of the structure of the model fitting problem. One goal of this paper is to overcome the technical obstacles for one class of models and fitting objectives.

A further obstacle to useful model fitting is more fundamental. Most models contain elements that are not derived entirely from mechanistic first principles. Instead, some parts of the model are phenomenological characterisations of a process or relationship. These terms introduce incidental assumptions into a model that have nothing to do with the biological mechanisms on which the model is based. It is usually assumed that these incidental assumptions will have little effect on the qualitative nature of the model's dynamics, but this is by no means guaranteed (see Wood and Thomas 1999, for an extreme example). In any case, incidental assumptions will produce quantitative effects and these may have implications when it comes to interpreting lack of model fit. Ideally, lack of fit should indicate that something is wrong with the biological assumptions of the model, rather than the incidental assumptions, but disentangling these two is very difficult. The issue is clearest when comparing the fit of two alternative models of a system. In this case interest focuses on distinguishing between the alternative mechanisms that the models embody, but model fit is contingent on incidental modelling assumptions as well as assumptions about biological mechanism. If the incidental assumptions differ between models it is hard to know whether a difference in fit is attributable to the differing biological assumptions made or merely to differences in the incidental modelling assumptions.

Further examples of the difficulty with incidental assumptions arise in pure estimation problems. For example, when estimating mortality rates from observed (structured) population time series it is usually the case that the form of the mortality rate term in the fitted model is not known *a priori*. Simply assuming a form means that estimates are conditional on an untestable assumption: a problem that compromises estimates and confidence intervals (see e.g. Wood and Nisbet 1991).

The problems introduced by model misspecification are well known in other contexts. Figure 1 shows an example in the context of linear regression: over-specification leads to apparently small confidence intervals but substantial bias (which can not be estimated - bias correction procedures assume a correct model and statistical consistency of the estimators of the model unknowns); too much flexibility in the specification leads to inflated confidence intervals. Similarly, non-parametric statistics is aimed at exactly the problem of not knowing the correct parametric model for the random component of data. In this statistical context the benefits of non-parametric methods are well known and usually accrue despite the relative robustness with which the central limit theorem endows equivalent parametric methods. When constructing models for the systematic component of data, rather than for the noise component, there is no equivalent of the central limit theorem: hence the benefits of avoiding baseless parametric representations of model components should be even more obvious. This has been recognized

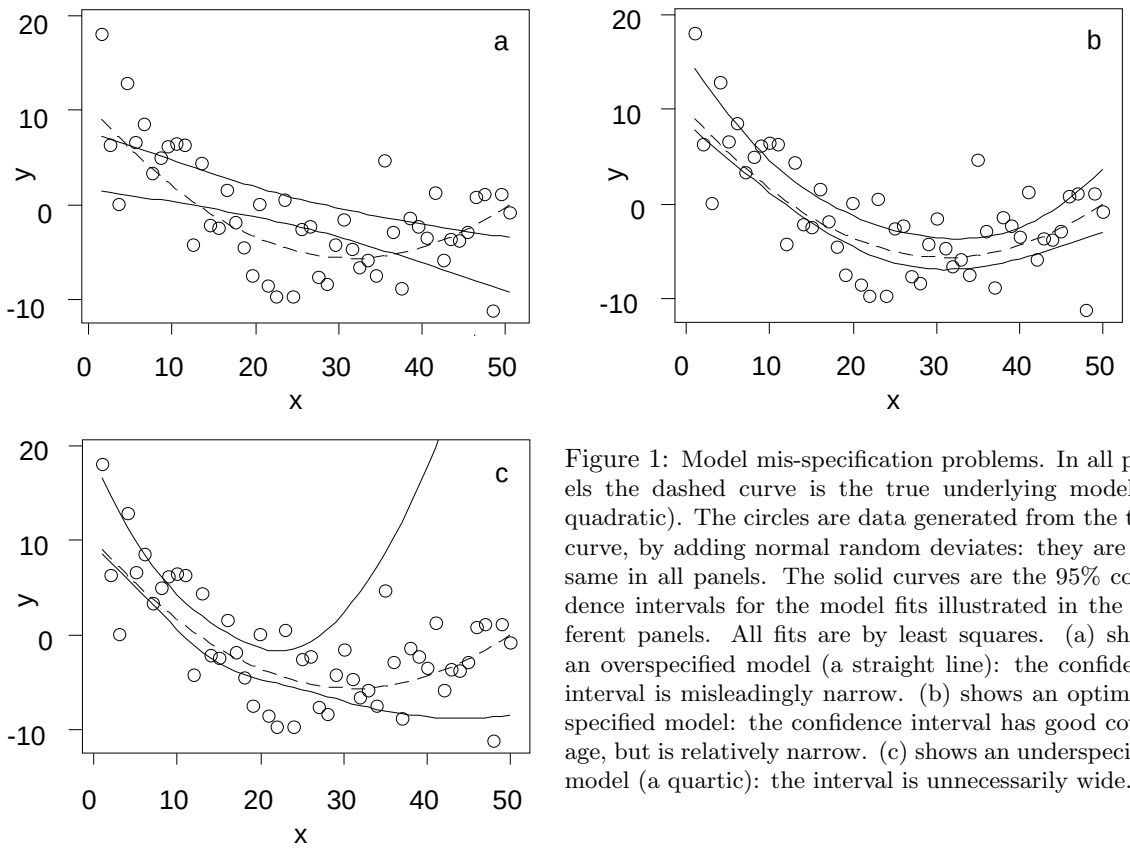

Figure 1: Model mis-specification problems. In all panels the dashed curve is the true underlying model (a quadratic). The circles are data generated from the true curve, by adding normal random deviates: they are the same in all panels. The solid curves are the 95% confidence intervals for the model fits illustrated in the different panels. All fits are by least squares. (a) shows an overspecified model (a straight line): the confidence interval is misleadingly narrow. (b) shows an optimally specified model: the confidence interval has good coverage, but is relatively narrow. (c) shows an underspecified model (a quartic): the interval is unnecessarily wide.

in conventional statistical modelling with the advent of non-parametric approaches such as Generalized Additive Models, splines and other formal smoothing methods (e.g. Hastie and Tibshirani 1990; Wahba 1990). In these methods models are specified in terms of linear combinations of fairly general *smooth functions* of the statistical covariates, hence avoiding the problems associated with the model mis-specification that must follow from assuming more arbitrary, fully parametric, model forms.

It seems sensible to take a similar approach when trying to reduce the problems of mis-specification in ecological dynamic models. One of the most obvious ways in which unwanted assumptions may be introduced into a model is via the specification of the functional forms relating model variables: if a relationship is poorly understood then the use of a parameter sparse functional form to represent it is likely to introduce spurious assumptions into the model. This can be avoided by employing rather general function specifications in such cases, so that problems are not introduced by the specification itself. In special cases this approach has been applied previously. Wood and Nisbet (1991) and Wood (1994) used the fitting of flexible and very general models of structured populations to extract death rate information from structured population data. Both studies used extensive Monte Carlo experiments to demonstrate the superiority of this approach to (published) methods employing more tightly specified models based on less biologically defensible assumptions. The key feature of the models used was that only quite well understood elements of the biology of the modelled systems were described in a prescriptive manner, while less well known parts of the biology were represented in a flexible non-parametric way. The hope is that, by taking some care to avoid producing a model that implicitly overstates how much is known about a system, it should be possible to produce a model structure that is capable of being a reasonable approximation to the truth. This improved model fidelity ought in turn lead to more reliable estimates when the model is fit to data. The practical utility of the approach has been further confirmed

since (e.g. Ohman and Wood 1996), and Wood (1994) also suggested how these methods could be generalized. In the context of measles epidemics Ellner *et al.* (1998) clearly and elegantly demonstrate the additional insight to be gained from what they term 'semi-mechanistic' models. Similarly, Bjørnstad *et al.* have use simple population dynamic models formulated as generalized additive models to analyse cod (1999) and Indian meal moth (1998) populations. But while the approach has been tried and proven in special cases, more general methods facilitating wider use are lacking: a deficit that this paper attempts to address.

To summarise: the ideal solution to the model mis-specification problem is to write down models which contain only what is actually known about the workings of a biological system. This ideal is impractical to achieve, but it is possible to move some way towards it by only specifying model relationships in general terms when concrete knowledge does not justify imposing extra structure in the form of detailed parametric specification. Such an approach has produced rich rewards when applied to more conventional statistical modelling (Wahba, 1990; Hastie & Tibshirani, 1990), and has shown considerable promise in the context of modelling ecological dynamics (Wood and Nisbet, 1991; Wood 1994, 1997; Ohman and Wood, 1996; Ellner *et al.*, 1997; Ellner, 1998, Bjørnstad *et al.* 1998, 1999). What is missing in the ecological context is general methodology to allow this approach to be employed in more than a few special cases. The partially specified modelling framework introduced in the next few sections of this paper is an attempt to fill this gap. In order to produce something that works I will consider a limited class of models and will only consider some fairly straightforward measures of fit: the latter restriction in particular is pragmatic and is not intended to imply that the measure of fit used is the only sensible measure.

Partially Specified Models

The idea is best introduced by example, so consider a simple predator prey model. Let P and N denote predators and prey respectively and α , β , γ and δ be parameters with obvious interpretations. The dynamics of predators and prey might be described by:

$$\frac{\mathrm{d}P}{\mathrm{d}t} = \alpha PN - \beta P \qquad \frac{\mathrm{d}N}{\mathrm{d}t} = \gamma N - \delta PN$$

I will refer to this as a *fully specified model*, since the interaction between predators and prey is spelled out quite clearly with some simple parametric model terms. Fitting such a model to time series of predator and prey abundances would involve finding values for α, β, γ, δ and (usually) the initial populations that best matched the data according to some criterion determined by the modeller.

Most ecologists do not believe that the proportional mixing assumption incorporated in the above model is actually correct and conceptually it would be preferable to work with a model that comes closer to embodying what *can* be stated with confidence. For example the *partially specified* model:

$$\frac{\mathrm{d}P}{\mathrm{d}t} = f(P,N) - \beta P \qquad \frac{\mathrm{d}N}{\mathrm{d}t} = \gamma N - \delta f(P,N)$$

retains considerable structure, but involves a much less specific assumption about the nature of the predation process. In reality a little more structure should probably be imposed in this case: for example, f should be smooth (meaning that its first few derivatives should be nowhere too large) and

$$\delta, \beta, \gamma, f > 0; \quad \frac{\partial f}{\partial N} > 0; \quad \frac{\partial f}{\partial P} > 0$$

Fitting this model involves somehow finding the parameter values and function f that result in the best fit of the model to data. Of course there are many alternative partially specified models for this situation, some more tightly specified and some less so.

This simple example is intended to introduce the concept of partial model specification: models are written with some components represented by unknown functions, rather than specific functional forms, and qualitative information is introduced as bound constraints on the functions and parameters of the model. The hope is that models specified in this fairly general way will be capable of closer approximation to reality than more tightly specified models. The approach gives the modeller increased control over the assumptions made in modelling by removing the need for the most common source of unwanted assumptions: namely the need to write down *something* expedient in place of the general '$f(\cdot)$' which may be all that is justified by biological knowledge.

The rest of this section will attempt to lay out a practical framework for partially specified modelling in more detail. There are many ways in which the framework could be extended, but I will only present what can currently be achieved technically.

Models

A wide range of population models can be written as a system of delay differential equations, with a finite number of discontinuities, including all ordinary differential equation models and discrete time models. In these models the state of a system at some time t can be encapsulated in the values of some state variables at t and at some previous times. Most of the time the rates of change of these state variables are smooth functions of the state variables and lagged state variables, but at some time points the state variables may change discontinuously. Let n_i be the value of the i^{th} state variable and $\mathbf{n}$ be the vector of all state variables at time t. Similarly let $\mathbf{n}_{t-\tau_i}$ be the values of the state variables at $t-\tau_i$ (τ_i may change with time and may be a state variable itself.) Then:

$$\begin{aligned}\frac{\mathrm{d}n_i}{\mathrm{d}t} &= g_i\left(\mathbf{n}, \mathbf{n}_{t-\tau_1}, \mathbf{n}_{t-\tau_2} \ldots, t\right) \\ &\quad \text{for all } t > 0, t \neq \{T_1, T_2, \ldots\} \qquad (1)\end{aligned}$$

where $\{T_1, T_2 \ldots\}$ is the set of points at which the state of the system changes discontinuously (the elements of this set may be state variable dependent). I will assume that g_i does not actually depend on the system state prior to $t = 0$, so that initial states, $n_i(0)$, rather than initial histories, are required to integrate the model: i.e. g_i is subject to the restriction that its partial differential with respect to any element of $\mathbf{n}_{t-\tau_i}$ is zero if $t < \tau_i$. The model may be supplemented by discontinuities:

$$n_i(T_j^+) = d_i(\mathbf{n}(T_j^-), j) \qquad (2)$$

where T_j^+ is the instant after T_j and T_j^- the instant before. The particular models given above and in the examples sections provide illustrations from the class of models.

Clearly most discrete time models and all ordinary differential equation models are special cases of this class of models. For example, by setting $g_i(\cdot) = 0$ for all i, we get the general class of models that can be written as systems of difference equations:

$$n_i(T_{j+1}) = d_i(\mathbf{n}(T_j), T_j)$$

This class includes matrix models and discrete difference equation models.

Similarly by having no discontinuities and no lags the general model becomes a model written as a system of ordinary differential equations:

$$\frac{\mathrm{d}n_i}{\mathrm{d}t} = g_i(\mathbf{n}, t).$$

Therefore everything done in this paper for models formulated as delay differential equations will apply without modification to difference equation models and differential equation models, except for some material on solving delay equations. In what follows the reader interested solely in the more restricted model classes can safely ignore the technical details associated only with delays.

Limitation to this class of models is largely pragmatic since fitting general partial differential equation models would usually require a prohibitive amount of computing with current technology. However, it is possible to solve many partial differential equations by discretisation into a series of ordinary (or even delay) differential equations (see Al-Rabeh, 1992 and the section *Example: marine copepods*), so the restriction is not overly onerous. Stochastic dynamics have been neglected because they introduce enough extra technical difficulty to double the length of this paper.

The class of models chosen covers a high proportion of the models actually used in ecology. Models written as systems of discrete equations or systems of differential equations are widely known and used. Delay differential equations (other than ordinary differential equations) are less widely used, but do allow a very wide range of situations to be modelled. In particular, a substantial amount of theory has been produced on how to employ delay differential equations to produce demographically sound, but computationally efficient, representations of (non-stochastic) population dynamics for organisms with moderately complex life cycles. Accessible introductions to the use of systems of d.d.e.'s to model age-structured populations and more general physiologically structured populations are given in Gurney and Nisbet (1998) and Nisbet (1997) and their application is amply illustrated by Gurney *et al.* (1983, 1986) Gurney and Nisbet (1985) or Nisbet and Gurney (1983), for example. It is worth noting that delay differential equation modelling is by no means retricted to the kind of heroic phenomenological characterisation that typifies some classic examples of their use (e.g. May 1974). For example, by employing mixtures of delay and ordinary differential equations it is straightforward to produce demographically rigorous models of populations of organisms in which individual development is heterogeneous in time and between individuals (see Blythe *et al.*, 1984; Macdonald, 1978, 1989).

The g_i's and d_i's in (1) and (2)will usually depend on unknown coefficients c_i and also some unknown functions f_i. Given particular f_i's and c_i's

the model can be solved to give estimates of the state variables at any time (or, more generally, estimates of functions of the states variables). Hence, given observations of some of the state variables (or functions of state variables) at some times, it should be possible to find the functions and coefficients that cause the model to best fit these data. Suppose that observations of state variables (or transformations of state variables) can be written as a vector $\mathbf{y}$. This vector will generally contain observations of several state variables at a number of times arranged in some order, the details of which are unimportant. What matters is that given c_i's and f_i's the population dynamic model can be solved (usually numerically) to produce a vector of model estimates $\boldsymbol{\mu}$ corresponding to the observations in $\mathbf{y}$. So the model can be thought of as a function(al) $\mathbf{M}$ mapping the unknown functions and coefficients of the model to the model predictions of the data.

$$\boldsymbol{\mu} = \mathbf{M}(f_1, f_2, \ldots, c_1, c_2 \ldots)$$

Given a correct model structure and the true values of the unknown functions and coefficients, and also neglecting all stochasticity except sampling error, the model states that $\boldsymbol{\mu} = E(\mathbf{y})$. Even with more realistic assumptions about stochasticity $\boldsymbol{\mu}$ will usually tend towards $E(\mathbf{y})$ as population size increases.

A complete model may contain some additional elements. The sampling distribution of the observed data, $\mathbf{y}$, may be specified and there is other information that might be included. For example, bound constraints may be available on some or all of the model parameters. Similarly, constraints may be imposed on the unknown functions in a model: the easiest to deal with are constraints that can be expressed in terms of linear functionals of the unknown functions (a *functional* is just a function of a function). For example, it may be important to insist that the function be monotonic, convex or positive, or some combination of these (i.e. f is such that $f' \geq 0$ or $f' \leq 0$ or $f'' \leq 0$ or $f \geq 0$ or some combination of these). The inclusion of this sort of qualitative information about model structure is another way of restricting the range of dynamics that the model can demonstrate, by as much as the modeller believes is justified by biological knowledge (for example: we may not know how mortality rate is related to population size, but we surely know that it is non-negative). A final conceptual element of a partially specified model is the belief that the unknown functions contained in the model should be smooth: how smooth will be discussed later, but it is usually implicit in the construction of the model at all, that its component functions should not be infinitely complex.

Fitting

Fitting partially specified models involves the conceptual difficulty that the unknown functions in a model must be represented somehow and that the *complexity* (or *flexibility*) of those functions must be chosen. Allowing too much flexibility can allow over-fitting (and over wide confidence intervals), while too little flexibility in the functions will lead to under-fitting and the unquantifiable bias that partially specified models are intended to reduce. (See Figure 1.)

The problem of balancing goodness of fit and smoothness of the unknown functions can be made quantitative by writing down an objective function that contains a term measuring badness of fit and a term measuring wiggliness of the unknown functions. An example of such an objective is:

$$\textit{minimise} \sum_{i=1}^{m_d} w_i(\mu_i - y_i)^2 + \sum_{i=1}^{m_f} \lambda_i \int \left[f_i''(x)\right]^2 dx \tag{3}$$

$$\begin{aligned} \textit{subject to } L_{ij} f_i &\geq b_{ij} \ \{i = 1 \ldots m_f, j = 1 \ldots k\} \\ \textit{and } \tilde{\mathbf{A}}_{\mathbf{c}}\mathbf{c} &\geq \mathbf{b}_{\mathbf{c}} \end{aligned} \tag{4}$$

The first term in (3) is model badness of fit measured as a weighted least squares term (m_d is the number of data), in simple cases the weights w_i might all be set to unity, or proportional to the reciprocal of the variance of y_i; the next term is a sum of 'wiggliness' measures for the model's unknown functions (of which there are m_f). Each wiggliness measure term in this summation is multiplied by a 'smoothness parameter' λ_i. λ_i controls the weight given to the goal of making f_i smooth relative to the goal of fitting the data closely (see Green and Silverman, 1994, for an accessible account of the use of penalized fitting objectives; Wahba, 1990, for more advanced examples or Villalobos & Wahba, 1987, for discussion of constrained penalized problems). Use of a weighted least squares term to measure badness of fit facilitates smoothing parameter selection, but other choices are possible. For example a (negative log) likelihood term might be used in place of the least squares term, although for likelihoods based on exponential family distributions the resulting objective would still be minimised by iterative solution of approximating problems of the

same form as (3) (the weights w_i would be adjusted at each iteration according to the model mean- variance relationship: again see Green and Silverman, 1994, chapter 5). Note that the objective is fairly general since the y_i's can be any transformations and/or combinations of the raw data so long as it can be predicted by the model: never the less, in some circumstances, it may be appropriate to use very different measures of fit tailored to the particular qualitative features of the biological system that the model is designed to investigate. Note also that the assumption that the f_i's are one dimensional is made for simplicity of presentation: equivalent multidimensional wiggliness measures are available if the f_i's are functions of more than one variable (see, e.g. Green and Silverman, 1994, chapter 7 or Wahba 1990).

The constraints (4), are the linear constraints applying to the f_i's and c_i's. The L_{ij}'s are linear functionals, and b_{ij}'s coefficients. For example, if one wanted to specify that the second constraint of f_1 is that its gradient is to be greater than one at the origin, then L_{12} would be the linear functional that differentiates f_1 at the origin and $b_{12} = 1$. Similarly $\tilde{\mathbf{A}}_\mathbf{c}$ and $\mathbf{b_c}$ are the matrix and vector containing coefficients relating to the linear constraints on the coefficients. Note that it is also possible to impose general linear constraints involving several functions and coefficients (for example $f_1(x) + f_2(x) < c_1$). Discussion of constrained optimization can be found in Gill *at al* (1981) and appendix B.

Before examining the crucial question of how to choose $\boldsymbol{\lambda}$ (and thereby the wiggliness of the unknown functions), consider how to represent the unknown functions in practice. The functions can be approximated using a suitable basis. This means constructing each f_i from the sum of some simple 'basis functions' ($\gamma_{ij}(x)$, say) multiplied by unknown parameters (α_{ij}, say) :

$$f_i(x) = \sum_j \alpha_{ij}\gamma_{ij}(x) \tag{5}$$

($\gamma_{i1}(x) = 1$, $\gamma_{i2}(x) = x$, $\gamma_{i3} = x^2$ ...is an example of a (bad) basis, a better basis is outlined in Appendix A). So the original population model will have an expression like the right hand side of (5) substituted where ever there is an unknown f_i in the original specification (in practice this can be done entirely automatically). The basis functions themselves have no unknown parameters, so finding the best fit f_i is reduced to finding the best fit parameters α_{ij}. Notice also that for any basis function representation like (5) the values and derivatives of the function at any point can be expressed as linear transformations of the parameters (for example, $f_i'(x) = \sum \alpha_{ij}\gamma_{ij}'(x)$): this is useful for turning inequality constraints on functions into general inequality constraints on parameters. The methods described here could be used with a variety of bases. A good choice is to use the basis that arises naturally in linear spline smoothing problems, as this yields an easily calculated form for the penalties, $\int[f'']^2$, and is known to have good approximation theoretic properties. Details can be found in Wahba (1990), Green & Silverman (1994) and Hastie & Tibshirani (1990), and Appendix A. Again, multidimensional functions can be used: the basis that arises from 'thin plate splines' is an obvious candidate, see Wahba (1990) or Green & Silverman (1994). If the unknown coefficients and the parameters for all the unknown functions are now collected into one vector, $\mathbf{p}^T = [\alpha_{11}, \alpha_{12}, \ldots, \alpha_{21}, \alpha_{22}, \ldots, c_1, c_2, c_3 \ldots]$, (^{T} denotes transposition) then the fitting problem (3, 4) can be written:

$$min.\ q(\mathbf{p}) = \sum_i w_i(y_i - \mu_i(\mathbf{p}))^2 + \sum_i \lambda_i \mathbf{p}^T \mathbf{C}_i \mathbf{p} \tag{6}$$

$$subject\ to\ \mathbf{A_c p} \geq \mathbf{b} \tag{7}$$

$\mathbf{C}_i$ is a matrix of coefficients that depend on the choice of basis, but not the parameters; $\mathbf{A}_c$ and $\mathbf{b}$ are a matrix and vector of coefficients defining the linear constraints (note that *any* basis function representation, which, like (5), is linear in its parameters, will allow the fitting problem to be written in this general form). Given particular values for the smoothing parameters, λ_i, this fitting problem is a constrained non-linear optimization problem for which solution methods will be given in the methods section.

The second conceptual issue is the choice of smoothing parameters. This is the key to using partially specified models. Without an objective means for choosing the amount of flexibility to allow unknown functions within a model one has done nothing but replace arbitrary or *ad hoc* choices of functional forms by arbitrary or *ad hoc* choice of smoothing parameters. Fortunately there has been a great deal of statistical work on this kind of model selection problem and some methods with good theoretical and practical properties exist. Generalized cross validation (GCV, Craven & Wahba, 1979) is the one that will be employed in this paper (see Green & Silverman 1994 for a clear introduction).

Cross validation can be motivated as follows: imagine fitting a model to all your data but one, and then measuring the square of the error in your model prediction of the missing datum; now calculate the average of such squared deviations across all data points: this gives a cross validation score. This *ordinary* cross validation score measures how bad your model is at predicting missing data: how bad it is at generalising. In the current context the cross validation score will depend on the choice of smoothing parameter. Low smoothing parameters will lead to complicated (i.e. flexible and/or wiggly) models that fit the noise in the data and consequently predict missing data badly. High smoothing parameters lead to simple models that don't match the data to which they are fitted very well and do no better on the missing data. Somewhere in the middle will be better choices.

Cross validation as described has some problems. Most worryingly, it is possible to perform simple transformations on some model fitting problems in such a way that the solution and structure of the fitting problem are unchanged, but the cross validation score ceases to contain any information about the optimum smoothing parameters (see Wahba 1990; p. 53 section 4.3). Furthermore, the cross validation score can be numerically expensive except in certain special cases that don't apply in the current context. Fortunately the problems with ordinary cross validation can be fixed with generalized cross validation. The ordinary cross validation score can be shown to be equivalent to a weighted sum of squares of the deviations of the fitted model from the data. Generalized cross validation averages the weights in this summation, so that each deviation gets the same weight. Writing $\hat{\mu}_i$ for the estimate of μ_i obtained by model fitting, the resulting score is:

$$V(\boldsymbol{\lambda}) = \frac{\sum w_i (y_i - \hat{\mu}_i)^2}{[\sum (1 - \partial \hat{\mu}_i / \partial y_i)]^2} \tag{8}$$

Another way of viewing this quantity is as the estimated error variance per error degree of freedom, since $\sum(1 - \partial\hat{\mu}_i/\partial y_i)$ is an estimate of the degrees of freedom associated with the error (this can be seen by analogy with general linear regression, if $\boldsymbol{\mu} = \mathbf{X}\mathbf{p}$, then the least squares estimate of $\boldsymbol{\mu}$ is $\hat{\boldsymbol{\mu}} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}\mathbf{y}$, $\partial\hat{\mu}_i/\partial y_i$ is just element i, i of $\mathbf{A} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}$. So $\sum \partial\hat{\mu}_i/\partial p_i$ is the trace of $\mathbf{A}$, which is well known to be the number of identifiable parameters in the original linear model).

The appealing theoretical property of GCV is that, for linear problems in the large sample limit, it has been proven to select the smoothing parameters that minimise $E[\sum(\hat{\mu}_i - E(y_i))^2]$ (again see Wahba 1990, chapter 4). That is, the criterion is attempting to find the model that best matches the true underlying values of the observed data. Notice that this will not be the model that fits the data most closely, since this would involve fitting the noise as well as the signal, implying that the signal would not be optimally fit. Note also that this property is contingent on the y_i's being statistically independent (i.e. observations on independent random variables). Extensive numerical experimentation seems to confirm that the theoretical promise of GCV is realised at realistic sample sizes (again see Wahba 1990, Green and Silverman 1994 and references therein). The principal difficulty with its use is computational: searching for the optimal $\boldsymbol{\lambda}$ has the potential to be very costly if there is more than one smoothing parameter (see Gu and Wahba 1991). Additionally there is no literature on non-linear multiple smoothing parameter estimation. An alternative to GCV for smoothing parameter selection is Akaike's Information Criterion (AIC, Akaike, 1973): Burnham and Anderson (1998) is a good reference. AIC behaves in a way that is quite similar to GCV and is based on the same basic principle of trying to pick out the smoothing parameters that yield a model that gets as close as possible to some underlying 'truth'. In the current context AIC is just as difficult as GCV to work with computationally. A philosophically different route would use a hypothesis testing approach to model selection in order to find the smoothing parameters: for example, by using likelihood ratio testing to find the simplest model not falsifiable by the data at the modellers favourite significance level: I do not pursue this approach further.

In summary, partially specified models are ecological models constructed so that part of the information that they embody is in the form of unknown smooth functions, unknown coefficients, and constraints on both of these: the idea is that this extra flexibility in the structure of the models being used allows models to be produced that are an honest reflection of what is known, or what you wish to assume, about a system, but no more. Fitting such models requires that the concept of 'goodness of fit' is extended to include more than simply 'fits the data as closely as possible': one needs instead to consider how the model can be made to approximate the underlying 'true' model as closely as possible. Table 1 provides a conceptual breakdown of

Step	Model	Objective
Formulate the model and the fitting objective mathematically	$\frac{\mathrm{d}n}{\mathrm{d}t} = \beta(n_{t-\tau})n_{t-\tau} - \delta n$	$\sum_i (n_{t_i} - y_i)^2 + \lambda \int [\beta''(x)]^2 dx$
Use basis functions, $\gamma_j(\cdot)$, to represent β. i.e. $\beta(x) = \sum_{j=1}^{k} b_j \gamma_j(x)$	$\downarrow$	$\downarrow$
Given the b_j's, δ, τ and initial conditions the model can be solved numerically and the objective evaluated along with derivatives w.r.t. the parameters.	$\frac{\mathrm{d}n}{\mathrm{d}t} = \sum b_j \gamma_j(n_{t-\tau})n_{t-\tau} - \delta n$	$\sum_i (n_{t_i} - y_i)^2 + \lambda \mathbf{b}^T \mathbf{C} \mathbf{b}$
Now iterate the following 2 steps to convergence:		
1.	Given an estimate of λ adjust b_j's, δ and τ to reduce the objective $\sum_i (n_{t_i} - y_i)^2 + \lambda \mathbf{b}^T \mathbf{C} \mathbf{b}$.	
2.	Adjust λ to reduce the GCV score.	

Table 1: Example of construction and fitting of a simple partially specified model similar to the adult competition model in the section *Blowflies II*. It is assumed that an adult population n suffers *per capita* death rate δ while net *per capita* fecundity is an unknown function of n: $\beta(n)$. Individuals take time τ to mature. The example does not consider inference or constraints.

the construction and fitting process for a very simple example.

A useful discussion of inference from these models benefits from some technical background, so it is postponed until the end of the methods section.

Methods

This section covers the technical issues involved in fitting and using partially (and fully) specified models in practice and presents the arguments for using the methods employed here rather than alternatives (particularly when these are more familiar and/or simpler). I have not documented every detail, but have covered those areas which require a non-standard or novel approach (*Numerical model solution*, *Calculating* $\mathbf{J}$, *Smoothing parameter selection*, *Model fitting methods*), as well as material required for understanding of the novel material which would otherwise involve a tedious amount of supplementary reading (*model fitting methods*, *calculating* $\mathbf{J}$). The material in this section can be ignored if the goal is to fit one particular model to one set of data and it is therefore acceptable to employ a considerable degree of trial and error in the fitting process. But more ambitious goals, such as the comparison of competing models, require methods that are efficient (i.e. quick), reliable (meaning that you can tell when a best fit has been achieved) and accurate (that is, with estimation uncertainty dominated by the statistical structure of the problem rather than by propagation of numerical errors).

Model fitting

In this section it will be assumed that $\boldsymbol{\lambda}$ estimates are provided and the aim is to find best fit parameters $\mathbf{p}$, *given* some set of smoothing parameters. The basic fitting strategy will be as follows:

1. Given some estimate (guess) of model parameter vector $\mathbf{p}$, numerically solve the model equations (delay or ordinary differential equations or difference equations), to obtain an estimate of $\boldsymbol{\mu}$.
2. By repeatedly solving the model with slight changes in parameters obtain an estimate of the matrix $\mathbf{J}$ where $J_{ij} = \partial \mu_i / \partial p_j$.
3. Use the current $\boldsymbol{\mu}$ and $\mathbf{J}$ estimates to construct (or update) a quadratic model of the fitting objective.

4. Finding the parameter vector that minimises the quadratic model, suggests a direction in which to change $\mathbf{p}$ in order to minimise the real fitting objective.

The estimate of $\mathbf{p}$ can be improved by iterating steps 1. to 4. to convergence, although extra steps are required to deal with constraints and in some circumstances $\boldsymbol{\lambda}$ selection steps will also be included in each iteration (see later).

Numerical model solution: There are 3 requirements for the numerical solution of the model: speed, accuracy and stability. Speed and accuracy suggest using an explicit integration scheme with adaptive time-stepping (Press *et al.* 1992, Hairer *et al.* 1987) Adaptive stepping also ensures that the explicit scheme maintains numerical stability, so that the model integration will not 'explode' and halt the fitting process if an unfortunate set of parameters is tried by the optimization algorithm. Sensible use of inequality constraints on parameters and functions during model formulation also helps to avoid really bad parameter choices during fitting. (Readers uninterested in d.d.e. models can now proceed to *Calculating* $\mathbf{J}$).

Speed and accuracy of an integration scheme for delay differential equations are surprisingly sensitive to some of the details of numerical analysis (Highman 1993b). When numerically solving delay differential equations it is necessary to store the past values of state variables (the n_i's of (1)). Any integration scheme only calculates values and derivatives at discrete times. Never the less, it will usually be the case that the scheme will require estimates of lagged state variables *between* the times that they were stored. So interpolation is required.

The order of accuracy of the interpolation scheme should be higher than that of the integrator: if it isn't then one of two things will happen. For some non-embedded time-stepping schemes the integrator will be forced to take very small steps, due to what it perceives as frequent discontinuities in (higher) derivatives of the lagged variables: this is inefficient because it will generally take more steps than the integrator of the correct order and each one of those steps involves more calculation than a single step of the lower order integrator. For embedded time-stepping schemes the continuity assumptions on which step-length selection is based will be violated, leading to the situation in which the accuracy of the numerical solution is no longer related to the integration tolerance used (Higham 1993a,b). These factors tend to be ignored in ecological applications with most workers using integrator/ interpolators of inconsistent order: in many cases this does no more than lose a little accuracy (that no-one will miss) and waste a little computer time (which, you can be sure, the machine would not have put to better use); for the current application it leads to serious problems when anywhere close to best fit parameters.

In the work reported here I used adaptive time-stepping with an embedded Runge Kutta 2(3) scheme due to Fehlberg (Hairer *et al.* 1987, p.170) and interpolated lagged state variables using cubic Hermite interpolation (Higham 1993a, Paul, 1992). The latter necessitates the storage of lagged variables and their gradients, but, since the gradients of state variables must be calculated anyway, this presents no difficulty. Note that an additional advantage of the use of this consistent integrator/interpolator pair is that the true accuracy of the numerical solution can be estimated by making use of the 'tolerance proportionality' of the estimated solution: since accuracy is linearly proportional to integration tolerance (Higham, 1993b) accuracy can be estimated by integration the model with 2 different tolerances. This is useful for setting convergence criteria when model fitting.

Calculating $\mathbf{J}$*:* For reasons that will hopefully become clear, it is necessary to calculate a 'Jacobian' matrix, $\mathbf{J}$, where

$$J_{ij} = \frac{\partial \mu_i}{\partial p_j}$$

$\mathbf{J}$ can be approximated by finite differencing using:

$$J_{ij} \approx \frac{\mu_i(\mathbf{p} + \Delta_j \mathbf{i}_j) - \mu_i(\mathbf{p})}{\Delta_j} \tag{9}$$

where $\mathbf{i}_j$ is a vector having zeroes everywhere except in entry j where it has a 1, and Δ_j is a small number. Through diligent study of the opening chapters of any number of numerical analysis textbooks the reader will be aware that a poor choice of Δ_j can cause problems. Obviously the finite difference approximation will be poor if Δ_j is too large (truncation error). Less obviously, if Δ_j is too small then the two estimates of μ_i will be so close that almost all the bits used to store them will be identical, leaving few or no bits storing the difference (cancellation error). In the current context getting Δ_j wrong leads to serious problems, but what counts as right can be sensitive to the local shape of the objective function. Hence the usual folklore for interval estimation is best ignored in favour of adaptively

adjusting the intervals in order to keep estimates of the average truncation error roughly equal to the average cancellation error (averages are over all estimated derivatives). Appropriate error estimation formulae can be found in Gill *et al.* (1981). Given the evaluations required for Δ_j control, a more accurate finite difference formula can be used at no extra cost:

$$J_{ij} \approx \frac{\mu_i(\mathbf{p} + \Delta_j \mathbf{i}_j) - \mu_i(\mathbf{p} - \Delta_j \mathbf{i}_j)}{2\Delta_j}$$

There is a second issue relating to the calculation of $\mathbf{J}$ that is specific to the fitting of (delay) differential equation models and requires a novel approach. The model integration scheme will introduce some error into the calculation of $\boldsymbol{\mu}$ and in the interests of computational efficiency this error may be allowed to be large relative to the machine precision (after all, the data will not be measured to a great number of significant figures). Unfortunately, moderate errors in $\boldsymbol{\mu}$ can entail serious errors in derivative estimates if the errors are independent between $\mu_i(\mathbf{p})$ and $\mu_i(\mathbf{p} \pm \Delta_j \mathbf{i}_j)$. If step size control is done separately for calculation of $\mu_i(\mathbf{p})$ and $\mu_i(\mathbf{p} \pm \Delta_j \mathbf{i}_j)$ then their error terms can become almost independent. The solution is to adopt exactly the same set of time steps for integrating to find $\mu_i(\mathbf{p} \pm \Delta_j \mathbf{i}_j)$ as was used for $\mu_i(\mathbf{p})$: in this circumstance most of the error in the two approximations will cancel when they are differenced.

Model fitting methods: Given $\boldsymbol{\mu}$ and $\mathbf{J}$, calculated as described in the previous two sections, model fitting is quite straightforward and will be based on the well tested approach of iteratively approximating the fitting objective by a (multidimensional) quadratic:

$$q(\mathbf{p}) \approx a + \mathbf{h}^T\mathbf{p} + \frac{1}{2}\mathbf{p}^T\mathbf{G}\mathbf{p} \qquad (10)$$

where a is a constant, $\mathbf{h}$ is the vector of derivatives of the objective with respect to the parameters ($h_i = \partial q/\partial p_i$), and $\mathbf{G}$ is the Hessian of the objective ($G_{ij} = \partial^2 q/\partial p_i \partial p_j$). Minimising the right hand side of (10) with respect to $\mathbf{p}$ suggests an updated parameter estimate of $\mathbf{G}^{-1}\mathbf{h}$ (although it is often better to use this estimate to define the direction along which to search for a reduction in $q(p)$, rather than using the estimate directly). There are several reasons for favouring methods based on a quadratic model of the objective function. Firstly, the least squares part of the objective can be expected to be close to quadratic (for linear models, least squares objectives are exactly quadratic) and the penalty terms are *exactly* quadratic. It would be rather inefficient to make no use of this information by using gradient descent or function evaluation methods, especially given the superior convergence rates of quadratic methods (Gill *et al.* 1981). However it is probably not sensible to attempt to approximate $\mathbf{G}$ directly by finite differencing: such an approach would be numerically costly and the difficulty of obtaining accurate estimates of second derivatives would be even greater than the difficulties encountered in estimating $\mathbf{J}$. The substantial extra numerical burden would certainly not yield benefits far from the optimum parameters where the quadratic approximation itself is not good. When close to the optimum the approximations that will be employed here should be at their best and the extra benefit of directly estimating $\mathbf{G}$ is likely to be slight. In contrast, the effort of evaluating $\mathbf{J}$ can always be justified by the fact that it guarantees that a descent direction can be found (or otherwise indicates a turning point).

A final important reason for favouring the quadratic model of the objective is the existence of well known and robust methods for minimising a quadratic model subject to linear constraints of the type used here. Appendix B outlines the basic principles underpinning such constrained optimization while Gill *et al.* (1981) provide much more detailed information on the topic.

Given the basic methods for constrained optimization outlined in Appendix B, two approaches to the minimisation problem seem to work effectively. The first is to use a Quasi-Newton method that builds up an approximation to $\mathbf{G}$ using the information contained in $\mathbf{J}$ evaluated at successive parameter vector estimates during the minimisation process. This method is good for badly behaved objective functions, and in the case in which the best fit model actually fits very badly. There are a number of less than optimally stable implementations described in the literature (e.g. Press *et al.*, 1992) and it is important to use as stable an algorithm as possible in the current context: a suitable method is described in Gill *at al.* (1981), with the necessary matrix factorisations given in Gill *et al.* (1974). Maximum stability of the method is important here because $\mathbf{J}$ is being approximated relatively crudely and excessive propagation of the resulting errors should be avoided (alternatively - stable optimization methods may allow use of less accurate $\mathbf{J}$ estimates with consequent saving of numerical effort).

An alternative, that can be very fast on sufficiently well behaved problems, is to approximate the non-linear model with an approximating linear model. The following steps are iterated:

1. Using the current parameter estimate $\mathbf{p}^{(k)}$ solve the population model to obtain $\boldsymbol{\mu}^{(k)}$ and $\mathbf{J}^{(k)}$.

2. Form 'pseudodata' $\mathbf{y}^{(k)} = \mathbf{y} - \boldsymbol{\mu}^{(k)} + \mathbf{J}^{(k)}\mathbf{p}^{(k)}$.

3. Minimise:

$$(\mathbf{y}^{(k)} - \mathbf{J}^{(k)}\mathbf{p})^T\mathbf{W}(\mathbf{y}^{(k)} - \mathbf{J}^{(k)}\mathbf{p}) + \sum \lambda_i \mathbf{p}^T\mathbf{C}_i\mathbf{p}$$

$$\text{Subject to } \mathbf{A}_c\mathbf{p} \geq \mathbf{b}_c$$

to find $\mathbf{p}^{(k+1)}$ (the problem being solved here is exactly quadratic).

($\mathbf{W}$ is a diagonal matrix with $W_{ii} = w_i$). In principle these steps are repeated to convergence, although in practice the algorithm is improved by including step length selection so that the routine does not always step to the minimum implied by the approximating model. Converged estimates will be denoted $\hat{\mathbf{p}}$, $\hat{\boldsymbol{\mu}}$, etc. This method is essentially a constrained version of the Gauss Jordan method (see, for example Press *et al.* 1992), but the presentation in terms of an approximating linear model facilitates the smoothing parameter selection method presented in the next section.

The fitting methods have been presented as if the weights w_i were fixed in advance and this will often be the case. However, if some mean- variance relation is known for the data then the fitting methods presented can be used iteratively as follows. The model is first fitted with uniform (or other specified) weights given to each data point. New weights are then produced which are inversely proportional to the variance predicted from the mean variance relationship given the μ_i estimates from the best fit so far. The model is then re-fitted, weights are estimated again and the process repeated to convergence of $\boldsymbol{\mu}^{(k)}$ (to $\hat{\boldsymbol{\mu}}$). This approach is appropriate for exponential family error distributions (for example Poisson, gamma, binomial) and the algorithm described is simply the one used for generalised linear model fitting (e.g. McCullagh and Nelder, 1989), although, since no simple transformation of the μ_i's will generally linearize the models used here, there is no link function in the current case.

Finally, note that I have restricted attention to methods appropriate to the case in which the objective function is fairly smooth, and where minimisation is not made difficult by multiple local minima. There are circumstances where these assumptions do not hold, and other approaches are needed. Appendix C describes an appealingly simple approach for dealing with difficult objective functions, that enables the methods described here to be used unmodified: the objective function itself is repeatedly perturbed by bootstrapping in a way that allows local minima to be escaped. For really difficult problems other approaches such as simulated annealing may have to be used, see Brooks and Morgan (1994), for example.

Smoothing parameter selection

The first step in smoothing parameter selection is to replace the GCV score function (8) with an approximation that is practical to work with. Using the linearisation of the fitting problem employed in the model fitting section:

$$V(\boldsymbol{\lambda}) = \frac{(\mathbf{y}^{(k)} - \mathbf{J}^{(k)}\mathbf{p})^T\mathbf{W}(\mathbf{y}^{(k)} - \mathbf{J}^{(k)}\mathbf{p})}{[tr(\mathbf{I} - \mathbf{A})]^2} \qquad (11)$$

where $\mathbf{A}$ is the so called 'influence matrix', or 'hat matrix', of the fitting problem: that is the matrix that has the partial derivative of the i^{th} element of $\mathbf{J}^{(k)}\mathbf{p}$ w.r.t. $y_j^{(k)}$ in its i^{th} row and j^{th} column. Explicitly $\mathbf{A} = \mathbf{J}^{(k)}\mathbf{Z}[\mathbf{Z}^T(\mathbf{J}^{(k)T}\mathbf{W}\mathbf{J}^{(k)} + \sum \lambda_i\mathbf{C}_i)\mathbf{Z}]^{-1}\mathbf{Z}\mathbf{J}^{(k)T}\mathbf{W}$, where $\mathbf{Z}$ (see Appendix B) is any matrix whose columns form the basis of a parameter space within which unlimited movement is possible without violating the model constraints that are exactly satisfied at the estimated best fit. Close to the optimum $\mathbf{p}$ and $\boldsymbol{\lambda}$ vectors, (11) should closely approximate (8).

The principal difficulty in using GCV is the fact that evaluation of the denominator of $V(\boldsymbol{\lambda})$ is potentially very expensive. Forming $\mathbf{A}$ explicitly will take $O(m_d^3)$ operations where m_d is the number of data, and since $\mathbf{A}$ depends on $\boldsymbol{\lambda}$, direct search for the minimizing $\boldsymbol{\lambda}$ would take $O(m_d^{3m_f})$ operations, where m_f is the dimension of $\boldsymbol{\lambda}$. In the single smoothing parameter case a number of methods have been developed that are much more efficient than direct evaluation of $\mathbf{A}$, but methods for multiple smoothing parameters have proved more difficult. Fortunately, Gu & Wahba (1991) produced a way of minimising a GCV score with respect to multiple smoothing parameters for certain spline models, and it is possible to generalise their method to the much wider class of problems giving rise to scores like the one used here (Wood, 2000). Once again a quadratic model of V is used to facilitate

minimisation. The challenging problem is to find an efficient way of evaluating the gradients and second derivatives of V with respect to the smoothing parameters: the painful details are in Wood (2000), but the essence of the method is to find a sequence of orthogonal transformations of $\mathbf{A}$ that enable cheap evaluation of the GCV score with respect to one 'overall' smoothing parameter, while allowing $\mathbf{A}$ to be decomposed into components that facilitate relatively cheap evaluation of the gradients and second derivatives of the GCV score with respect to the (logs of the) λ_i's. (In many cases, the same basic computational strategy can also be used to perform smoothing parameter selection by AIC, where difficulty is caused by the need to calculate the effective number of parameters in the model, which is $tr(\mathbf{A})$.)

Use of (11) to find the smoothing parameters must proceed iteratively. When used with the iterative least squares scheme of the model fitting section, it is most efficient to re-estimate smoothing parameters after each step of the iterative scheme. In the case in which (11) is being used with the Quasi-Newton method it is better to alternate smoothing parameter estimation with full minimisation by Quasi-Newton, allowing the model parameters $\mathbf{p}$ to converge at each step. This is because the Quasi- Newton method is attempting to build up a quadratic model of the objective $q(\mathbf{p})$, so it is best not to change $q(\mathbf{p})$ during the minimisation.

Inference

Hypothesis testing (model comparison) and confidence interval estimation can be approached in two ways: by bootstrapping (Efron and Tibshirani 1993, Davison and Hinkley 1997) or by using the asymptotic distributions implied by the quadratic fitting objective.

There are several flavours of bootstrap that can be employed. The simplest is the non-parametric bootstrap. Replicate data sets are produced by sampling with replacement from the collection of y_i's with each selected y_i being accompanied by all the auxiliary information that goes with it: e.g. which stage of the population it refers to, the time at which it was sampled, and so on. In practice this means that each replicate dataset contains some of the original observations more than once, while some do not appear at all. The model is refitted to each replicate dataset and in this way an approximate distribution for the $\mathbf{p}$ estimates can be developed. Note that it is not sensible to perform cross validation on these replicates: consideration of the 'leave-one-out' idea underlying GCV suggests that smoothing parameters would be systematically under-estimated using the bootstrap replicate dataset. Hence the original smoothing parameter estimates must be used for each model fit. An advantage of the non-parametric approach is the automatic preservation of correlation structure in the residuals.

The second bootstrapping possibility is to bootstrap the residuals. In its simplest from this means that the residuals from the original model fit are collected, and treated as observations from a single distribution. Replicate data sets are produced by sampling with replacement from the residuals and adding the resampled residuals back on to the best fit model estimates, $\hat{\mu}_i$. Again the model is refitted to each replicate dataset, and after enough replicates an approximate distribution for the $\mathbf{p}$ estimates will be produced.

An approach that has advantages for small sample sizes is to bootstrap parametrically. This involves specifying a probability model for the residuals and estimating its parameters from the observed residuals. Replicate datasets are then produced by simulating residuals from the parametric residual model and adding these to the $\hat{\mu}_i$'s.

The second strategy for inference is to use asymptotic results based on the quadratic model employed in minimisation. To do this requires estimation of the covariance matrix for the parameter estimators. Assuming that the weights w_i have been chosen to be inversely proportional to the sample variances it is clear that the covariance matrix for the data is $\mathbf{V}_y = \mathbf{W}^{-1}\sigma^2$. σ^2 can be estimated in the usual way by:

$$s^2 = \frac{(\mathbf{y} - \hat{\boldsymbol{\mu}})^T \mathbf{W} (\mathbf{y} - \hat{\boldsymbol{\mu}})}{tr(\mathbf{I} - \mathbf{A})}$$

(i.e. the residual sum of squares divided by the residual degrees of freedom). There are then two possibilities for an estimate of the covariance matrix for the parameters $\mathbf{V}_p$. The conventional approach uses the fact that at convergence of the fitting algorithm it is possible to write:

$$\hat{\mathbf{p}} \approx \mathbf{k} + \mathbf{B}\mathbf{y}$$

where $\mathbf{k}$ and $\mathbf{B}$ are a vector and matrix whose exact form depends on the fitting algorithm used. Standard results (e.g. Meyer, 1975, p417) then yield the covariance matrix estimate:

$$\mathbf{V}_p \approx \mathbf{B}\mathbf{B}^T s^2$$

An alternative is to use the co-variance matrix derived for spline smoothing by Wahba (1983) using a Bayesian argument. The analogue in the current case is simply:

$$\mathbf{V}_p \approx \mathbf{Z}(\mathbf{Z^T G Z})^{-1}\mathbf{Z}^T s^2$$

in terms of the general quadratic model used for minimisation (again the columns of $\mathbf{Z}$ form the null space of any constraints satisfied exactly by the best fit model: see Appendix B). Note however that this approximation can perform badly if $\mathbf{G}$ is estimated as part of a Quasi-Newton optimization, since curvature in some directions in the model parameter space can be poorly explored. Hence it may be more useful to use:

$$\mathbf{V}_p \approx \mathbf{Z}[\mathbf{Z}^T(\hat{\mathbf{J}}^T\mathbf{W}\hat{\mathbf{J}} + \sum \lambda_i \mathbf{C}_i)\mathbf{Z}]^{-1}\mathbf{Z}^T s^2$$

implied by the iterative least squares scheme. The 'Bayesian' results tend to give higher variance estimates than the conventional results, by virtue of attempting to account for mis-specification in the final model. However in simulation studies using linear spline models the coverage properties of interval estimates based on these results have proved very good (Wahba 1990). At least in the large sample limit one expects $\hat{\mathbf{p}} \sim N(\mathbf{p}, \mathbf{V}_p)$, and inference can be based on this result.

Identifiability

Optimization methods work best if model parameters are identifiable from the outset. Thus it is best to remove parameter co-linearity from the model before attempting to fit it. However, there are automatic methods for dealing with the problem. The most direct method is to attempt to identify parameter dependence from the Jacobian $\mathbf{J}$. This can be done by linearly regressing each column of $\mathbf{J}$ on all the other columns - a high r^2 indicates problems and suggests that the parameter relating to that column should be constrained at a fixed value (which is easily done, as seen in the model fitting part of the methods section). The difficulty comes in deciding at exactly what r^2 value one should treat a parameter as non-identifiable, since the columns of $\mathbf{J}$ are only estimated to relatively low precision. It is possible, but non-trivial, to base the criterion on the calculated accuracy of the integration scheme and finite difference approximation. On the other hand, the very facts that $\mathbf{J}$ is only approximate and co-linearity is hard to detect usually mean that the optimization scheme will not fail because of singularity problems: parameters will almost always appear 'nearly' co-linear, rather than exactly so. In this case inequality constraints can alleviate potential problems. For example, if a non-identifiable parameter has upper and lower bounds supplied then it will usually become fixed at one of these bounds quite rapidly, thus reducing potential numerical problems. However, models with unidentifiable parameters will present difficulties when it comes to the calculation of confidence intervals for parameters and may also reduce the reliability of smoothing parameter estimates.

Visualisation

Non-linear fitting methods tend to be as good as their starting values, and in many cases may display multiple local minima (for a good example of this see the Nicholson's blowfly example, below). For this reason it is a bad idea to fit models blind. It is vital to be able to see a fit progressing, in order to check that starting values are reasonable and that an obviously false local minimum has not been found. It is therefore important that fitting should be done in an environment where it is easy to modify and test starting values for parameters. In the case of unknown functions, this may mean relatively sophisticated programming to allow the user to manipulate starting functions, but within one of the many windowed environments now available, this is not overly difficult to achieve.

Examples

This section contains examples of the application of the methods described thus far, chosen to illustrate different aspects of the PSM approach. I start with a section that uses simulated data to examine the efficacy of the framework when fitting partially specified models, with multiple unknown functions, using a variety of modelling formalisms. A real example is then given, applying the method to modelling of laboratory *Daphnia* cultures before examining model comparison by statistical hypothesis testing using fully and partially specified models of Nicolson's blowfly data. The simplest model examples are in the blowfly section. Discrete matrix models, partial, ordinary and delay differential equation models are covered in the first example section. The models presented in all sections are relatively simple. Obviously this simplicity is somewhat at odds with the stated aim of reducing biologically spurious assumption in the modelling

process, but the hope is to be able to illustrate the approach without becoming swamped in a mass of modelling detail.

Marine Copepods: a comparative example with simulated data

Consider the problem of inferring underlying mortality and birth rates in a structured copepod population. This problem has produced more methods than there are data to fit (Asknes *et al.*, 1997); a fact for which the current author is as responsible as anyone. Copepods develop through a number of sequential life history stages that can be identified in population samples. If a model can be developed to predict the population dynamics of each stage from knowledge of the birth rate to the population and the death rates afflicting each stage, then in theory it should be possible to infer these rates by fitting the model to data. This problem allows demonstration of the efficacy of the smoothing parameter estimation methodology and the use of these methods with models formulated as difference equations, ordinary differential equations, delay differential equations and even partial differential equations. It also permits exploration of the sensitivity of inferred functions to changes in modelling formalism.

Consider an 11 stage population. Suppose that the population of stage i at time t is $n_i(t)$, that $d_i(t)$ is the *per capita* death rate in the stage, $R_1(t)$ is the recruitment (birth) rate into the first stage and τ_i is the mean duration of the i^{th} stage. Also define the subsidiary variables $\gamma_i = 1/\tau_i$ and the survival rate from time a to time b: $S_i(a,b) = exp(-\int_a^b d_i(x)dx)$. There are several competing structured population models that might be used for the stage populations, for example:

1. A matrix population model of the type advocated by Caswell (1989):

$$\mathbf{n}(t+1) = \mathbf{D}(t)\mathbf{n}(t) + \mathbf{r}$$

where $\mathbf{n}$ is the vector of stage populations, $\mathbf{r} = (R_1(t), 0, 0, \ldots, 0)^T$ and all elements of $\mathbf{D}(t)$ are zero except $D_{ii}(t) = S_i(t, t+1)(1-\gamma_i)$ for $i = 1, \ldots, 11$ and $D_{i,i-1} = S_{i-1}(t, t+1)(\gamma_{i-1})$ for $i = 2, \ldots 11$. The interested reader should consult Caswell (1989) for justification of this model structure. Clearly this model is a system of difference equations and hence one of the general class covered by this paper.

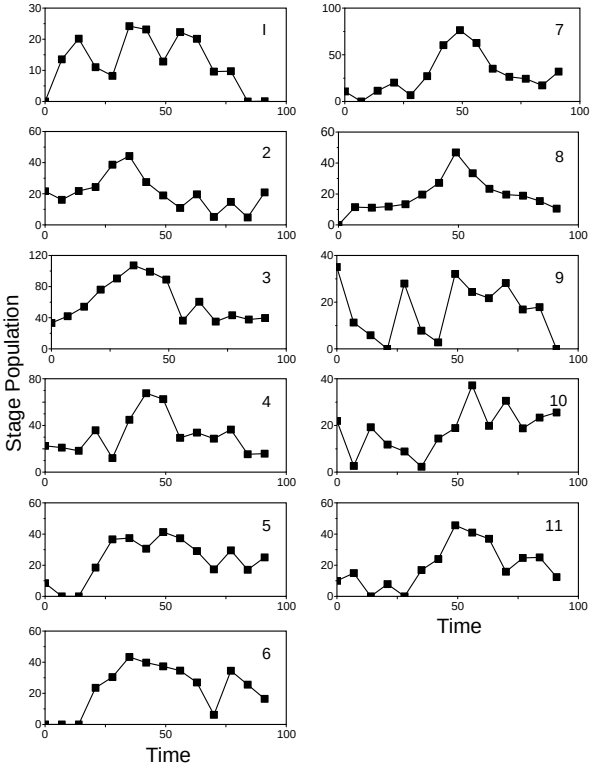

Figure 2: Typical replicate copepod data set simulated from the delay differential equation based structured population model given in the section *Marine Copepods: a comparative example with simulated data*. Stage durations used were: 0.75, 1.4, 4.55, 2.8, 2.5, 1.7, 3.5, 3.1, 3.2, 3.7 and 4.7. The functions used to simulate were a Gaussian plus a constant for recruitment, a decaying exponential for δ_1 and a constant for δ_2 (see figure 3).

2. A system of ordinary differential equations:

$$\frac{\mathrm{d}n_i}{\mathrm{d}t} = r_i(t) - \gamma_i n_i - d_i(t)n_i \quad i = 1, \ldots, 11$$

where $r_1(t) = R_1(t)$ and $r_i(t) = \gamma_{i-1}n_{i-1}(t)$ for $i = 2, \ldots, 11$. This model describes a rather extreme form of distributed stage durations in which individual stage durations are random variables following a negative exponential distribution.

3. A system of delay differential equations:

$$\frac{\mathrm{d}n_i}{\mathrm{d}t} = r_i(t) - r_i(t-\tau_i)S_i(t-\tau_i, t) - d_i(t)n_i \qquad i = 1, \ldots, 11$$

where $r_1(t) = R_1(t)$ and $r_{i+1}(t) = r_i(t - \tau_i)S_i(t - \tau_i, t)$. This system of equations describes the other extreme of stage duration

distribution, in which all individuals take exactly τ_i days to pass through a stage. See Gurney and Nisbet (1998) for a clear exposition of this sort of model, or see Gurney *et al.* (1983).

4. The McKendrick equation for an age structured population:

$$\frac{\partial f}{\partial t} + \frac{\partial f}{\partial a} + d(a,t)f = 0$$

where $f(a,t)$ is the population (or expected population) per unit age interval at age a and time t and $n_i(t)$ is given by the integral of f over the age range covered by stage i at time t. For this model $f(0,t) = R_1(t)$. Strictly this model is not in the class covered in this paper, but its numerical solution by the 'method of lines' (see e.g. Al-Rabeh, 1992) yields a set of ordinary differential equations which is in the class of models considered here. Discretizing in the age direction yields a system of equations (in an obvious notation) of the form:

$$\frac{\mathrm{d}f_j}{\mathrm{d}t} = -\frac{f_{j+1} - f_{j-1}}{2\Delta a} - d_i(j\Delta a, t)f_j(t) \qquad j = 1, 2, \ldots$$

where $f_0(t) = R_1(t)$ and the equation at the upper age boundary employs a backwards difference rather than a centred difference. $n_i(t)$'s are obtained by appropriate summations over the $f_j(t)$'s. For the results given here $\Delta a = 0.05$.

In all cases I set up the model initial conditions to be consistent with the initial recruitment and mortality rates, assuming that the population was at equilibrium (subject to those rates) prior to the initial time. I have kept these models very simple: clearly a realistic model would involve a more sensible distribution of stage durations - something which is easy to achieve within the d.d.e. framework (see Blythe *et al.* 1984). Note that all the models have been defined to be as similar as possible given their different structures. None the less models 1 and 2 are fairly closely related as are models 3 and 4, but 3 and 4 differ quite markedly from 1 and 2.

In some circumstances it may be reasonable to treat the stage durations as known, the birthrate to the first stage as an unknown function, the death rate in the first 6 stages (nauplii) as an unknown function of time ($\delta_1(t)$, say) and the death rate in the remaining 5 stages as a different unknown function of time($\delta_2(t)$, say). This is because

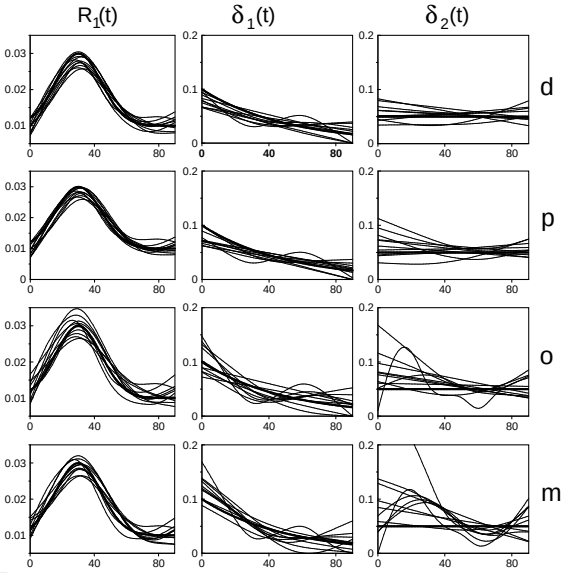

Figure 3: Reconstructed functions from 11 stage simulated copepod population data as described in figure 2 and *Marine Copepods: a comparative example with simulated data*. The column headings relate to the functions being reconstructed, while the row letters refer to the model used to fit the data: d - delay differential equations; p - partial differential equations; o - ordinary differential equations; m - discrete time matrix model. The heavy line in each panel is the true function used in simulation. The 10 fine lines in each panel are the replicate reconstructions (same true functions for each, but different sampling errors).

of substantial physiological and behavioural differences between the naupliar and copepodite stages. To demonstrate the smoothing parameter selection method I simulated an 11 stage population from a fully specified version of model 3 above. The simulated population was sampled (without error) and the samples were then perturbed with additive normal random errors (standard deviation 8). A typical resulting data set is shown in figure 2. I generated 10 replicates in total.

I then fitted partially specified models with 3 u.f.s to the resulting data treating stage durations as known. In this case the objective function consists of the sum of a weighted least squares term and three wiggliness penalties, each with its own smoothing parameter. The objective was subject to linear constraints imposing the condition that the three unknown functions must be positive. Each simulated data set was fitted with the 4 different partially specified models obtained by using the 4 different modelling approaches given above. Figure 3 shows the reconstructed rates superimposed on

the true rates used in simulation, for each model for each of the 10 replicates. All the models were fitted using iterative least squares with $\boldsymbol{\lambda}$ selection by GCV. A maximum of 10 degrees of freedom was allowed for each unknown function (this restriction is purely pragmatic - the number is higher than the effective degrees of freedom selected by GCV, but low enough that the computations are still reasonably quick, since there are only 10 parameters per unknown function).

Figure 3 illustrates several points. The recruitment function is well reconstructed in all cases, and the first mortality function is also reasonable. The final mortality function fares less well, particularly using the two models which have quite different assumptions to the model used for simulations. However, it is worth noting that the confidence intervals for this last function will be correspondingly wide: since the penalized likelihood approach applies a zero penalty to a straight line the model fitting will fit a line with a slope no matter how little statistical significance can be attached to that slope: this problem is not fundamental, one could examine the GCV score of a model with and without a slope, and select between them on this basis but I have not done so in this example. The tendency to raised mortality rates in the first half of the time range for δ_2 under models 1 (o) and 2 (m) is directly attributable to the fact that both these models allow some individuals very rapid transfer through the stages, in a way that the model used for simulation does not (indeed the hint of this effect in the partial differential equation model (4) is probably the result of numerical diffusion allowing a similar phenomenon). The general comparability of the reconstructed functions under substantially different models is encouraging: although the quantitative results depend on the formalism used for demographic book-keeping, the qualitative results do not seem greatly sensitive in this case and reconstructions are all quite close to the truth.

This example also demonstrates the effectiveness of the smoothing parameter selection method, and one of the important advantages of partially specified models. All the fitted functions have far fewer effective degrees of freedom than the maximum allowed in the fitting (10 in this example) and the fitted functions are mostly reasonable reconstructions of the truth. The importance of efficiency in the smoothing parameter selection is emphasized by the partial differential equation model - solution of this involved in excess of 600 coupled ordinary differential equations, yet fitting and model selection was completed in 10-20 minutes of computer time on a low specification Pentium II 400 Mhz PC. Notice as well that reconstructions are poorer where the most directly influential population data is low so that the signal to noise ratio is poor: the late parts of δ_1 and the early parts of δ_2. Confidence intervals estimated for these rates will reflect this: indicating clearly where the inferred knowledge of rates is poor. This contrasts with the situation that pertains for a fully specified model, where the convenient fiction that a good parameter sparse model is known tends to lead to much narrower confidence limits than can really be justified by data or knowledge - this confidence being based almost entirely on extrapolation of the fitted model from data rich periods and stages, to data poor portions of the data set.

Daphnia

The examples presented in the previous section used simulated data to compare structurally simple partially specified models constructed using a number of alternative formalisms. I now turn to an example using a more complicated partially specified structured model of some *Daphnia* population data kindly supplied by E. McCauley. The data consist of time series of adult and juvenile populations from a laboratory culture. The data were modelled using the moderately detailed structured population model given in table 2, which was supplied by R.M. Nisbet and is a modified version of the model of McCauley *et al.* (1996), which should be referred to for a detailed explanation and justification. Whilst most of the parameters of this model had been independently obtained (see McCauley *et al.* 1996), α and τ_A, were left as free parameters. To simplify presentation, the model used here assumes that the rate of food supply to the experimental system was kept constant (although in reality it was introduced in frequent pulses). There was some evidence that food quality had been changing throughout the experiment, although no detailed information was available about how it had changed. The aim was to produce a model in which food quality was described by an unknown, non-negative, function of time. In this way it could be ascertained how much of the model mismatch could be explained by food quality variation and the nature of any food quality variation could be examined. Food was provided to the population 3 times a week, usually on the basis of 2,2 and 3 day intervals. This has the potential to drive perceived food quality on a weekly

State variables	
Meaning	Equation
Juveniles	$\mathrm{d}J/\mathrm{d}t = R_J(t) - M_J(t) - \delta_J(t)J(t)$
Adults	$\mathrm{d}A/\mathrm{d}t = M_J(t) - M_J(t-\tau_A)P_A - \delta_A(t)A(t)$
Juvenile Survival	$\mathrm{d}P_J/\mathrm{d}t = \begin{cases} -P_J\delta_J & t < \tau_J \\ P_J(\delta_J(t-\tau_J)h(t)/h(t-\tau_J) - \delta_J(t)) & t \geq \tau_J \end{cases}$
Adult Survival	$\mathrm{d}P_A/\mathrm{d}t = \begin{cases} -P_A\delta_A & t < \tau_A \\ P_A(\delta_A(t-\tau_A) - \delta_A(t)) & t \geq \tau_A \end{cases}$
Juvenile duration	$\frac{\mathrm{d}\tau_J}{\mathrm{d}t} = \begin{cases} 1 - h/h_m & t < \tau_J \\ 1 - h(t)/h(t-\tau_J) & t \geq \tau_J \end{cases}$
Auxiliary functions	
Meaning	Equation
Juvenile ration	$\rho_J(J,A,t) = \frac{F(t)V}{\tau_I(J+\alpha A)}$
Adult ration	$\rho_A(J,A,t) = \frac{\alpha F(t)V}{\tau_I(J+\alpha A)}$
Initial inoculum	$i(t) = r_0 e^{-t}$
Juvenile recruitment	$R_J = i(t) + \rho_A \frac{N_A}{\rho_{A0}}$
Juvenile development	$h(J,A,t) = max\left(\frac{h_m \rho_J}{\rho_J + \rho_{J0}}, h_{min}\right)$
Adult death rate	$\delta_A = \delta_{Am} e^{-\rho_A/\rho_{A1}}$
Juvenile death rate	$\delta_J = \delta_{Jm} e^{-\rho_J/\rho_{J1}}$
Juvenile maturation	$M_J(t) = \begin{cases} 0 & t < \tau_J \\ \frac{R_J(t-\tau_J)P_J h(t)}{h(t-\tau)} & t \geq \tau_J \end{cases}$

Other definitions			
Food quality	$F(t)$	*Maximum adult death rate*	δ_{Am}
Microcosm volume	V	*Maximum Juvenile death rate*	$\delta_{Jm}V$
Transfer interval	τ_I	*Ration for* $1/e$ *adult mortality cut*	ρ_{A1}
Initial inoculum	r_0	*Ration for* $1/e$ *juvenile mortality cut*	ρ_{J1}
Minimum development rate	h_{min}	*Juvenile ration halving development*	ρ_{J0}
Maximum development rate	h_m	*Adult food to egg ratio*	ρ_{A0}
Adult Longevity	τ_A	*Adult to juvenile food ratio*	α

Table 2: *Daphnia* model definitions.

basis - so the unknown function in the model was given a maximum number of degrees of freedom of 60, in order to be able to accommodate a weekly cycle if necessary.

w_i's for the adults were set to 9 times those for the juveniles (this weighting was selected to give the least bad residual plots) and the model was fitted by constrained minimisation of a weighted least squares objective penalized by a smoothness constraint on the food quality function. This was done by iterative least squares. The smoothing parameter controlling the trade-off between fidelity to the population data and smoothness of the food quality function was chosen by GCV, as outlined in the methods section.

Figures 4a and 4b show the best fitting model and the original data plotted together. Figure 4c shows the best fitting $F(t)$ with asymptotic confidence bands (which should be treated with caution, given the residuals). Notice that the model is incapable of reproducing the initial transient in the data, and that later fluctuations seem to be matched by allowing $F(t)$ to exhibit considerable temporal variability.

This example illustrates several points. Firstly, it demonstrates that the fitting framework described can be used to fit real data with multiple time series and irregular sampling and that the fact that the model is moderately structurally complex does not cause difficulty. Secondly, it demonstrates an important point about partially specified models: the model was allowed a maximum of 63 free parameters and still failed to fit the data overly well. The well worn adage that 'if you give me four free parameters I will draw you an elephant, and given 5 I can make it wiggle its tail', only contains any truth if you have freedom to choose the model which will incorporate these 5 parameters. In this case the model structure imposes very strong constraints on what the model population can do, irrespective

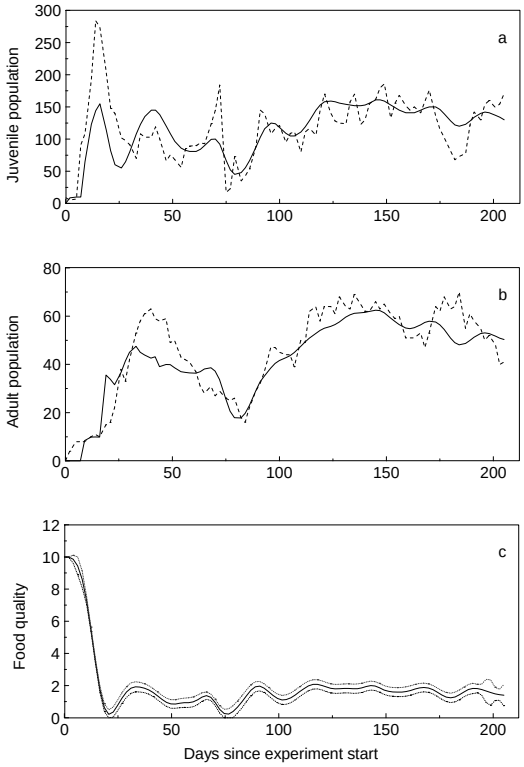

Figure 4: Partially specified model fit to *Daphnia* population data. In (a) and (b) the dashed line joins the population data and the solid line is the model fit. (a) shows the fit to the juvenile population. (b) shows the fit to the adult population. (c) is the fitted food function (a function of time), the solid line is the best estimate and the dashed line an asymptotic 95% confidence interval, calculated from the quadratic approximation to the model fitting objective.

of the values taken by its parameters. The third related point concerns smoothing parameter selection: the fitted model in fact suggests a relatively slow rate of change for food quality, so that the final food quality function has far fewer effective degrees of freedom than the maximum possible number of 60: the GCV criterion has in effect decided that no real improvement would be achieved by allowing the food quality to vary more rapidly. This illustrates that GCV itself avoids over-fitting. Indeed, if food quality is allowed a full 60 degrees of freedom (by setting the smoothing parameter to zero), then it varies widely and rapidly, but the 'improvement' in fit is marginal.

Nicholson's Blowflies I: a fully specified example

In order to introduce simple statistical hypothesis testing with population models, consider comparing two alternative hypotheses about the mechanisms driving the dynamics of Nicholson's famous blowflies (1954a,b) using fully specified models (Kendall *et al.* 1999 discuss this problem in some detail). The data to be fitted are numbers of adult blowflies *Lucila cuprina* in laboratory cultures. In the cultures examined here the adults were supplied with protein at a fixed and limiting rate, but all other resources were supplied in abundance. As can be seen from figure 5, the population cycled.

The first hypothesis is that the regulatory process driving the blowfly cycles is competition amongst adults affecting adult fecundity, so that (given suitable starting conditions) the adult population can be described by the delay differential equation:

$$\frac{\mathrm{d}A}{\mathrm{d}t} = SA(t-\tau)e^{-A(t-\tau)/A_0} - \delta A(t)$$

where S is a compound variable made up of adult fecundity multiplied by juvenile survival, τ is the delay from egg laying to adulthood, A_0 is the reciprocal of the exponential decay rate of fecundity with adult population, and δ is the adult death rate. This model was first proposed by Gurney *et al.* (1980) at the same time as a structurally equivalent model was produced by Readshaw and Cuff (1980)

The second hypothesis is that regulation is through the effect of juvenile competition on adult size and hence fecundity: this mechanism can give rise to the following formulation (after Gurney and Nisbet, 1985):

$$\begin{aligned}
\frac{\mathrm{d}J}{\mathrm{d}t} &= B(t) - \Delta B(t-\tau)e^{-\tau\delta_J} - \delta_J J \\
\frac{\mathrm{d}A}{\mathrm{d}t} &= \Delta B(t-\tau)e^{-\tau\delta_J} - \delta_A A \\
\frac{\mathrm{d}B}{\mathrm{d}t} &= \Delta B(t-\tau)e^{-\tau\delta_J} W - \delta_A B \\
\frac{\mathrm{d}W}{\mathrm{d}t} &= g(t) - \Delta g(t-\tau) \\
g(t) &= \frac{g_m}{1+J/J_0}
\end{aligned}$$

where $\Delta(t)$ is zero for $t < \tau$ and one otherwise. J and A are adult and juvenile populations, B is birth rate, W is weight at maturation to adulthood. τ is development time, δ_A is adult death rate, g_m is a parameter made up from maximum juvenile growth rate and adult fecundity per unit weight, J_0 controls the rate of decrease growth rate with juvenile population, μ is maintenance cost and δ_J juvenile death rate. Given the experimental set up, this second hypothesis is almost certainly incorrect,

but in the current context this is useful, since we can be fairly sure what the correct answer should be.

Both models were fitted to Nicholson's data by using constrained quasi-Newton to minimise the sum of squares of differences between model trajectories and adult population counts; parameters were constrained to be positive, but of course there are no wiggliness penalties in the fitting objective. The best fits obtained for the two models are shown in figure 5. The r^2 values are 0.69 and 0.60 for the adult competition model and juvenile competition models respectively. It might be useful to access the statistical significance of these results, but statistical comparison of the models is a non-standard problem.

To motivate the approach taken here it is worth considering what statistical hypothesis testing does, in quite general terms. Statistical hypothesis testing amounts to comparing two models of how a set of data has been generated. The model constituting the null hypothesis is a restricted version of the model constituting the alternative hypothesis: we are comparing a simple model and a more complicated model of data. Because the complicated model is an extended version of the simple model it will always be able to fit any set of data at least as well as the simple model. Hypothesis testing asks the question: *if the simple model is true, what distribution should I expect for the improvement in fit of the complicated model relative to the simple model?* All that then remains is to decide whether the observed difference in fit is consistent with this distribution, and hence with the null hypothesis.

The general method of hypothesis testing is well covered in Silvey (1975). Hypothesis testing using this approach in combination with standard distributional results is familiar in the context of analysis of variance or likelihood ratio tests (in generalized linear modelling for example, McCullagh and Nelder, 1989). If the assumptions required to use standard distributional results are not met then an obvious computer intensive approach can be taken: generate replicate data from the best fit of the simpler model, and build up an empirical distribution for the difference in fit between the two models, by fitting both to each of these replicates. Parametric or semi-parametric bootstrapping is the obvious way to achieve this. A detailed discussion of this approach in the context of linear regression can be found in Davison and Hinkley (1997 section 6.3.2).

The current example does not quite fit into the general hypothesis testing framework because the worse fitting model is not obviously a restricted version of the better fitting model. Hence we don't know that the better fitting model would fit more closely even if the worse fitting model were true and this precludes the use of general methods such as the likelihood ratio test for comparing these models (again Silvey, 1975, provides the necessary background for a mathematical appreciation of this point). Never the less, the general question of what sort of improvement in fit of the better fitting model over the worse fitting model is expected if the worse fitting model is true retains exactly the same meaning in the current non-nested case that it holds in the case of nested models. By the same token the answer holds the same scientific implications in this case as it holds in the case of nested models. Appreciation of this latter point suggests that it is worth using a computer intensive approach to obtain an approximate *p-value* to attach to the improvement in fit of the adult competition model.

I therefore simulated data by assuming the best fit juvenile growth limitation model to be true. Replicate data sets were generated by sampling with replacement from the residuals of the best fit juvenile growth model and adding these resampled residuals to the model predictions of the population at each sampling time. Both models were fitted to each replicate and their difference in goodness of fit was assessed. By generating a large number of replicate datasets under the null hypothesis (that the juvenile growth model generated the data), it is possible to obtain an approximate distribution of the difference in fit between the two models *assuming that the null hypothesis is true*. The proportion of differences in fit that are at least as large as the difference observed when fitting the original data, provides a *p-value* to associate with testing the null hypothesis that the juvenile growth model generated the data, against the alternative that the adult competition model did so. Figure 5c gives a histogram of the simulated distribution of badness of fit under the null (measured by residual sum of squares), and the true difference.

The adult competition model clearly fits better than the juvenile growth model, and this would tend to lend support to the adult competition hypothesis as opposed to the juvenile competition hypothesis. Of course this support is equivocal: given the simple structure of the models used and the fairly arbitrary specification of the key functional relationships within both models it is not clear to what extent the test results are related to the ability of the models' incidental assumptions to match reality, rather than relating to the core ecological

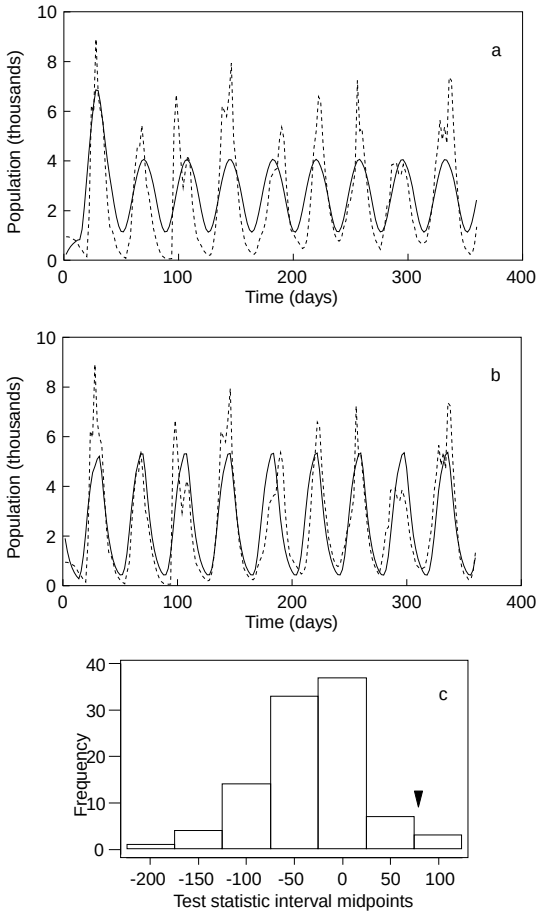

Figure 5: Comparison of fully specified model fits to Nicholson's Blowfly data. (a) The best fit of the adult fecundity model (solid line) to the experimental data (joined by dashed line). (b) The best fit of the juvenile competition model (solid line) to the same data (again joined by a dashed line). (c) A histogram of the difference in mean square deviations of models from data, for each of 99 replicate non-parametrically bootstrapped datasets under the null hypothesis that the adult fecundity model is as good a description of the underlying dynamics as the juvenile competition model. The triangle shows the difference in mean square errors for the original data.

assumptions underpinning the two models. This deficiency will be partially rectified in the next section. Note also that, as in most statistical hypothesis testing, the question posed is quite limited - attention has been restricted to just two alternative mechanisms *as embodied in two simple models* - hence it is only the *relative* merits of the two alternatives that are assessed. Furthermore the bootstrap analysis just described must be treated with caution, since the replicates under the null do not preserve the covariance structure in the residuals, and it is not clear under what circumstances this structure should be preserved (also note that the differences in goodness of fit is unlikely to be a pivotal quantity in the statistical sense). Hence a complementary model comparison should also be performed, by non-parametrically bootstrapping from the data, to obtain a bootstrap estimate of the distribution of goodness of fit difference between the two models. An example of this approach will be given in the next section, when partially specified models for the blowfly population are considered.

Of course, none of this hypothesis testing machinery is necessary if we merely seek to find which model is"best" with respect to criteria like AIC or GCV, and do not seek to attach a p-value to the choice. AIC in particular is quite often used for selecting between non-nested models, but the validity of the approach does not seem to be settled in the literature. Chapter 6 of Burnham and Anderson (1998) should probably be read critically before deciding to compare non-nested models in this way. Murata *et al.* (1994) give a clear statement of the concerns about the practice raised by a careful examination of the derivation of AIC. However, if an AIC model selection approach is taken then Burnham and Anderson (1998) suggest some interesting approaches for assigning confidence levels to the model choice.

This example demonstrates the utility of the fitting approach for fully specified models, which are, after all, only special cases of partially specified models. On the other hand the example can also be used to illustrates some of the difficulties with model fitting by trajectory matching. The fact that a number of cycles are to be matched, means that the objective function is almost certain to have several local minima. To see this, imagine doubling the frequency of the model cycles - every other cycle would still line up with a cycle in the real data - to move away from this situation is bound to involve increasing the objective function. One way around this difficulty, that works well in practice, is to supplement the fitting objective with some squared terms that penalise deviation of the model ACF from the data ACF at a number of lags. Giving sufficient weight to this part of the objective often results in rapid attainment of approximately the right cycle period, after which the extra terms can be removed. Kendall *et al.* (1999) present several interesting approaches to both model fitting and inference with respect to this example, including the use of stochastic population dynamic models.

Blowflies II: partially specified model comparison

Finally, consider the blowfly example again, but this time let the analysis proceed with partially, rather than fully, specified models. Consider the adult fecundity model. An appropriate partially specified statement of this model is:

$$\frac{\mathrm{d}n}{\mathrm{d}t} = \beta(n_{t-\tau})n_{t-\tau} - \delta(n_t)n_t$$

where β is a monotonically decreasing non-negative function of density and δ is a non-negative monotonically increasing function of density.

The partially specified version of the juvenile growth limitation model is a little more complicated:

$$\begin{aligned}
\frac{\mathrm{d}J}{\mathrm{d}t} &= B_t - \Delta B_{t-\tau} \\
\frac{\mathrm{d}A}{\mathrm{d}t} &= \Delta B_{t-\tau} - f(B_t)A_t \\
\frac{\mathrm{d}B}{\mathrm{d}t} &= \Delta B_{t-\tau}W_t - f(B_t)W_t \\
\frac{\mathrm{d}W}{\mathrm{d}t} &= g(J_t) - \Delta g(J_{t-\tau})
\end{aligned}$$

Now g, the juvenile growth rate, can be expected to decline monotonically with juvenile density (while remaining non-negative), and f the adult *per capita* death rate will be a monotonically increasing function of B (which is proportional to adult biomass). Δ again takes the value 1 when $t > \tau$ and 0 otherwise. For practical fitting the unknown functions in each of these models were each allowed at most 10 degrees of freedom (again, this is well above the degrees of freedom selected by GCV, while 10 parameters per unknown function still allows relatively quick computation). The function domains were selected with some experimentation to make sure that the they were matched to (but were a bit larger than) the domains suggested by the best fit models.

When fit to Nicholson's data these partially specified models fit slightly better than their fully specified counterparts. r^2 values show only slight improvement to 0.71 and 0.61, for the adult fecundity and juvenile competition versions, respectively, but the juvenile competition model now has a best fit trajectory that is not obviously worse than the adult fecundity model.

Note that although the models are both capable of producing the double humped peak evident in parts of Nicholson's data, the best fitting models, do not display such dynamics. Figures 6a and 6b show the best fits of the 2 models and the best fit forms for f and g in both models. τ and the initial adult population were also fitted as a free parameters for both models, and the juvenile growth limitation model had an additional free parameter: the initial adult fecundity.

The significance of the goodness of fit improvement of one model over the other was assessed by non-parameteric bootstrapping, in order to obtain a 98% confidence interval for the difference in model goodness of fit (as measured by difference in residual sum of squares divided by 1 million). In all 99 replicates the adult fecundity model fitted better than the juvenile growth limitation model: so if bootstrap percentile confidence intervals (Efron and Tibshirani, 1993) are used the 98% C.I. for the differences in model fits does not include zero. If basic bootstrap confidence intervals (Davison and Hinkley, 1997) are used then the 98% confidence interval includes zero, but the 95% C.I. does not. It is reasonable to conclude that the adult fecundity model is better at the 5% level.

Note that interpretation of statistical model comparison is complicated by the fact that it is not obvious to what extent the models can be treated as 'nested'. To see the problem, consider nested models in linear regression. Using percentile confidence intervals, the bootstrap analysis just performed would never reject the model $E(y) = a + bx + cx^2$ in favour of $E(y) = a + bx$, no matter how spurious the quadratic term. Using basic bootstrap intervals the quadratic model *could* be rejected, but there is no reason to suppose that the *p-value* would relate to the probability of rejecting a correct null hypothesis! These considerations would tend to favour use of the bootstrap analysis employed for the fully specified blowfly models: but the error structure clearly violates the assumptions underlying that approach suggesting that it should not be used alone. Despite the problems the combination of the 'bootstrap under the null' and non-parametric bootstrap provides a useful tool.

As well as illustrating another method of model comparison, this example demonstrates the utility of partially specified models, even when a fully specified model is available. In this case it was possible to check whether the functions chosen for use in the fully specified models produced spurious constraints on the population dynamics that restricted one or both of the models in a manner that had nothing to do with the biology that the models were attempting to describe. For the juvenile competition model there is some evidence for this being the case, since

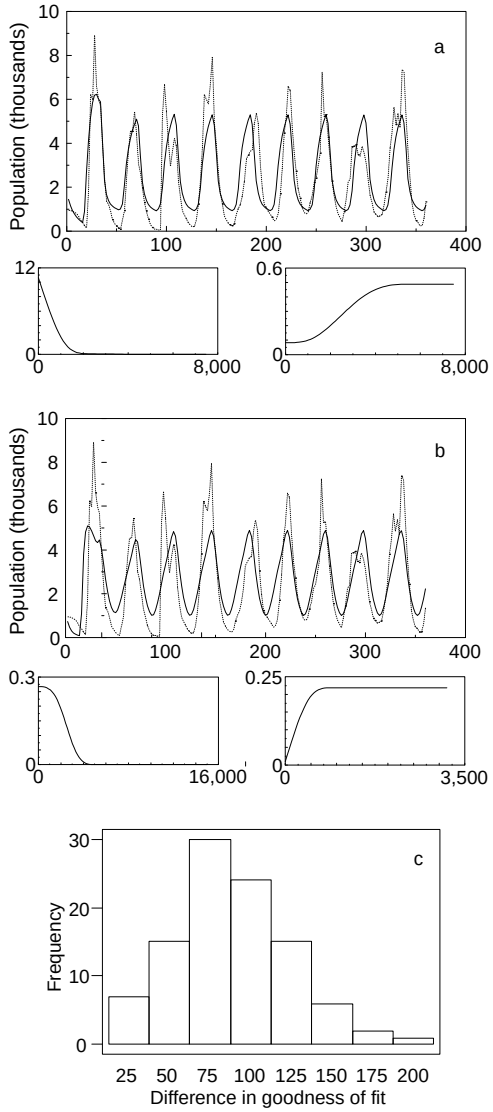

Figure 6: Partially specified blowfly models. (a) shows the best fitting population trajectory for the adult fecundity model ($r^2 = 0.71$), in the large panel: the dashed line joins the actual data, the solid line is model fit. The two small panels are the best fit adult fecundity as a function of adult density (left) and the best fit adult death rate as a function of adult density. (b) is similar to (a), but shows the best fit growth limitation model ($r^2 = 0.61$). The left small panel shows juvenile growth rate as a function of juvenile abundance, while the right small panel is adult *per capita* death rate as a function of total adult fecundity (a surrogate for biomass). (c) is a histogram of the difference in badness of fit between the two models in 99 non-parametric bootstrap re-samples. Badness of fit is measured by error sum of squares over 10^6: the range of bootstrapped values went from 15 to 197, the original value was 98.

the partially specified model yields dynamics which appear to be a more reasonable match than the fully specified equivalent. However, it still appears that the adult fecundity model is significantly better, so that confidence in this conclusion is strengthened. In short, within the limitations of the rather simplistic models employed for this example, the analysis based on partially specified models comes closer to testing for differences between competing mechanisms, rather than testing for differences between alternative sets of incidental assumptions.

Discussion

The approach described in this paper provides a way of formulating ecological models that precisely reflect the information about a system that the modeller wants to include, while avoiding assumptions that the modeller doesn't wish to include in the hope of reducing reliance on biologically spurious assumptions. More importantly it provides a practical way of using such models by providing fitting methods, as well as model selection and inference techniques. In the context of statistical estimation the benefits of an approach that avoids arbitrary parametric models are well known from work on GAMs and Spline models (Hastie and Tibshirani 1990, Wahba 1990) and the methods can be viewed as an extension of these techniques which introduces mechanism and non-linearity into the models employed. In ecological contexts the benefits of partial specification have been recognized and clearly demonstrated in a number of special cases (Bjørnstad *et al.* 1999, 1998, Ellner *et al.* 1998, 1997, Wood 1997, 1994, Ohman and Wood, 1996 and Wood and Nisbet 1991): what this paper does is to move from special cases to a general framework, as well as solving the crucial problem of choosing the optimal model complexity, without which the exercise would merely have replaced one set of arbitrary model assumptions with another (by complexity I mean the complexity of the component functions of the model, as controlled by their associated smoothing parameters $\boldsymbol{\lambda}$).

The methods described have several practical benefits for fully and partially specified model fitting and comparison: when used for 'trajectory matching' the objective function makes good use of information over all available timescales, which has distinct advantages for short noisy datasets; missing values, uneven sampling and unobserved state variables present no technical difficulties; the efficient, reliable and robust fitting methods make it possible to employ modern computer intensive inference

methods with many models, and to have some confidence in the notion that differences in model fit reflect real differences in model performances, rather than differences in how much help was given to the fitting routine.

Of course the approach is far from perfect. Production of a method that allows formal inference and model selection by a means other than armwaving has necessitated some restrictive choices. I have considered a class of models that is quite specific to ecology, and have not considered stochastic dynamics. The latter omission can lead to difficulty when analysing long time-series for cyclic systems with considerable process noise. In this circumstance the process noise may produce phase drift in the dynamics, which a deterministic model can not match. This implies that some care will be required in selecting the 'data' to be matched in the objective function employed here: if the model is not designed to exactly mimic phase characteristics of the modelled system dynamics, then the 'data' to be fitted needs to be relatively phase insensitive. The obvious way to approach this problem is to transform cyclic timeseries data to obtain, for example, a data vector to be matched which consists of the mean, sample variance, and ACF or PACF of the original data. The difficulty with this approach is lack of independence of the resulting data, which will strip $\boldsymbol{\lambda}$ selection methods of theoretical justification.

Chaotic systems are also problematic: it's fairly obvious that they should never be fit by trajectory matching - the possibility of entirely spurious fit to signal and noise is ever present, and yet any chaotic fit is contingent on the ecologically nonsensical notion that the system's parameters really did not change at all over the course of the data, and that there was absolutely no process noise. Happily the very fragility of a chaotic trajectory match means that it's very difficult to find one - in the chaotic regime the objective function becomes far too nightmarish a landscape for the methods employed here, designed as they are for relatively gentle slopes and rolling valleys. Pragmatically, this means that a trial step into the chaotic region of parameter space almost never leads to decrease in the objective function, and if the minimisation routine does get stuck there the problem is easily diagnosed. Even with non-chaotic systems the fitting objective will sometimes be too irregular for the methods used here: Appendix C describes an effective approach when local minima are relatively small scale nuisance features of the objective, but in more severe cases other techniques, such as simulated annealing (e.g. Press *et al.* 1992, Brooks and Morgan, 1994) may be the only way to make progress.

Discrete time cyclic systems can also display problems of their own: such systems do not necessarily yield smooth changes in frequency as parameters change... a feature which has the advantage of acting against phase drift in the data itself, but the disadvantage of promoting local minima in the fitting objective. Discrete models also display 'aliasing' effects when the cycle period is not an integer multiple of the model time-step: the resulting small scale irregularities can also cause local minima. Fortunately, both of these difficulties can often be overcome by the use of "bootstrap restarting" during model fitting as described in Appendix C.

In the light of figures 5 and 6 it is important not to overstate the difficulty of fitting cyclic data by the methods suggested here, and in any case most data to be fitted will not be long cyclic series. Never the less there are alternative model fitting methods that avoid what problems there are. One approach is to use auto-regressive methods (e.g. Caswell and Twombly 1989, Wood 1997). At its most general this simply means using the model to predict the next data point or two on the basis of current (and perhaps past) data points. The model parameters that do the best job at this are considered to fit the data best. Continuous models can be dealt with by the simple expedient of smoothing the data and seeing how well the model predicts the smoothed gradients when fed the smoothed data (see Ellner *et al.* 1997). A disadvantage of simple regression approaches is the need to observe all the state variables, although more sophisticated approaches avoid this by iterating the original model in order to statistically build up a dynamically equivalent model which predicts observed data solely from previously observed data (e.g. Ellner *et al.* 1998). A more fundamental problem with using regression type approaches with partially specified models is the difficulty in selecting the flexibility of model functions: it is not hard to think up *ad hoc* techniques, but an efficient objective method with a firm theoretical basis is not yet available. Also, although regression methods are a sensible way of dealing with process error, they fare badly in the face of sampling error by focusing on a feature of the data that is likely to have the lowest signal to sampling error ratio. To see this, imagine a population whose dynamics are a sine wave corrupted by sampling error of comparable magnitude to the wave amplitude. A regression method would clearly yield very

poor results in this case, while trajectory matching would give quite reasonable estimates. Inference is also complicated in the regression case, partly because both predictor and response variables are subject to measurement error.

An appealing extension of the partially specified modelling framework would be the direct inclusion of stochasticity into the modelled dynamics. Over short timescales one approach would be to include process errors directly. Many noise models can be characterised by a single scale parameter, so that it is possible to generate a set of zero mean pseudo-random variates from the noise model assuming one value for the variance, and simply scale these variates in order to obtain a set of variates consistent with the same model with a different variance. Interpolating these variates with a smooth curve, yields a realisation of a noise process that can be incorporated into a continuous model (S.P. Ellner, pers. comm.), and the variance of which can be controlled by a single scale parameter. By averaging over a fairly large number of replicates of such a process the expected population and other useful moments of the population's statistical distribution can be obtained, given any set of model parameters. Smoothness of the fitting objective is maintained by only changing the scale of the variates, rather than generating new variates for each new set of parameters in the fitting process. This smoothness should allow all the methods reported here to operate correctly.

A further option for modelling stochasticity in dynamics is simply to add a perturbation function of time to the model structure as an unknown to be estimated, and to treat this as a 'random effect' within the fitting framework. In some cases this results in a straightforward penalized regression problem in which the smoothing parameter to be estimated is proportional to the reciprocal of the variance of associated with the random function (further details of this approach will be published elsewhere).

One non-obvious use of the current framework is to simultaneously fit dynamics and separate parameterization data. This is usually easy to achieve: for example it is always possible to define a state variable that simply takes the value of a parameter of interest, and the deviation of this state variable from independent measurements of the state variable is then easily included in the fitting objective. More sophisticated generalisations of this approach would be useful. Raftery *et al.* (1995) have attempted just such a synthesis of multiple sources of parameterisation data and population dynamic data using simple deterministic demographic models for Bowhead whales and it provides clear benefits in the context of prediction.

But is model fitting really a good thing at all? A model that gets it right without fitting, purely on the basis of independent parameter estimates, is always impressive. By contrast fitting is often frowned upon, as if the validity of a model is undermined by it. Why is this? Partly it is because a good ecological model needs to be relatively robust to parameter variation to be taken seriously: if a model can only match reality for parameters lying within some very narrow band then it is probably wrong - the real quantities that the model parameters represent will have varied quite substantially over the time period of data collection. At the same time the parameter values required to make the model fit ought to agree with independent measurement. A model that fits data on the basis of independently measured parameter values obviously meets this last criterion, but must almost always have met the robustness criterion as well - otherwise it would be most unlikely that a good fit would have been obtained by simply using the independent point estimates for the parameters. So models that fit well without parameter adjustment are likely to be good, but this should not be taken to imply that a model is likely to be wrong if it does have to be fitted. Dynamics can alter substantially within the region of parameter space that is consistent with independent parameterisation data and to reject a model because it did the wrong thing at a single point somewhere in the middle of that region is a nonsense. To really test models, they should be fitted to data, while checking for agreement with independent parameter estimates and separately checking robustness. The methods presented here make this feasible, while markedly increasing the scope and quality of model fitting based inference in those applications which provoke no controversy, such as hypothesis testing and estimation. A Windows package implementing the methods is available free of charge from the author.

Acknowledgements

Roger Nisbet provided ideas and encouragement over the rather lengthy period that this work took, as well as test driving the code implementing the methods, and commenting on the manuscript. Bruce Kendall provided many helpful suggestions on the paper and discussions with him and Steve Ellner heavily influenced the discussion section.

One anonymous referee provided a large number of helpful suggestions which improved the introduction and the examples sections; a later referee and F.R. Adler suggested further very helpful improvements. Peter Turchin suggested matching ACFs. Interactions with members of the NCEAS complex population dynamics group have been very useful: Cheryl Briggs, Steve Ellner, Bruce Kendall, Ed McCauley, Bill Murdoch, Roger Nisbet and Peter Turchin. Ed McCauley and Roger Nisbet provided the *Daphnia* data and model. Joe Horwood provided the initial problem, and other encouragement. Thanks to all of them. Parts of this work were conducted as part of the Complex Population Dynamics Working Group at the National Center for Ecological Analysis and Synthesis at Santa Barbara, funded by NSF DEB-94-21535, and the work was started at the NERC Centre for Population Biology.

Literature Cited

Al-Rabeh, A. 1992. Towards a general integration algorithm for time dependent one-dimensional systems of parabolic partial-differential equations using the method of lines. Journal of Computational and Applied Mathematics **42(2)**: 187-198.

Akaike, H. 1973. Information theory as an extension of the maximum likelihood principle. Pages 267-281. *in* B.N. Petrov and F. Csaki editors. Second International Symposium on Information Theory. Akademiai Kiado, Budapest.

Asknes, D.L., C.B. Miller, M.D. Ohman and S.N. Wood. 1997. Estimation techniques used in studies of copepod population dynamics - A review of underlying assumptions. Sarsia **82**:279-296.

Bjørnstad, O.N., M. Begon, N.C. Stenseth, W. Falck, S. M. Sait and D.J. Thompson. 1998. Population dynamics of the Indian meal moth: demographic stochasticity and delayed regulatory mechanisms. Journal of Animal Ecology **67**: 110-126.

Bjørnstad, O.N., J.M. Fromentin, N.C. Stenseth, and J. Gjøsæter. 1999. A new test for density-dpendent survival: the case of coastal cod populations. Ecology **80(4)**:1278-1288.

Blythe, S.P., R.M. Nisbet and W.S.C. Gurney. 1984. The dynamics of population- models with distributed maturation periods. Theoretical Population Biology **25(3)**:289-311.

Brooks, S.P. and B.J.T. Morgan. 1994. Automatic starting point selection for function optimization. Statistics and Computing **4**: 173-177.

Burnham, K.P. and D.R. Anderson. 1998. Model Selection and Inference: A Practical Information-Theoretic Approach. Springer-Verlag, New York.

Caswell, H. 1989. Matrix Population Models. Sinauer, Sunderland. Mass.

Caswell, H. and S. Twombly. 1989. Estimation of stage-specific demographic parameters for zooplankton populations: methods based on stage-classified matrix projection models. Pages 94-107 *in* L. McDonald, B. Manly, J. Lockwood and J. Logan, editors. Estimation and analysis of Insect Populations Springer-Verlag.

Craven, P. and G. Wahba. 1979. Smoothing noisy data with spline functions. Numerische Mathematik **31**: 377-403.

Davison, A.C. and D.V. Hinkley. 1997. Bootstrap Methods and their Application. Cambridge University Press, Cambridge.

Efron, B. and Tibshirani. 1993. An Introduction to the Bootstrap. Chapman and Hall, London.

Ellner, S.P., B.E. Kendall, S.N. Wood, E. McCauley, C.J. Briggs. 1997. Inferring mechanism from time-series data: Delay-differential equations. Physica D **110**:182-194.

Ellner, S.P., B.A. Bailey, G.V. Bobashev, A.R. Gallant, B.T. Grenfell and D.W. Nychka. 1998. Noise and nonlinearity in measles epidemics: combining mechanistic and statistical approaches to population modelling. American Naturalist **151(5)**:425-440.

Gill, P.E. , G.H. Golub, W. Murray and M.A. Saunders. 1974. Methods for Modifying Matrix Factorizations. Mathematics of Computation **28**:505-535.

Gill P.E., W. Murray and M.H. Wright. 1981. Practical Optimization. Academic Press, London.

Green, P.J. and B.W. Silverman. 1994. Nonparametric Regression and Generalized Linear Models. Chapman and Hall, London.

Gu. C. and G. Wahba. 1991. Minimizing GCV/GML scores with multiple smoothing parameters via the newton method. SIAM Journal on Scientific and Statistical Computation **12**: 383-398.

Gurney, W.S.C., R.M. Nisbet and S.P. Blythe. 1980. Nicholson's blowflies revisited. Nature **287**:17-21.

Gurney, W.S.C., R.M. Nisbet and J.H. Lawton. 1983. The Systematic formulation of tractable single species population models incorporating age structure. Journal of Animal Ecology **52**:479-495.

Gurney, W.S.C. and R.M. Nisbet. 1985. Fluctuation periodicity, generation separation, and the expression of larval competition. Theoretical Population Biology **28**:150-180.

Gurney, W.S.C. R.M. Nisbet and S.P. Blythe. 1986. The systematic formulation of models of stage structured populations. Pages 474-494 *in* J.A.J. Metz and O. Diekmann, editors. The Dynamics of Physiologically Structured Populations. Springer Verlag Berlin.

Gurney, W.S.C. and R.M. Nisbet. 1998. Ecological Dynamics. Oxford University Press.

Hairer, E., S.P. Norsett and G. Wanner. 1987. Solving Ordinary Differential Equations I. Springer-Verlag, Berlin.

Hastie, T.J. and R.J. Tibshirani. 1990. Generalized Additive Models. Chapman and Hall, London.

Higham, D.J. 1993a. Error control for initial value problems with discontinuities and delays. Applied Numerical Analysis **12**: 315-330.

———. 1993b. The tolerance proportionality of adaptive ODE solvers. Journal of Computational and Applied Mathematics **45**:227-236.

Kendall, B.E., C.J. Briggs, W.W. Murdoch, P. Turchin, S.P. Ellner, E. McCauley, R.M. Nisbet and S.N. Wood. 1999. Why do populations cycle? A synthesis of statistical and mechanistic modeling approaches. Ecology **80(6)**:1789-1805.

May, R.M. 1974. Stability and Complexity in Model Ecosystems. Princeton University Press.

MacDonald, N. 1978. Time lags in biological models. Springer-Verlag, Berlin.

———. 1989. Biological Delay Systems. Cambridge University Press.

McCauley, E., R.M. Nisbet, A.M. deRoos, W.W. Murdoch and W.S.C. Gurney. 1996. Structured Population Models of Herbivorous Zooplankton. Ecological Monographs **66**:479-501.

McCullagh, P. and J.A. Nelder. 1989. Generalized Linear Models. Chapman and Hall, London.

Meyer, S.L. 1975. Data analysis for Scientists and Engineers. John wiley and Sons, New York.

Murata, N., S. Yoshizawa and S. Amari. 1994. Network Information Criterion - Determining the Number of Hidden Units for an Artificial Neural Network Model. IEEE Transactions on Neural Networks **5(6)**:865-871.

Nicholson, A.J. 1954a. Compensatory reactions of populations to stresses and their evolutionary significance. Australian Journal of Zoology **2**: 1-8.

———. 1954b. An outline of the dynamics of animal populations. Australian Journal of Zoology **2**: 9-65.

Nisbet, R.M. and W.S.C. Gurney. 1983. The systematic formulation of population models for insects with dynamically varying instar durations. Theoretical Population Biology **23**:114-135.

Nisbet R.M. 1997. Delay-Differential Equations for Structured Populations. Pages 89-118 *in* S. Tuljapurkar and H. Caswell, editors. Structured-Population Models in Marine, Terrestrial, and Freshwater Systems. Chapman & Hall, New York.

Ohman, M.D. and S.N. Wood. 1996. Mortality estimation for planktonic copepods: *Pseudocalanus newmani* in a temperate fjord. Limnology and Oceanography **41(1):**126-135.

Paul, C.A.H. 1992. Developing a delay differential equation solver. Applied Numerical Mathematics **9**:403-414.

Press, W.H., S.A. Teukolsky, W.T. Vetterling and B.P. Flannery. 1992. Numerical Recipes in C. Cambridge University Press, Cambridge.

Raftery, A.E., G.H. Givens and J.E. Zeh. 1995. Inference from a Deterministic Population Dynamics Model for Bowhead Whales. Journal of the American Statistical Association **90**:402-416.

Readshaw, J.L. and W.R. Cuff. 1980. A model for Nicholson's blowfly cycles and its relevance to predation theory. Journal of Animal Ecology **49**:105-1010.

Silvey, S.D. 1975. Statistical Inference. Chapman and Hall, London.

Villalobos, M. and G. Wahba. 1987. Inequality constrained multivariate smoothing splines with

application to the estimation of posterior probabilities. Journal of the American Statistical Association **82**:239-248.

Wahba, G. 1983. Bayesian "Confidence Interval" for the Cross-validated Smoothing Spline. Journal of the Royal Statistical Society **B 45**:133-150.

———. 1990. Spline Models of Observational data. SIAM Philadelphia.

Watkins, D.S. 1991. Fundamentals of Matrix Computations. John Wiley and Sons, New York.

Wood S.N. and R.M. Nisbet. 1991. Estimation of Mortality Rates in Stage-Structured Populations. Springer Verlag.

Wood, S.N. and M.B. Thomas. 1999. Super sensitivity to structure in biological models. Proceedings of the Royal Society **B 266**: 565-570.

Wood, S.N. 1994. Obtaining birth and mortality patterns from structured population trajectories. Ecological Monographs **64** 23-44.

———. 1997 Inverse problems and structured population dynamics. in Structured-Population Models. Pages 555-586 *in* S. Tuljapurkar and H. Caswell, editors. Structured-Population Models in Marine, Terrestrial, and Freshwater Systems. Chapman & Hall, New York.

———. 2000. Modelling and smoothing parameter estimation with multiple quadratic penalties. in press Journal of the Royal Statistical Society **B**.

APPENDIX A: Spline Functions

A cubic spline, $f(x)$, say, is the smoothest curve (in the sense of minimum integrated square of second derivative) through a set of n points (x_i, p_i), say (so $p_i = f(x_i)$). There are many alternative ways of representing splines using different sets of basis functions, for example, defining parameters a, b and c_i's:

$$f(x) = a + bx + \sum_{i=1}^{n} c_i |x - x_i|^3$$

is one representation (see, e.g. Wahba 1990), although extra conditions are required to ensure its uniqueness. The representation actually used in the examples in this paper treats the p_i's as parameters of the spline. Specifically I used:

$$f(x) = m_i\zeta_{0i}(x) + m_{i+1}\zeta_{1i}(x) + p_i\psi_{0i}(x) + p_{i+1}\psi_{1i}(x) \quad x_i \le x \le x_{i+1}$$

where m_i is the second derivative of f at x_i (i.e. $m_i = f''(x_i)$) and the functions are: $\zeta_{0i}(x) = [(x_{i+1} - x)^3/h_i - h_i(x_{i+1} - x)]/6$, $\zeta_{1i}(x) = [(x - x_i)^3/h_i - h_i(x - x_i)]/6$, $\psi_{0i}(x) = (x_{i+1} - x)/h_i$ and $\psi_{1i}(x) = (x - x_i)/h_i$, where $h_i = x_{i+1} - x_i$. m_1 and m_n are set equal to zero to yield the so called "natural" spline. The remaining m_i's ($[m_2, m_3 \ldots, m_{n-1}]^T = \mathbf{m}$) are completely determined by the p_i's via the linear equation:

$$\mathbf{Bm} = \mathbf{Hp} \quad (\textit{that is } \mathbf{m} = \mathbf{B}^{-1}\mathbf{Hp})$$

where the $(n-2) \times (n-2)$ matrix $\mathbf{B}$ and the $(n-2) \times n$ matrix $\mathbf{H}$ have zeroes everywhere except as follows: $H_{i,i} = 1/h_i$ $H_{i,i+1} = -(1/h_i + 1/h_{i+1})$ $H_{i,i+2} = 1/h_{i+1}$ $B_{i,i} = (h_i + h_{i+1})/3$ $1 \le i \le n-2$ and $B_{i,i+1} = B_{i+1,i} = h_{i+1}/6$ $1 \le i \le n-3$ ... this representation is quite convenient for computational purposes, but it is straightforward, if tedious, to demonstrate that it can be re-written in the form:

$$f(x) = \sum_{i=1}^{n} p_i \gamma_i(x)$$

under a suitable (but rather long winded) definition of the basis functions $\gamma_i(x)$. Another convenient fact is that:

$$\int [f''(x)]^2 \, dx = \mathbf{p}^T\mathbf{H}^T\mathbf{B}^{-1}\mathbf{Hp}$$

APPENDIX B: Constrained Optimization

This appendix sketches the principles underpinning constrained optimization with a quadratic model. Suppose that we wish to minimise the r.h.s. of (10) subject to the linear equality constraints $\mathbf{A}_f\mathbf{p} = \mathbf{0}$, where $\mathbf{A}_f$ is an $(m \times r)$ matrix. To do this we need to find a form for $\mathbf{p}$ that will ensure that it never violates the constraints. Suppose that it is possible to find a matrix $\mathbf{Q}$ such that $\mathbf{A}_f\mathbf{Q} = \mathbf{T}$, where $\mathbf{T}$ is a matrix whose first $r - m$ columns are zero. This means that if we write the first $r - m$ columns of $\mathbf{Q}$ in a matrix $\mathbf{Z}$, then $\mathbf{A}_f\mathbf{Z} = \mathbf{0}$. Hence for any $r - m$ dimensional vector $\mathbf{p}_z$: $\mathbf{A}_f\mathbf{Z}\mathbf{p}_z = \mathbf{0}$. So writing $\mathbf{p} = \mathbf{Z}\mathbf{p}_z$, ensures that $\mathbf{p}$ will never violate the constraints. It turns out that $\mathbf{Q}$ is quite easy to construct using 'Householder' rotations applied to $\mathbf{A}_f$ (see Watkins, 1991). Substituting for $\mathbf{p}$ in (10) yields:

$$q(\mathbf{p}_z) \approx a + \mathbf{h}^T\mathbf{Z}\mathbf{p}_z + \frac{1}{2}\mathbf{p}_z^T\mathbf{Z}^T\mathbf{G}\mathbf{Z}\mathbf{p}_z$$

Differentiating with respect to each element of $\mathbf{p}_z$ and setting the results to zero yields the minimum:

$$\hat{\mathbf{p}}_z = (\mathbf{Z}^T\mathbf{G}\mathbf{Z})^{-1}\mathbf{Z}^T\mathbf{h}$$

Inequality constraints are dealt with iteratively. Start with a parameter vector that violates none of the constraints and then find the direction in which to alter the parameter vector to minimise the model. If proceeding to this minimum would violate an inequality constraint then a step is taken to the constraint and it is subsequently treated as an equality constraint with minimisation proceeding as indicated above. This process is repeated iteratively, until a minimum can be reached that violates no constraints. By this stage several constraints may be treated as equality constraints and tests have to be made to ensure that they all need to be retained before a minimum can be accepted.

It is always tempting to try and perform constrained optimization by using penalty function methods, which add a penalty to the objective function for violating constraints - such methods are simple to implement, but tend to spoil the quadratic model, particularly in the vicinity of constraint boundaries and also tend to require *ad hoc* adjustment of the strength of the penalties. The methods used here avoid making problems more non-linear than they already are.

The practical details of how constrained optimization is done efficiently and robustly are quite involved, and the interested reader should consult Gill *et al.* (1981) and references therein.

APPENDIX C: Bootstrap Restarting

An often effective approach to dealing with irregular objective functions that may have multiple local minima is to use *bootstrap restarting* in conjunction with a minimisation method designed for smooth objective functions. Given the relationship between bootstrapping and statistical inference, a fitting objective based on a bootstrap resample from the original data is likely to share the statistically important features of the original objective function, while differing in details, such as the location of small scale local minima. Having applied a minimisation method to the original fitting objective $q(\mathbf{p})$ in order to obtain best fit parameter estimates $\hat{\mathbf{p}}$, it is sometimes possible to improve these estimates by iterating the following simple steps:

1. Sample with replacement from the original data $\mathbf{y}$, and use this bootstrap resampled data, $\mathbf{y}^*$ in place of the original data to construct a perturbed objective function $q^*(\mathbf{p})$ (each resampled y_i will be accompanied by all associated information, such as sample time, life history stage etc.). Starting from $\hat{\mathbf{p}}$ apply the minimisation method to q^* to obtain parameter estimate vector $\mathbf{p}^*$.

2. Starting from $\mathbf{p}^*$, apply the minimisation method to the real objective $q(\mathbf{p})$ to obtain a parameter estimate $\tilde{\mathbf{p}}$, if $q(\tilde{\mathbf{p}}) < q(\hat{\mathbf{p}})$ then set $\hat{\mathbf{p}}$ to $\tilde{\mathbf{p}}$.

The basic idea is that while the global minimum of the true and bootstrap objectives will be in much the same location, they will have different local minima - hence the bootstrap steps can free the optimization method from local minima of the true objective, by moving the parameters away from this minimum. The approach is preferable to other strategies for randomly jumping out of local minima in that the size and direction of the step out of the minima automatically takes account of the shape of the objective. Of course for a smooth well behaved objective function, without local minima, the method produces no improvement. Note also that general convergence criteria for this approach are difficult to formulate.