Aspects of Applied Biology 53: 41-49

Semi-parametric population models

By S N WOOD

Mathematical Institute, University of St Andrews, North Haugh,
St. Andrews, Fife KY16 9SS UK   snw@st-and.ac.uk

Summary

An example of an attempt to use simple models to predict the outcome of releasing a fungal pathogen as a bio-control agent against grasshoppers is used to illustrate the possibility of extreme sensitivity of model predictions to the details of model specification. A partial solution to this problem is to formulate models in a more general way than is conventional in population dynamic modelling, but is similar in spirit to the method of generalized additive modelling. Component functions of a model, whose exact form is unknown *a priori* can be represented as non-parametric smooth functions, thereby avoiding some of the incidental assumptions associated with assuming arbitrary parameter sparse functional forms. After a brief description of this approach to modelling it is applied to the grasshopper control example.

Keywords: partially specified model; ecological risk assessment; ecological prediction; impact assessment; GAM; spline model; grasshopper; pest control

Introduction

Population dynamic models are useful as cartoons of biological systems since they are a good bit more tractable than the real thing, and therefore easier to understand. In this role models have a long history of providing insight into how biological systems might work, as well as giving much unwholesome amusement to their creators. Before the advent of cheap computing it was seldom sensible to work with anything but the simplest models, since the heroic efforts required to reach an analytic understanding of more complex models seldom repays with *breadth* of understanding. This is because the inclusion of sufficient model complexity to describe one system in detail tends to preclude the use of the model for other systems, so that generality seems to decline with effort expended. More recently, however, cheap computer power has opened the way to the use and understanding of more realistic models, and this in turn opens up the possibility of statistically validating such models against data, and of using models to make quantitative predictions as well as gain general insights.

In this paper I present a case study based on prediction of the success or failure of biological control of grasshoppers, showing how prediction using non-linear population dynamic models can show alarming and counterintuitive sensitivity to apparently insignificant details of model specification (Wood & Thomas, 1999). Consideration of the mechanism of this sensitivity suggests an alternative and hopefully safer approach to model building based on *partially specified* models, in which model elements that are poorly known *a priori* are represented in a fairly general way as unknown functions, thereby avoiding some of the spurious assumptions associated with more conventional model formulation. Having outlined the key ingredients of this approach, I then apply it to the grasshopper control example.

Grasshoppers, pathogens and extreme sensitivity to model formulation

Insect pests are estimated to consume between 20 and 30 percent of worldwide crop production each year (around US$300 billion annually, Hill 1997), so that the prediction of success or failure of control programmes is of some applied importance, while even small increases in the efficiency of control have the potential to provide substantial benefits. Grasshoppers and locusts have been significant agricultural pests for some time (Exodus chapter 10, Steedman 1990) and this section concerns a simple model designed to investigate the scope for successful control of the rice- grasshopper *Hieroglyphus dagenensis* by introduction of the entomopathogen *Metarhizium flavoviride*.

H. dagenensis is active during a rainy seasons of some 3 to 4 months duration, spending the rest of its life cycle as resting eggs, during the dry season. In the active part of the life cycle it is suceptible to infection by the fungal pathogen *M. flavoviride*, which typically kills the host after an incubation period of some 12 days. The dead host is not immediately infectious but becomes so as the fungus builds up in and on the cadaver. Typically it reaches maximum infectivity a week or so after death, with a slow decline thereafter: there is substantial infectivity 6 weeks after death (see figure 1 Thomas, Wood & Lomer 1995), and ground that contained infectious cadavers during the previous wet season still shows residual infectivity after an intervening dry season (Thomas, Gbongboui & Lomer 1996). Grasshoppers that survive the rainy season lay eggs before dying, and these eggs lie dormant until the next years rains. Early attempts to model this system (Thomas *et al.* 1995; Wood & Thomas 1996) employed the simple assumption that the probability of an individual becoming infected is directly proportional to the density of infectious material in its environment. However, this ‘proportional mixing’ assumption is untenable in the face of experimental data in which density of infective material is manipulated directly (see Wood & Thomas 1999). Unsurprisingly it seems that infection risk is a saturating function of the density of infectious agent.

Given this sketch of the biology, a reasonable model for the grasshopper- pathogen dynamics within a rainy season can be formulated in terms of 3 state variables: H is the healthy host density (m^{-2}) while A_0 and A_1 are two dummy variables used to obtain the desired build up and decline of individual cadaver infectivity. A_0 can be thought of as an index of the total density (per m^2) of resources within infected cadavers that has yet to be turned into pathogen, while A_1 is an index of the total density of pathogen. The rate at which H declines per unit time is given by the product of a saturating function of A_1, $f(A_1)$, and host density itself, so that:

$$\frac{\mathrm{d}H}{\mathrm{d}t} = -f(A_1)H$$

The rate of change of pathogen density is governed by the rate at which cadaver resources are converted to pathogen, less the rate at which pathogen becomes inactive. This leads to:

$$\frac{\mathrm{d}A_1}{\mathrm{d}t} = c(A_0 - A_1)$$

(a full justification of the equations for A_1 and A_0 is given in Wood & Thomas 1996 or Thomas *et al.* 1995). The rate of change of the 'cadaver resource' term A_0 is given by the rate at which healthy hosts die less the rate at which resources are converted to pathogen. The rate at which hosts die is just the rate at which they got infected τ days earlier, where τ is the incubation period. Hence:

$$\frac{\mathrm{d}A_0}{\mathrm{d}t} = c[f(A_1(t-\tau))H(t-\tau) - A_0]$$

Between seasons, H is multiplied by a finite rate of increase, F and supplemented by a small amount of immigration m, while A_0 and A_1 are multiplied by the between season pathogen survival rate, γ. In the work reported here $F = 4$, $\tau = 12\mathrm{d}$, $m = 0.1\mathrm{m}^{-2}$ and $\gamma = 0.02$.

Without a detailed mechanistic model of the infection process it is difficult to know what form to give the function $f(\cdot)$, so in Wood and Thomas (1999) three alternative saturation functions were tried: the well known Michaelis-Menton equation, $f(x) = ax/[b + x]$; a form used by Briggs & Godfray (1995), $f(x) = k\log[1+\alpha x/k]$; and a form due to Hochberg (1991), $f(x) = \beta x^{1+q}$.

To estimate the parameter c and the parameters of each $f(\cdot)$ the within season population dynamic model given above was fitted directly to experimental data. In these experiments (performed in Northern Benin in 1994) cohorts of healthy grasshoppers were exposed to infectious cadavers in field cages for three days before being incubated in the lab, where the proportion surviving was noted. Different cohorts were exposed to four cadaver densities at 6 times after cadaver death, so that the experiment provided information on both the time course of infectivity and the shape of the relationship between pathogen density and infectivity. Details are given in Wood & Thomas (1999). The model can be used to predict survival for each of the 24 cohorts, and its parameters can be optimized to maximise the accuracy of this prediction. Uncertainty in model parameters was estimated by using parametric bootstrapping (see Davison & Hinkley, 1997) to generate 99 replicate parameter sets for each of the three forms of $f(\cdot)$.

Figure 1a shows the best fit functions for the three alternative formulations, and figure 1b superimposes the corresponding 98% confidence bands. From these plots the alternative representations of saturating infectivity appear practically indistinguishable, and it might reasonably be supposed that it would not much matter which is chosen. Figure 2 shows that this is in fact not the case. The Hochberg model predicts sustained control after a single pathogen application, while the other two models predict that repeated pathogen re-introduction is needed to maintain control. Furthermore the predictions of each model are consistent across all 99 bootstrap replicate parameter sets - the Hochberg model always predicts sustained control, the other two models always predict the necessity of repeated spraying.

What this example shows is the potential for predictions to go badly wrong as a result of apparently innocuous modelling assumptions. Looking more carefully at the models it appears that the Hochberg model gives a rather big boost to the pathogen population at very low densities: but if the Hochberg model had been the only one used there would have been nothing

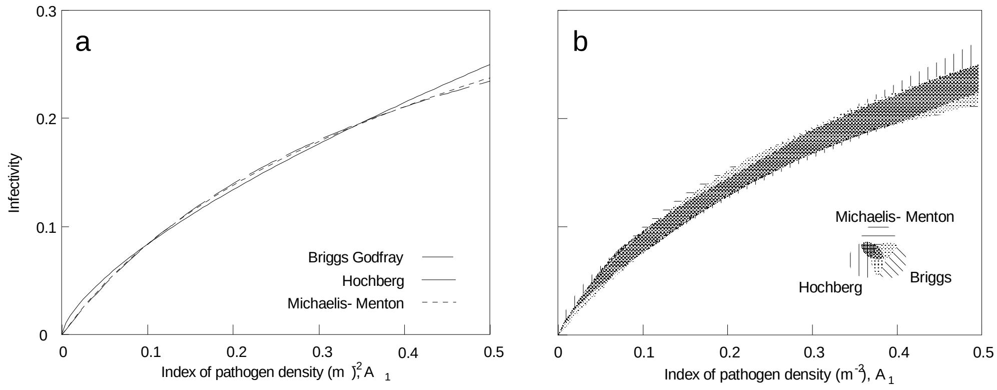

Figure 1: (a) Shows each of the 3 alternative functional forms describing infectivity with the parameters that achieved the best fit to experimental data. (b) Shows overlapping 98% confidence bands for each of these three functions, obtained by parametric bootstrapping from the experimental data. Direction of shading distinguishes the different bands. Notice both the closeness of the best fit functions and the almost complete overlap of the confidence bands.

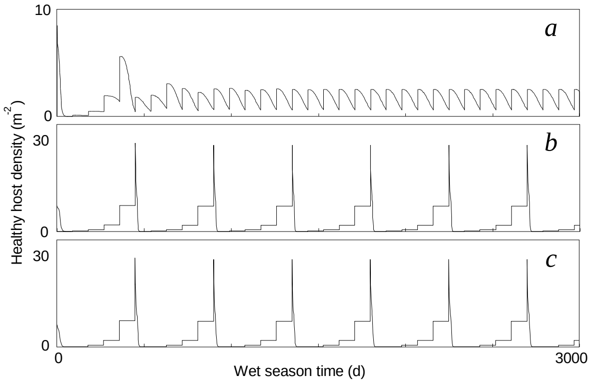

Figure 2: Typical population dynamics predicted by grasshopper- pathogen models fitted to data and incorporating (*a*) Hochberg's function, (*b*) the Michealis-Menton equation and (*c*) the Briggs-Godfray equation as descriptions of the relationship between pathogen density and infectivity. The predicted dynamics were qualitatively similar across all 99 bootstrap generated parameter sets tried for each model. The range of A_1 is chosen for consistency with what occurs in the models discussed as well as consistency with experimental data.

to caste suspicion on its validity. Indeed the bootstrapping exercise quantifying the uncertainty in the parameters of the model would tend to suggest rather robust predictions. Clearly in this example the difficulties arise because the population dynamic model is much more sensitive to tiny changes in the functional form of $f(\cdot)$, than to variability in $f(\cdot)$'s parameters *given* a functional form. In order to understand this phenomenon it helps to describe it in quite general terms.

Each particular functional form for $f(.)$ can be thought of as defining a space of functions, with different parameter values yielding different elements of the space. Clearly the model predictions do not seem terribly sensitive to variation within any one of the three spaces defined in this way (at least provided variation is on a scale consistent with the experimental data to which the model has been fitted). Now suppose that we could define the true space of possible forms of $f(\cdot)$. The population dynamic model is quite likely to be highly sensitive to variation in some directions within this space, and quite insensitive to variation in other directions, in the same way that non-linear models usually show sensitivity to some parameters and robustness to others. All that is required to get the phenomenon observed here is that the chosen functional forms restrict variation in this space to directions to which model predictions are insensitive while the differences *between* functional forms defines a direction in the space to which predictions are highly sensitive. Figure 3 attempts to illustrate these ideas, which are developed more fully in Wood & Thomas (1999). What this general discussion suggests is that the phenomenon has the potential to occur in other models and is unlikely to be a pathological quirk of this particular host pathogen model.

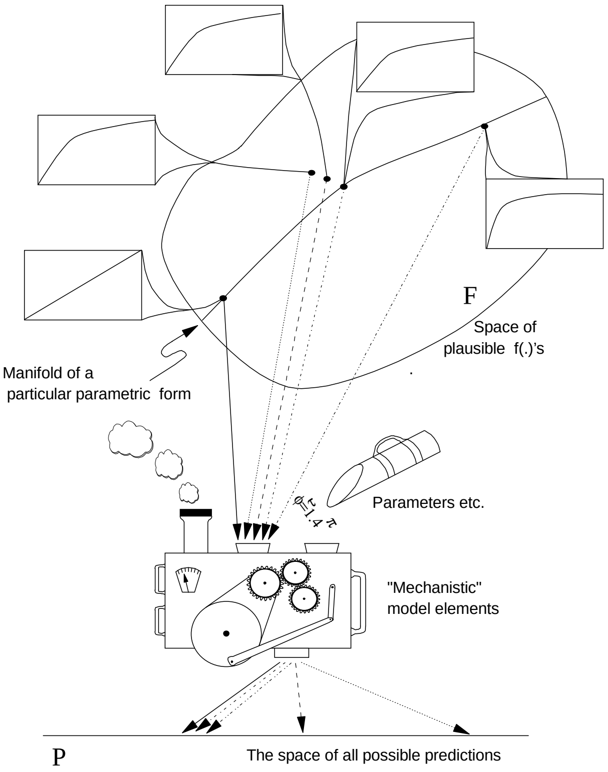
F
Space of plausible f(.)'s
Manifold of a particular parametric form
$\phi = 1.4$ τ π
Parameters etc.
"Mechanistic" model elements
P
The space of all possible predictions

Figure 3: Schematic diagram of the origins of extreme sensitivity to specification. The population dynamic model can be thought of as having mechanistic elements that transform model inputs in the form of component functions and parameters into model outputs in the form of predictions. The diagram illustrates how movement within the space of plausible $f(\cdot)$'s can cause radically different changes in prediction according to the direction of movement: it's quite possible for movement along the manifold of a particular functional form to cause little change in predictions, while movement off that manifold can cause big changes. (The type of dashing on arrows identifies which function produces which prediction.)

Semi-parametric population dynamic models

The previous section illustrated a problem in modelling for predictive purposes: when describing elements of the modelled system that must be characterised without a great deal of mechanistic understanding, it is possible to introduce apparently minor incidental assumptions that can have a surprisingly serious impact on predictions. Part of the difficulty is that although it isn't always clear exactly how to write down a model, the act of writing down *something* tends to suggest that more is known about the system than is actually the case - and it's this addition of 'extra' information in the form of unsubstantiated model structure that has the potential to cause difficulties.

One way to reduce these problems is suggested by the explanation of the extreme sensitivity problem given at the end of the previous section. Why not construct our model not as a 'fully specified' model in which everything is assumed known but for a few unknown parameters, but as a 'partially specified' model in which unknown functions are left as just that? In the context of the grasshopper- pathogen model, this means that the specification of $f(\cdot)$ becomes something like '$f(\cdot)$ is a smooth monotonically increasing function, for which $f(0) = 0$'. The space of possible $f(\cdot)$'s can be chosen to be large enough to avoid too much artificial restriction on the form of $f(\cdot)$. The population dynamic model can then be thought of as a mapping from the unknown parameter and function to predictions corresponding to the data that the model is intended to fit:

$$\boldsymbol{\mu} = \mathbf{M}(f, c)$$

where $\mu_i = E(y_i)$ and y_i is the ith observation to be fitted by the model (E denotes 'expected value'). So in principle model fitting can proceed by attempting to minimise:

$$\sum_{i=1}^{n}(y_i - \mu_i)^2$$

(or some other measure of fit). However, there are two problems to be overcome to make this practical. Firstly it is necessary to represent $f(\cdot)$ in some tractable manner. The way to do this is to set up a basis for $f(\cdot)$ so that it can be represented in terms of a finite number of unknown parameters $\mathbf{c}$ and some 'basis functions', $\gamma_i(x)$, which have no unknown parameters. i.e. let

$$f(x) = \sum_{i=1}^{p} c_i\gamma_i(x)$$

A familiar example of a set of basis functions is $\{\gamma_i(x) = x^i : i = 0 \ldots p\} = \{1, x, x^2, \ldots, x^p\}$ which is used to represent polynomial functions ($f(x) = c_0 + c_1x + c_2x^2 + \ldots c_px^p$). Althugh easy to understand, the polynomial basis is a rather poor choice as it tends to lead to unstable estimates, and it is better to use the cubic spline basis, which has much better approximation theoretic properties (see Wahba 1990; deBoor 1978). Use of this basis function representation means that the problem of finding the unknown function has been reduced to the problem of finding the vector of its coefficients relative to the basis: $\mathbf{c}$, and this is a fairly ordinary non-linear optimization problem (Gill, Murray & Wright 1981). Provided p is big enough the space of functions that can be represented is quite large.

The second problem is the measure of fit. A sufficiently complicated $f(\cdot)$ will be liable to cause the model to overfit the y_i's. That is, the model will fit both the signal and the random error in the data. One way around this problem is to penalise model complexity as part of the lack of fit measure, so that simple smooth f's are favoured relative to more wiggly and complex

f's. A suitably modified fitting objective is:

$$minimise \ \sum_{i=1}^{n}(y_i - \mu_i)^2 + \lambda \int [f''(x)]^2 dx$$

where the first term measures infidelity of the model to the data and term measures wiggliness of $f(\cdot)$ (formally, the integrated square curvature of $f(\cdot)$). λ controls the trade off between the competing objectives of close fit to data, and smoothness of $f(\cdot)$. High λ gives poor predictions of data but a simple model, while low λ tends to promote close matching of the data with a complicated form for f. Given λ and the basis function representation of $f(\cdot)$ this modified objective can be minimised numerically (see Gill *et al.* 1981).

Clearly some means of choosing λ is now required, and a plausible method is cross validation. This works by leaving out one y_i at a time and fitting the model to the remaining data. The squared difference between the model prediction of the missing y_i and y_i itself is then calculated. The average of these squared differences across all missed out data provides a measure of how badly the model is doing that can be used to select λ. High λ tends to mean that the model doesn't match the data to which it has been fitted very closely and it does no better with missing data; low λ means that every random fluctuation in the data is fitted, which leads to a very poor match to missing data. Middling values of λ which correctly partition signal and noise, will tend to give the best cross validation scores. The version of cross validation just described has some technical problems (see Wahba, 1990), but a modification 'generalized cross validation' (GCV) solves these, and results in the score:

$$V(\lambda) = \frac{\sum(y_i - \mu_i)^2}{[\sum(1 - \partial\mu_i/\partial y_i)]^2}$$

which is minimised to choose λ. The term in the denominator is the square of the estimated error degrees of freedom for the model.

A general framework for this sort of approach is given in Wood (*submitted a*): it is possible to work with multiple unknown functions in a model although GCV is quite challenging in this case (Wood *submitted b*); it is also possible to impose linear inequality constraints on functions and coefficients, and to use general exponential family likelihoods in place of the least squares term in the fitting objective. There are some numerical difficulties waiting to trap the unwary in attempting to use these methods, but these are also documented (along with solutions) in Wood (*submitted a*).

In short a population dynamic model to be fitted to data can be written down in terms of unknown functions and unknown parameters rather than just unknown parameters, leading to increased flexibility in model structure, and a reduced chance of arbitrary errors of misspecification. The unknown functions can then be represented using spline functions, and the model fitted to data with complexity of the unknown functions penalized - the degree of penalization is chosen by minimising a GCV criterion.

A semi-parametric approach to the grasshopper pathogen system

As an illustration of the use of the partially specified models introduced in the last section, I reformulated the population dynamic model for *H.dagenensis* and *M. flavoviride* as a partially specified model, representing $f(\cdot)$ as a monotonically increasing smooth function, passing through the origin and having at most 20 degrees of freedom. The idea is that by allowing the model so much potential flexibility mis-specification errors resulting from more restrictive choices of $f(\cdot)$ will be avoided. Hence the form of $f(\cdot)$ that results from fitting should be dominated by what the experimental data suggests is appropriate, rather than by prior assumptions.

Model parameter c and the unknown function were estimated by fitting directly to the experimental data described earlier (and in Wood & Thomas, 1999) using a constrained Quasi-Newton approach (Gill *et al.* 1981). The complexity of $f(\cdot)$ was chosen by GCV. Figure 4a shows (pointwise) 98% bootstrap confidence limits for $f(\cdot)$ given the GCV choice of smoothing parameter. Figure 4d shows corresponding typical predicted population dynamics over multiple rainy seasons. All bootstrap replicates gave similar dynamics. Hence the partially specified model is in qualitative agreement with the Michaelis Menton and Godfray Briggs formulations. However the partially specified model also allows further investigation of other uncertainties in $f(\cdot)$. Figures 4b and 4c show the confidence bands corresponding to smoothing parameters of 10% and 1% of the GCV optimum. These more flexible functional forms, exhibit greater variability, but again all produced dynamics similar to figure 4d, somewhat reinforcing the conclusion that sustained control is very unlikely in this system.

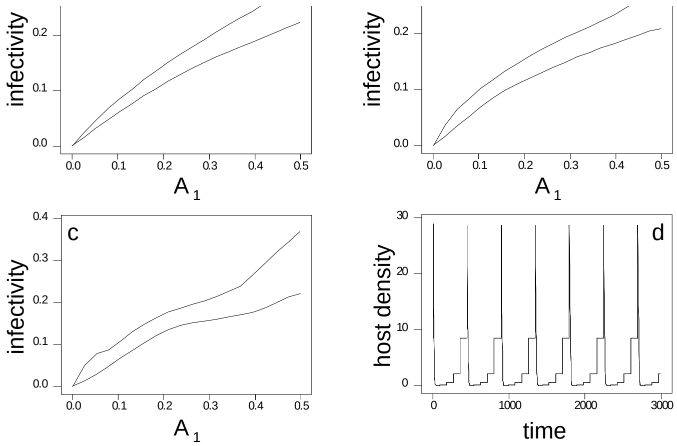

Figure 4: *a*- *c* show bootstrap 98% confidence bands for the unknown function $f(\cdot)$ when the smoothing parameter λ was set to 1, 0.1 and 0.01 × the optimum λ chosen by GCV. *d* shows the form of the dynamics that resulted for all 99 bootstrap replicates of each smoothing parameter choise - there were small quantitative differences between replicates, but no qualitative difference.

Discussion

The example presented at the beginning of this paper shows that prediction with population dynamic models can be dangerous, in that predictions can display great sensitivity to apparently trivial variation in model specification. Consideration of the origins of this sensitivity suggest an alternative approach to modelling that may go some way to reducing the chances of such problems, namely the formulation of partially specified models in which some poorly known model elements are specified only in rather general terms as being ‘smooth functions’. This approach is in tune with model statistical modelling using GAMs (see Hastie & Tibshirani, 1990) and has previously been applied in some special cases where the models have convenient forms (e.g. Wood & Nisbet 1991, Wood 1994, Ellner *et al.* 1997,1998). Recently a much more general framework for such models has been worked out (Wood, *submitted a*) with software also available (Wood 1999). The application to the problem with which this paper started illustrates the utility of the approach: we still cannot be sure of making the right prediction, but at least the chances of being badly wrong should be reduced as a result of removing some of the scope for model mis-specification.

References

de Boor, C 1978. *A practical guide to splines* Springer Verlag, New York.

Briggs, C.J., Godfray H.C.J. 1995. The dynamics of insect- pathogen interactions in stage structured populations *The American Naturalist* **145:** 855-887.

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

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, D.W. Nychka 1998. Noise and nonlinearity in measles epidemics: combining mechanistic and statistical appraoches to population modelling. *The American Naturalist* **151(5):** 425-440.

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

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

Hill, D.S. 1997 *The Economic Importance of insects.* Chapman & Hall

Hochberg, M.E. 1991. Non-linear transmission rates and the dynamics of infectious disease *Journal of theoretical Biology* **153:** 310-321

Steedman, A. 1990. *1990 Locust Handbook, 3rd edn* Chatham: Natural Resources Institute.

Thomas, M.B., C. Gbongboui, C.J. Lomer 1996. Between-season survival of the grasshopper pathogen Metarhizium flavoviride in the Sahel. *Biocontrol Science and Technology* **6**: 569-573.

Thomas, M.B., S.N. Wood, C.J. Lomer 1985. Biological control of Locusts and Grasshoppers using a fungal pathogen: the importance of secondary cycling. *Proceedings of the Royal Society of London (B)* **259**:265-270

Wahba, G. 1990.*Spline Models of Observational data.* SIAM Philadelphia.

Wood, S.N., M.B. Thomas 1996. Space, time and the persistence of virulent pathogens. *Proceedings of the Royal Society of London (B)* **263**:673-680

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

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

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

Wood *submitted a* Partially specified ecological models. *Ecology*

Wood *submitted b* Modelling with multiple quadratic penalties. Journal of the Royal Statistical Society (B)

Wood 1999 `http://www.ruwpa.st-and.ac.uk/simon/ddefit.html`