9

Balancing Statistics and Ecology: Lumping Experimental Data for Model Selection

Nelly van der Hoeven, Lia Hemerik and Patrick A. Jansen

ABSTRACT

Ecological experiments often accumulate data by carrying out many replicate trials, each containing a limited number of observations, which are then pooled and analysed in the search for a pattern. Replicating trials may be the only way to obtain sufficient data, yet lumping disregards the possibility of differences in experimental conditions influencing the overall pattern. This paper discusses how to deal with this dilemma in model selection. Three methods of model selection are introduced: likelihood-ratio testing, the Akaike Information Criterion (AIC) with or without small-sample correction and the Bayesian Information Criterion (BIC). Subsequently, we apply the AICc method to an example on size-dependent seed dispersal by scatterhoarding rodents.

The example involves binary data on the selection and removal of *Carapa procera* (Meliaceae) seeds by scatterhoarding rodents in replicate trials during years of different ambient seed abundance. The question is whether there is an optimum size for seeds to be removed and dispersed by the rodents. We fit five models, varying from no effect of seed mass to an optimum seed mass. We show that lumping the data produces the expected pattern but gives a poor fit compared to analyses in which grouping levels are taken into account. Three methods of grouping were used: per group a fixed parameter value; per group a randomly drawn parameter value; and some parameters fixed per group and others constant for all groups. Model fitting with some parameters fixed for all groups, and others depending on the trial give the best fit. The general pattern is however rather weak.

We explore how far models must differ in order to be able to discriminate between them, using the minimum Kullback-Leibler distance as a measure for the difference. We then show by simulation that the differences are too small to discriminate at all between the five models tested at the level of replicate trials.

We recommend a combined approach in which the level of lumping trials is chosen by the amount of variation explained in comparison to an analysis at the trial level. It is shown that combining data from different trials only leads to an increase in the probability of identifying the correct model with the AIC criterion if the distance of all simpler (=less extended models) to the simulated model is sufficiently large in each

T.A.C. Reydon and L. Hemerik.,(eds.), Current Themes in Theoretical Biology,
233-265.

trial. Otherwise, increasing the number of replicate trials might even lead to a decrease in the power of the AIC.

Keywords: AIC, *Carapa procera*, Kullback-Leibler distance, Likelihood-Ratio test, model selection, *Myoprocta acouchy*, non-central chi-square distribution, power, Red acouchy, scatterhoarding, seed dispersal, seed size.

9.1 INTRODUCTION

It is quite common in ecology to have several candidate models for describing ecological observations (Hilborn and Mangel, 1997). In some cases, models are based on different assumptions about the underlying mechanism, whereas in others, models are used to describe the relationship between factors. Both cases however, require the identification of the model best conforming to the observations.

Several criteria exist to determine which model fits best, for instance likelihood-ratio (LR) testing, the AIC (Akaike Information Criterion) and the BIC (Bayesian Information Criterion) (see Burnham and Anderson, 2002; Hilborn and Mangel, 1997; Linhart and Zucchini, 1986; Borowiak, 1989 for extensive reviews of model discrimination methods). After an initial comparison of the three methods (LR, BIC and AIC) we focus in this paper on the AIC that treats all models as equivalent and allows comparison of nested and non-nested models. Thus, the AIC assumes that each model can be the true model and none of the models is preferred.

Ecological experiments often accumulate data for model fitting by carrying out several independent trials, each containing a limited number of observations, which are then pooled and analysed for a pattern. Replicating trials may be the only way to obtain sufficient data, yet lumping is not a priori admissible. If conditions between trials differ, simply lumping all trials is even a priori inadmissible. Such situations require the model be fitted to the data of each trial separately, each with different model parameters. This will, however, affect the ability to distinguish between models (the identifiability), and the possibility to derive general conclusions from the properties of the best fitting model. We consider a model identifiable if the probability of being the best-fitting on its own simulated data exceeds 80%. An alternative approach is to assume that the parameters in each trial are independent drawings from some probability distribution.

This paper explores the consequences of data lumping for model selection using data on seed selection by scatterhoarding rodents as an example. The question to be answered is whether there is an optimum size for seeds to be selected and dispersed by these rodents. In our example, it is biologically unrealistic as well as technically difficult to provide a single animal with

>1000 marked seeds at a given time, while it is ecologically desirable to consider selection by different individuals. The only way to detect a trend was to carry out many independent replicate trials with small batches of seeds, spaced apart in time and space, and involving different individual rodents. The challenge is to balance statistical requirements with ecological feasibility.

We start with the description of three methods for model selection (Section 2). In Section 3, we apply two of these methods to a data set on seed dispersal by scatterhoarding rodents. Next, we have fitted the same models to simulated data in order to obtain an impression of the identifiability of the chosen models for certain combinations of parameter values, that is which percentage of the simulation runs are classified correctly (Section 4). Finally, conclusions of the model fitting both the experimental data and the simulated models are given and discussed (Section 5).

9.2 METHODS FOR MODEL DISCRIMINATION

Hypothesis testing

Models that are to be compared are often nested: one model (the nested model) is a special case of another, more complex model with one or more of the parameters of the complex model fixed. For example, the linear model $y = a + bx$ is a special case of the quadratic model $y = a + bx + cx^2$ with $c = 0$. If these models are compared with the usual hypothesis testing method, the null hypothesis is that the simplest model is true, unless the observed data are much more likely under the more complex model. A general method to test the simple model against the more complex is the Likelihood-Ratio test (LR test). This test compares the ratio of the maximum likelihood (ML) for the two models to a critical value. Instead of the ratio between the ML's the difference between the log of both ML's can be used. Twice the difference between these maximized log-likelihoods is approximately χ^2 distributed. This means that for large numbers of observations the α-critical value for 2×(the difference in maximized log-likelihoods) is approximately $\chi^2_{\alpha\,\nu}$ with ν the difference in the number of parameters of the extended (k_2) and the more simple model (k_1), so $\nu = k_2 - k_1$. For a small number of observations, the χ^2 approximation may not hold.

So, in general let L_1 and L_2 be the maximum of the likelihood function for the simple and the extended model. Then, for large numbers of observations

$$T = 2\times(\ln(L_2) - \ln(L_1)) \to \chi^2_\nu. \qquad (9.1)$$

The AIC: finding the model giving the best approximation

One approach to discriminate between models, described by Akaike (1974), is to assume that there is some - unknown - “real” model, and that the model having the minimum distance to that unknown real model is the best approximation. It uses the so-called Kullback-Leibler (K-L) distance (Kullback and Leibler, 1951) as a measure of the distance between models. For continuous models, the K-L distance of the approximate model *g* with parameter θ to the real model *f* is

$$I(f,g_\theta)=\int f(y)\ln\left(\frac{f(y)}{g(y\,|\,\theta)}\right)\mathrm{d}y\,. \tag{9.2}$$

This distance is related to the information lost by using model *g* with parameter θ instead of the real model *f.* It indicates how good model *g* with parameter θ approximates model *f.* Note that *I*(*f*,g) ≠ *I*(g,*f*), that is, the K-L distance is not commutative and therefore is not a real distance.

For discrete models with *k* possible outcomes y_i ($i = 1, \ldots, k$), the K-L distance can be written as

$$I(f,g_\theta)=\sum_{i=1}^{k}p(y_i\,|\,f)\ln\left(\frac{p(y_i\,|\,f)}{p(y_i\,|\,g_\theta)}\right). \tag{9.3}$$

In general, the real model, *f,* will be unknown. Fortunately, when two models, g_1 and g_2 have to be compared, the difference $I(f,g_1) - I(f,g_2)$ does not depend on the real model *f.* Using this, Akaike (1974) developed the AIC (An Information Criterion, better known as Akaike’s Information Criterion) which is defined as

$$AIC = 2[k - \ln(L)] \tag{9.4}$$

where *k* denotes the number of estimated parameters and *L* is the maximum of the likelihood function. The model with the minimum *AIC* is considered to be the best fitting model. This approach allows a simple ranking of the models and is also appropriate for comparing non-nested alternatives. Using the *AIC*, Model 2 is preferred above Model 1 if $AIC_1 - AIC_2 > 0$, so if

$$T = 2[\ln(L_2) - \ln(L_1)] > 2(k_2 - k_1) = 2\nu\,. \tag{9.5}$$

A correction term should be added to the *AIC* if the number of parameters, *k*, is large, or the number of observations, *n*, is small. There is no universal best correction term, but the corrected *AIC*, *AICc* as given by Hurvich and Tsai (1989),

$$AICc = AIC + C(k,n) = 2[k - \ln(L)] + \frac{2k(k+1)}{n-(k+1)}, \tag{9.6}$$

performs reasonably well for most models (Burnham and Anderson, 2002).

Using the *AICc*, Model 2 is preferred over Model 1 if

$$\begin{aligned} T \;&>\; 2\,(k_2-k_1)\left(1+\frac{n\,(1+k_1+k_2)\;-\;(k_1+1)(k_2+1)}{(n-\;(k_1+1))(n-\;(k_2+1))}\right) \\ &=\; 2\;\nu\,(\,1+\text{correction}\;\;\text{term}). \end{aligned} \qquad (9.7)$$

The correction term only depends on the number of parameters in both models (k_1, k_2) and the number of observations (n).

The BIC: finding the true model within a set of models

There may be a reason to believe a priori that one of the models in a set of models is true. The BIC described by Schwarz (1978) is a selection criterion for identifying such a true model with an as large as possible probability. The BIC is also based on twice the log ML's, and uses a correction term increasing with the number of observations,

$$BIC = k\,\ln(n) - 2\ln(L) \qquad (9.8)$$

The model with the minimum *BIC* is considered to be the best fitting model. Using the *BIC*, Model 2 is preferred above Model 1 if $BIC_1 - BIC_2 > 0$, so if

$$T = 2[\ln(L_2) - \ln(L_1)] > (k_2 - k_1)\ln(n) = \nu\ln(n)\,. \qquad (9.9)$$

The *BIC* is a consistent estimator for the model type: if the number of observations becomes very large, the probability that the correct model is identified increases to 1. It should be noted however, that to meet the condition "very large" extremely large sample sizes are indeed required. For instance, identifying the correct model with high probability requires a very large number of observations. Umbach and Wilcox (1996), for example, needed as much as 125,000 simulated observations to reach a power of 0.79.

Comparison between the three methods

LR, AIC, AICc and BIC use the same test statistic T to find the best approximate model. If the extended model has one extra parameter ($\nu = 1$), the χ^2 approximation for the LR test criterion at $\alpha = 5\%$ leads to rejection of the more simple model if $T > 3.84$. The threshold value for T increases with an increasing degree of freedom (see Figure 9.1). The AIC considers all models equivalent and for $\nu = 1$ chooses the more extended model if $T > 2$. The threshold value for T increases linearly with higher values of ν (see Figure 9.1). For a difference of one parameter, the AIC criterion will choose the extended model with (approximately) probability 0.16 if the simple model is true.

The critical value for T in the AICc criterion is more complex. If the simple model has two parameters and the extended model three, the AICc criterion chooses the more extended model if $T > (2n(n-1))/((n-3)(n-4))$. The critical value for T in a trial of only five observations ($n = 5$), for example, is 20. The AICc criterion becomes less strict for the extended model with increasing n, and for $n > 12$, the AICc criterion is less strict than the LR one. Figure 9.1 shows the threshold values of the *AICc* for T with $n = 25$ or 100 and a 2-parameters simplest model. In this figure, the BIC criterion is given for the same numbers of observations.

In contrast to the AICc, the BIC criterion becomes stricter for the extended model as the number of observations increases (Figure 9.1). If the difference in the number of parameters is one ($\nu = 1$), the *AIC* and *BIC* are almost identical for $n = 8$, and the results of the BIC criterion and the χ^2 approximation of the LR test are about the same for $n = 47$. The preference for the more parsimonious model with increasing difference in number of parameters increases faster for the AIC, the AICc and the BIC than for the LR test (Burnham and Anderson, 2002).

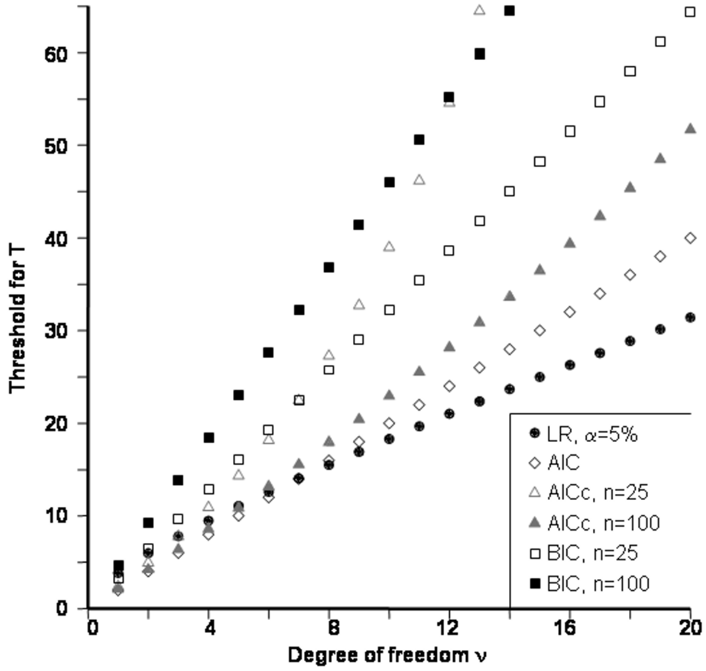


Figure 9.1. The threshold value for the test statistic T as a function of the difference in the number of parameters (degrees of freedom ν) of the two compared models. For the AICc the number of parameters in the simpler model is set at 2.

The differences in critical value for model selection illustrate a fundamental difference between the three methods. Classical hypothesis testing with a likelihood-ratio test assumes that the most simple model is true unless the observed values are very unlikely (probability less than α, with $\alpha = 0.05$ as most common choice). Using the AIC or AICc criterion, it is assumed that none of the models is true, but it is tried to minimize the (K-L) distance to the real, unknown, model in order to choose the model giving the best prediction for new data sets. Using the BIC criterion, it is assumed that one of the models is true, and the probability of choosing that true model is maximized. Note that only the difference between the log-likelihoods is of interest in each method. Therefore, all *AIC*, *AICc* and *BIC* can be decreased by a constant. The smallest *AIC* (*AICc*, *BIC*) is often subtracted from all *AIC* (*AICc*, *BIC*) values, making the smallest *AIC* (*AICc*, *BIC*) 0.

The Kullback-Leibler distance between models

Choosing among models requires quantification of the difference between them. The fundamental distance measure for the AIC is the Kullback-Leibler (K-L) distance. The K-L distance between models can be determined for one realization, but also cumulative for a combination of *n* observations, which is of interest for model selection. Then, the K-L distance between the models is the expectation, given the extended model, of the difference between the simpler model and the more extended model in the log-likelihoods for *n* combined realizations. It depends, among others, on the values of the independent variables in the observations.

In an example, we will show how the K-L distance between models can be determined if *n* realizations of the model are observed. We assume that some discrete variable, *y*, is observed, and that *y* can have *m* different realizations, $w_1, w_2, \ldots, w_m$. The probability to attain w_j, can be described by some model and depends on the independent variable, *x*, and a model parameter. We consider two models *f* and *g* with parameters φ and θ, respectively, and *n* independent observations. Thus, for a certain *x* and θ, model *g* gives the probability that the realization for the observed variable *y* is w_j. This probability is written as $P_g(y = w_j \mid x, \theta)$. Suppose that *n* independent discrete observations are obtained, each with its own value for *x* and all with the same set of *m* different possible realizations. Then, the K-L distance of model *g* for all *n* observations together (g^n) to model *f* for the same combination (f^n) is

$$I^n(f,g) = \sum_{i=1}^{n}\sum_{j=1}^{m} P_f(y = w_j \mid x_i, \varphi)\ \ln\left(\frac{P_f(y = w_j \mid x_i, \varphi)}{P_g(y = w_j \mid x_i, \theta)}\right). \qquad (9.10)$$

The K-L distance can be calculated for fixed parameters φ and θ and a specific set of independent variables $\breve{x} = (x_1, x_2, \ldots., x_n)^T$. So if f^n is completely defined, i.e. parameter φ and the independent variable $\vec{x}$ are fixed, then the K-L distance of g'' for each possible value of its parameter θ can be calculated. Thus, the minimum K-L distance of g'' to this specific version of f'' can be determined.

The minimum K-L distance gives an indication of how easily model *g* will be preferred over model *f* if this specific version of model *f* is true. If none of the models is true, as the AIC criterion assumes, the minimum distance of model *g* to the best approximating version of model *f* can be used as an indication of how easily model *g* will be preferred over model *f*. Note that the minimum K-L distance is 0 if model *f* is nested in model *g*.

The term

$$\sum_{j=1}^{m} P_f(y = w_j \mid x_i, \varphi)\ \ln\left(\frac{P_f(y = w_j \mid x_i, \varphi)}{P_g(y = w_j \mid x_i, \theta)}\right) \tag{9.11}$$

in equation (9.10) depends on the values of x_i. Adding an extra data point will lead to an increase in the K-L distance depending on the position of the independent variable in that data point. However, if the models *f* and *g* are reasonably smooth and the frequency distribution of the independent variables is (nearly) unaffected by addition of extra data points, *I*(*f*,*g*) will increase nearly proportional to the number of observations (see e.g. Linhart and Zucchini, 1986). In other words, the minimum distance of g'' to f'' becomes proportional to *n* for large *n* if the distribution of the independent variable does not depend on the number of observations. This result is clearly only intended for large samples. For the first few data points, it might easily be possible to estimate parameter θ of *g* so that the probability $P_g(y \mid x_i, \theta) = P_f(y \mid x_i, \varphi)$ in the few data points x_i. This will generally be true for linear models if the number of data points does not exceed the number of parameters of *g*.

9.3 EXAMPLE: SEED SIZE DISCRIMINATION BY SCATTERHOARDING RODENTS

Methods

Ecological background

The dispersal phase is one of the most critical phases in plant life history. Plants have evolved a wide variety of mechanisms to have their seeds dispersed. Many nut-bearing tree species depend on scatterhoarding birds or rodents for dispersal. Such animals bury seeds as food supplies in numerous spatially scattered caches in the soil surface. This behaviour provides effective dispersal because some seeds are left to germinate and establish seedlings (Vander Wall, 1990). Non-scatterhoarded seeds, in contrast, probably die underneath the parent tree due to fungi, invertebrates and non-hoarding mammals (Jansen, 2003).

The benefits of scatterhoarding have given rise to the idea that the production of large, nutritious seeds in nut-bearing tree species has evolved in response to feeding preferences of scatterhoarding animals (Smith and Reichman, 1984). Large seeds are more nutritious and may therefore be more suitable for hoarding than smaller seeds. Indeed, several studies have shown that scatterhoarding animals disperse large seeds further than small ones (e.g. Hallwachs, 1994; Jansen *et al*., 2002; Vander Wall, 2003). However, there must be a point beyond which seeds become too large to efficiently be handled by a given animal taking into account its limited body mass and mouth width. Therefore, there should be an optimum seed size for dispersal by a given scatterhoarding animal (Jansen *et al*., 2002).

Data

Jansen (2003) experimentally studied the effect of seed size on dispersal by scatterhoarding rodents in the Nouragues rainforest reserve in French Guyana, South America (4°02'N and 52°42'W). During five consecutive years (1996-2000), numerous cafeteria plots were laid out in the territories of Red acouchy (*Myoprocta acouchy*), a cavi-like scatterhoarding rodent. Each plot contained 25 (1996-1997) or 49 (1998-2000) individually marked seeds of the canopy tree *Carapa procera* (Meliaceae), numbers that agree with the approximate daily production by average individuals of this species. Seed batches were assembled as to have seed mass within plots ranging from 3 to 60g, offering acouchies a wide choice. Seed removal from the experimental plots was monitored at days 1, 2, 4, 8, 16, 32, 64 and 128 after the start of the experiment. Moreover, the plots were also continuously monitored on video during the first day or first few days. Seeds that were eaten on the plot were included in the removed seeds, with the annotation of being eaten. See Jansen (2003) for further details.

The data set used in this paper consists of 66 plots (trials) with complete data on seed masses and seed removal. The structure of this data set allows us to apply our model selection methods at four different levels: (1) all trials lumped; (2) trials grouped in years of poor and rich fruiting; (3) trials grouped by year; and (4) individual trials. Moreover, there was variation among plots within and between years. Plots were laid out at different sites, under different forest conditions and in different rodent territories. Moreover, years differed in fruit availability. Seeds were abundant during the even years and seeds were scarce during the odd years. This distinction is important, because feeding preferences are more pronounced under seed abundance, allowing animals to be more choosy, than under conditions of scarcity (Jansen *et al.*, 2002).

The models

We modelled the probability of seed removal as a function of seed mass using a hierarchical set of models (Huisman *et al.,* 1993),

Model I: $$p(x) = \frac{1}{1+e^{a}}, \tag{9.12a}$$

Model II: $$p(x) = \frac{1}{1+e^{a+bx}}, \tag{9.12b}$$

Model III: $$p(x) = \frac{1}{1+e^{a+bx}} \cdot \frac{1}{1+e^{c}}, \tag{9.12c}$$

Model IV: $$p(x) = \frac{1}{1+e^{a+bx}} \cdot \frac{1}{1+e^{c-bx}}, \tag{9.12d}$$

Model V: $$p(x) = \frac{1}{1+e^{a+bx}} \cdot \frac{1}{1+e^{c+dx}}. \tag{9.12e}$$

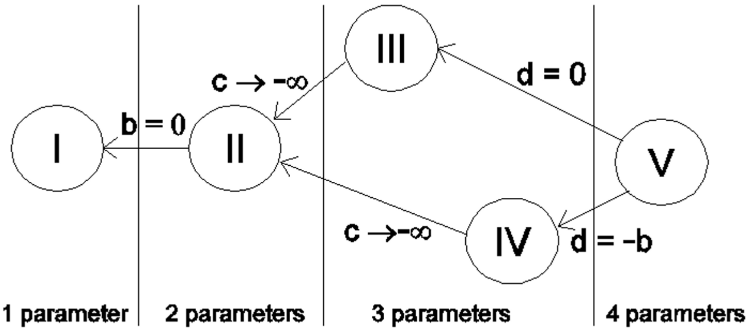


Figure 9.2. The relation between the five models fitted.

Here, *x* is the 10-logarithm of seed fresh mass and *p*(*x*) is the probability that a seed with log-mass *x* is removed. Model I describes a constant probability, independent of the seed mass. Model II describes a probability that increases gradually from 0 to 1 (or decreases from 1 to 0). Model III describes a gradual increase (or decrease) of the probability from 0 to some intermediate value. Finally, Models IV and V both describe an optimum relationship, Model IV being symmetric and Model V asymmetric. Note that the models are functions of the log of the seed mass, so that the symmetry of Model IV is in the log of the seed mass, not in the seed mass itself. Figures of the five models are given in Appendix A, and the relationship between them is shown in Figure 9.2.

We used the AICc criterion to select the model best describing the data. Each data point is denoted as (x_i,y_i), where x_i is the log of the seed mass and y_i equals 1 if the seed is removed, and 0 if it is not. For each model, the likelihood L of the data is

$$L = \prod p_\theta(x_i)^{y_i}(1 - p_\theta(x_i))^{1-y_i}, \tag{9.13}$$

and the log-likelihood ln (L) is

$$\ln(L) = \sum y_i \ln(p_\theta(x_i)) + (1 - y_i)\ln(1 - p_\theta(x_i)). \tag{9.14}$$

The value of p depends on the variable x and on the parameter value θ. The ML estimator of θ is the value of θ that maximizes the likelihood or log-likelihood.

We fitted the five models at four levels: (1) to the pooled data; (2) to poor and rich years separately; (3) to years separately; and (4) to individual trials. Furthermore, we also fitted the models as random effect models. That is, we assumed that for each trial the parameters were independent drawings from a normal distribution and estimated the mean and standard deviation of these parameters. If the model had more parameters, we assumed that the parameters were independent. Random effect models were fitted at three levels: (1) to the pooled data; (2) to poor and rich years separately; and (3) to years separately. Note that some trials showed no variation because all seeds were removed.

We also fitted some mixed effect models to the same three levels as the random effect models. Here, we assumed that the slope parameters (b and d) had fixed values. For each trial, the parameters determining the position of the model (a and c) were randomly drawn from some normal probability distribution. Finally, we fitted special versions of Models II and IV. Here, we assumed that the slope parameter b and the maximum M (Model IV only) were constant. This was done: (1) for all trials; (2) for the trials in poor and rich years separately; or (3) for the trials in one year. The position of the inflection point (Model II, $-a/b$) or top (Model IV, $(c-a)/2b$) was fitted for each trial separately. The random effect, mixed effect and special effect models were only considered with the AICc as selection criterion.

We wanted to distinguish certain basic relations between seed mass and the probability of seed removal. The five hierarchical models allow us to assess: (1) whether any such relationship exists (Model II versus Model I); (2) whether there is an upper limit < 1 to the probability of seed removal (Model III versus Model II); and (3) whether the probability of seed removal is maximal at intermediate seed mass or rather monotonously increasing or decreasing with seed mass (Models IV or V versus Model II). These relations are only of interest within the normal range of seed masses, i.e. 3-50g in our example.

Analysis

The size of the effect to be detected

First, we determined which effect we wanted to be able to detect. This rather arbitrary process lead to the following choices:

- Model II versus Model I: If the differences in log-odds at the smallest and largest seed mass is greater than 2, we wish to be able to assess an increasing (or decreasing) trend in probability. Then the slope parameter b should be less than -1.64. Also, we are not interested in assessing a monotone increase if it is an increase from almost never to very rarely (the maximum probability should be over 0.3) or an increase from in most cases to almost always (the minimum probability should be at most 0.7). For the minimum detectable slope this leads to $a \in (-0.07, 3.63)$. Figure 9.3a shows three versions of Model II, which we wish to be able to distinguish from Model I.
- Model III versus Model II: We wish to be able to assess whether the upper limit of the probability is at most 0.8 ($c > -1.39$), if that upper limit is approached sufficiently closely and if the conditions under which Model II can be distinguished from Model I are met. "Approaching the upper limit sufficiently closely" is operationalized as a log-odds distance from the upper limit of less than 0.5, i.e. if the upper limit is 0.8, the maximum probability reached in the range of possible seed masses is at least about 0.7. In Figure 9.3b some possible versions of Model III are given, which we wish to be able to distinguish from Model II.

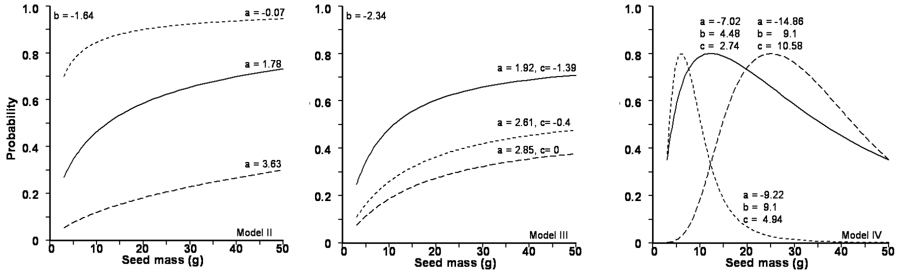


Figure 9.3. Examples of Model II (Figure 9.3a), Model III (Figure 9.3b) and Model IV (Figure 9.3c) which we wish to be able to distinguish from simpler models or models with the same number of parameters.

- Models IV or V versus Model II: We wish to be able to recognize a maximum in the probability if the differences is greater than two between the log-odds of that maximum, as attained at intermediate seed mass, and the log-odds of the probabilities at the two limits of the seed mass range. Furthermore, the top should be well within the range of the seed mass, say between 6 and 25 g. The minimum probability at both borders of the seed mass range should be less than 0.7 and at the top at least be 0.3. In Figure 9.3c three versions of Model IV are drawn. We wish to be able to distinguish these from Model II.

Levels of lumping data

Our first analysis was to compare fitting results of all trials lumped together, and the trials lumped for poor and rich years separately. Fitting Model II, for example, to trials lumped for poor and rich years separately, can be considered as fitting the model to the complete data set with an extra factor for poor or rich years. Model II then becomes

$$p(x) = \frac{1}{1 + e^{a_1 + a_2 z + (b_1 + b_2 z)x}} \tag{9.15}$$

where z is the factor for the year type ($z = 1$ in rich years, and $z = 0$ in poor years) and x is the log of the seed mass. Figure 9.4 shows the data for the probability of seeds being removed and the corresponding best fitting models. The AICc values for all models are given in Table 9.1a (first two lines).

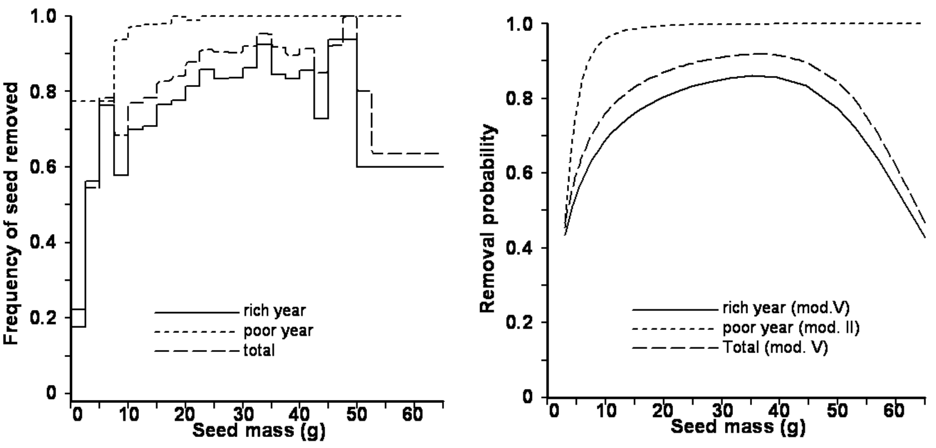


Figure 9.4. (a) The frequency of seed removal per size class, with all trials lumped and with trials lumped for rich and poor years separately. The size classes have a width of at least 5g. Size classes with less than 10 observations were lumped. (b) The corresponding models that gave the best fit.

Table 9.1. AICc values for the hierarchical set of Models I-V fitted to seed removal data with different levels of data lumping and model types. AICc values are given for the fixed effect models (a), the random effect models (b), the mixed effect models (c), and the special models (d). Levels of lumping were: all trials lumped together, trials lumped for poor and rich years separately, trials lumped for all five years separately and trials all considered separately. Note that AICc values were standardized by subtracting the smallest AICc value (the special version of model II with rich and poor years fitted separately). The smallest AICc values for each level are printed in bold.

AICc	Model I	Model II	Model III	Model IV	Model V
(a) Fixed effect					
All data together	712.5	619.5	620.2	618.5	**613.2**
Split in rich/poor years	502.1	**410.4**	413.2	412.1	411.4
Years apart	422.4	**318.5**	325.1	324.7	330.5
All trials apart	95.0	**44.0**	156.9	162.8	299.9
(b) Random effect					
All data together	175.1	72.4	76.6	**70.3**	76.7
Split in rich/poor years	146.3	38.8	39.0	**37.7**	49.3
Years apart	140.0	**39.3**	52.0	51.6	73.0
(c) Mixed effect					
All data together		89.0	**68.5**	76.1	70.5
Split in rich/poor years		49.8	**38.7**	42.0	42.0
Years apart		**46.1**	49.2	51.4	59.2
(d) Special models					
Slope/top for all data together		9.3		181.2	
Slope/top for rich and poor years		**0.0**		23.7	
Slope/top for each year		3.4		8.5	

Table 9.2. Frequency of best-fitting individual trials for five hierarchical models. All trials are considered. Numbers of trials in which all seeds were removed (no variance) are given between brackets.

Model	Number of trials for which the model is best fitting	
	Removed seeds	
	Poor year	Rich year
I	17 (17)	26 (20)
II	4	15
III	0	3
IV	0	1
V	0	0
Total	21 (17)	45 (20)

Clearly, the parameters in the poor and rich years do not have the same values. Moreover, the best fitting model differs between all years lumped together and years grouped into poor and rich years. The best model for all trials lumped shows an optimum seed size for removal. Consideration of rich and poor years separately however, reveals that an optimum seed mass for removal exists only in rich years. Poor years show rather an exponential rise to a maximum removal probability.

We then investigated how further reduction of the level of trial lumping affected the results. We extended the models with dummy variables, as in equation (9.15), to find out whether the parameters differed between years or even between individual trials. Especially the latter increased the number of parameters considerably.

The results are shown in Table 9.1a (lines 3-4). Clearly, fitting the models to trials separately yields considerably lower AICc values than fitting to lumped trials, despite the large number of extra variables involved. Seed removal is best described at the trial level by Model II, indicating higher probability of removal with increasing seed mass.

Subsequently, we determined which of the five models best fitted each trial individually. The distribution of best fitting models among trials is given in Table 9.2. Simply counting how many times each of the models turns out to give the best fit would have resulted in a constant probability per trial to be removed (Model I). However, this does not guarantee that it is the best model (Hemerik *et al.*, 2002; Hemerik and van der Hoeven, 2003). None of the five models will always be identified as the best fitting model even for data simulated with that very model (see Appendix A).

Random and mixed effect models

Another approach to account for differences between trials is to fit the five models as random effect or mixed effect models. In random effect modelling, we assume that all parameters for each trial are independent drawings from a normal distribution. In mixed effect modelling, we assumed that the parameters *a* and *c*, which determine the position of the model, were randomly drawn for each trial from some normal probability distribution, while the slope parameters (*b* and *d*) had fixed values. The resulting AICc values are given in Tables 9.1b and 9.1c, respectively.

The best fitting random effect model was Model IV with trials lumped for rich and poor years. The AICc value was even lower for this model than for the fixed effect Model II in which trials were treated separately. Figure 9.5 shows the envelopes containing 80% and 95% of the probabilities according to this model. In contrast the mixed effect models performed poorly. They never fitted better than the random effect models (Table 9.1), and rarely better than the best fitting fixed effects model. Mixed effect models had lower AICc values than fixed effect models only with trials lumped for rich and poor years.

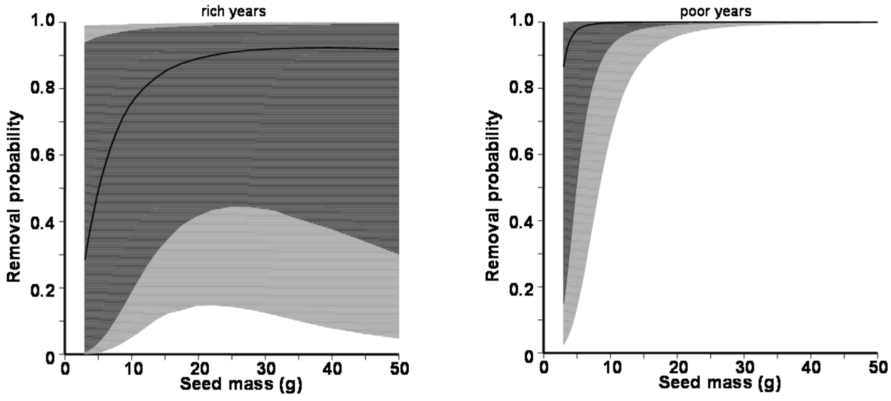


Figure 9.5. The probability of seed removal as a function of seed fresh mass according to the random effect version of Model IV. The black line indicates the removal probability with all parameters at their mean value. The dark grey envelope contains 80% of all possible realizations of parameter combinations, the grey area 95%.

Some special models

We have now seen that the best fixed effect description is obtained by fitting models to each trial separately. For the removal data, a slightly better but not very informative fit is reached by the random effect version of Model IV fitted to the data of rich and poor years separately. Our main question however, is whether there is a general relationship between seed size and the probability of seed removal (and subsequent dispersal). The two logical alternative relationships are an increase and an optimum. To investigate this, we fitted special versions of Models II and IV. Here, we assumed that the slope parameter b and the maximum M (Model IV only) were the same for all trials, while the position of the inflection point (Model II, $-a/b$) or the optimum (Model IV, $(c-a)/2b$) were fitted for each trial separately.

The AICc values for these special models are given in Table 9.1d. The lowest AICc values by far were for Model II with slope parameter b (–3.1), or even better, with slope parameter for rich ($b = –2.7$) and poor years ($b = –8.6$) separately. The models for each trial are shown in Figure 9.6 (a and b). For six out of the 45 rich trials and 17 out of the 21 poor trials, the inflection point is way below the observable range of seed mass (3 to 60 g), leading to a removal probability of nearly 1, independent of the seed mass.

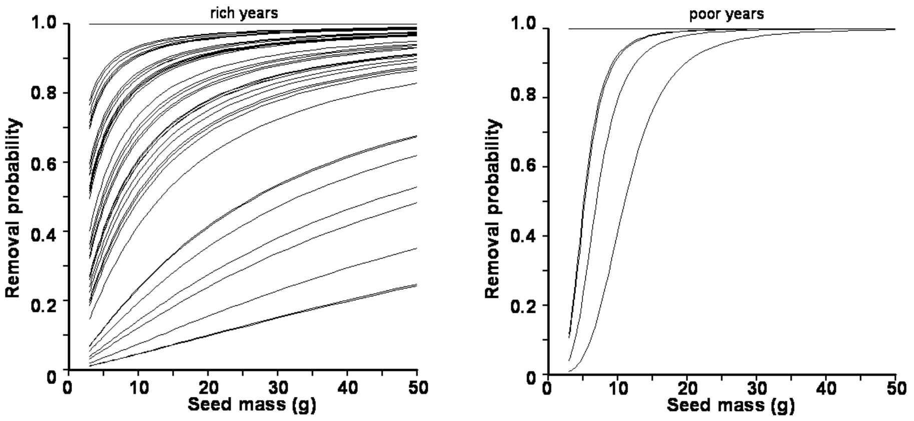


Figure 9.6. The probability of seed removal according to the special version of Model II with a fixed slope parameter for all trials in rich and in poor years respectively (Figures 9.6a and 9.6b). The upper horizontal line represents trials in which all seeds were removed, six trials in rich years (Figure 9.6a) and 17 trials in poor years (Figure 9.6b).

Information loss through fixed effect modelling with lumped data?

To investigate how much information was lost by lumping data, we calculated what percentage of the variance explained by the best fitting fixed effect model was also explained at higher levels of lumping (Burnham and Anderson, 2002). We calculated the ratio of (1) the difference between twice the log-likelihoods of an intermediate model and the simplest model (Model I with all data lumped) and (2) the difference between the best fitting model (all trials separated, Model II) and the simplest model (equation (9.16)). Let $\ln(L_b)$ be the log-likelihood of the best fitting model, $\ln(L_s)$ the log-likelihood of the most simple model and $\ln(L_i)$ the log-likelihood of the intermediate model. Then the multiple coefficient of determination, R^2 is

$$R_i^2 = \frac{2\ln(L_i) - 2\ln(L_s)}{2\ln(L_b) - 2\ln(L_s)}. \qquad (9.16)$$

R_i^2 can be interpreted as the fraction of the structural information in the best fitting model, which is also contained in the intermediate model (i).

Calculating the R_i^2 for the best fitting model gives 17%, 53% and 68% explained for complete lumping, lumping in poor and rich years, and lumping per year, respectively. These percentages indicate that lumping trials in rich and poor years conserves about 50% or more of the information. Figure 9.7

shows the best fitting models for the probability of removal with at least 50% of the information retained.

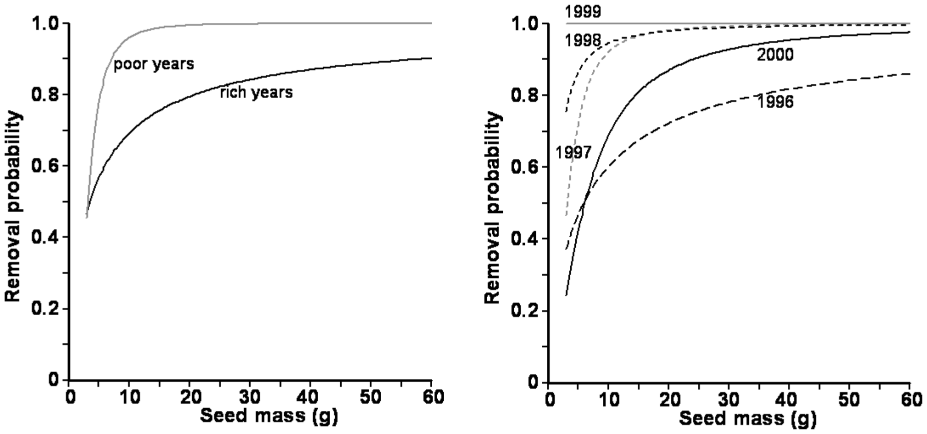


Figure 9.7. The best fitting fixed effect models with data lumped into rich and poor years (a) with data lumped per year (b). Black lines represent rich years, grey lines poor.

The Likelihood-Ratio approach for the fixed effect models

An alternative for using the AICc criterion is a stepwise test of a simpler model against a one-step more complex fixed effect model. Figure 9.8 shows all possible pathways of hypothesis testing in the case of our five models. There are two main pathways. The first (sequence 1) is to test whether the data can be split into groups. Subsequently, if further splitting is not significant and thus not allowed, models are tested in order of increasing complexity. The second (sequence 2) is to test the models in sequence of increasing complexity for the lumped data, and then, for the most complex model allowed, test whether the data can be split into groups.

Note that two alternatives are tested against Model II. Testing both at the 5% significance level will lead to a larger than 5% probability that Model II is rejected under the null hypothesis. Here, we have chosen to ignore this fact because a standard Bonferroni type correction would be far too conservative.

Both main sequences lead to the same conclusion, viz. that the best model is Model II for all trials separately (Figure 9.8). Note however, that Model II is rejected in favour of Model V ($p = 0.0058$) when tested at level (1) (all data lumped).

Here, we do not explore the LR approach further because our aim is to find the model that best describes our data, rather than to choose the simplest possible model.

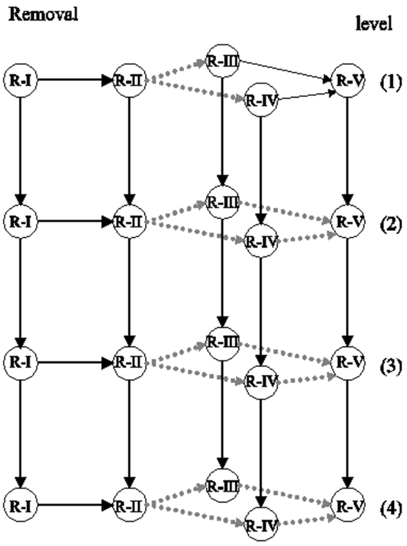


Figure 9.8. Pathways of pairwise testing for selection of the best-fitting model for seed removal by scatterhoarding rodents. Each model is tested against a one step more extended version using the χ^2 approximation of a likelihood-ratio test. Note that the extension can either be in the direction of a more complex relationship (horizontal, for example Model II instead of Model I), or in the direction of splitting the data in extra classes (vertical). Levels of lumping are: (1) all trials lumped; (2) trials lumped within rich and poor years; (3) trials lumped per year; and (4) all trials separately. Dotted grey arrow: simpler model cannot be rejected ($\alpha = 0.05$), thin arrow: the significance level between 0.05 and 0.005, intermediate arrow: the significance level between 0.005 and 0.0005, fat arrow: the significance level less then 0.0005. Left: LR test for seed removal, right: LR test for seeds being found cached.

9.4 SIMULATION

Methods

We have seen that models selected on the basis of lumped trials may differ considerably from models selected at the trial level. The best fitting fixed effect models at the trial level only indicated a simple increase of seed removal with seed mass, whereas the fixed models indicate an optimum when all data are lumped together, and the random effect models indicate an optimum both when all data are lumped together as well as when the data are split in rich and poor years. Other research has indicated that rodents discriminate against both small and large seeds, resulting in an optimum seed size for dispersal (Jansen *et al.*, 2002; Jansen, 2003). We therefore investigated whether trials with as few as 25 or 49 observations are at all suitable for accurately discriminating at the single trial level between the five models studied.

We used the five models with a wide range of parameter values. For each combination, we simulated 1000 data sets of 25 or 49 observations (seeds) with masses in a geometric series (log-masses arithmetic). The log-masses were centred and the width of the series of the log-masses was taken from -0.6109 to +0.6109. Then we determined which of the five models fitted best to each of these simulated data sets, using the AICc criterion, and counted how often the fitted model was indeed the model by which the data were generated. We used the AIC(c) as selection criterion because we wish to compare the results with the model selection in the experiments, where none of the five models will be completely true. We simulated the models with fixed parameter values because we were only interested in the frequency of correct model selection in one single trial.

We also calculated the minimum K-L distance of each simpler model to the simulated model for each parameter set for the Models II, III, IV and V. This minimum distance is considered as a measure of the distance between the simulated model and the other model. Examples of simulation and fitting results are given in Appendix A.

Simulation results

The simulations showed that 25 or even 49 seeds per plot provide too few observations per trial to accurately distinguish the five models for realistic values of the parameters. The following points emerge:

- Models II and III can be distinguished from Model I more easily if the slope parameter *b* is larger (in absolute value) and if the point of inflection is more in the centre of the data.
- Model IV can be distinguished more easily if the slope parameter *b* is large.
- Model I is chosen more often if the top of Model IV is closer to the median of the data points ([*a*–*c*] small), whereas Model II is chosen more often if the top moves farther away from the median of the data point (abs($a - c$) becomes larger).
- Models IV and V are chosen only rarely if a simpler model (or Model III) is true.
- If Model IV is true, the best fitting model is often Model III instead of Model IV. Only if the slope parameter *b* is very large, will Model IV be chosen as best model more often than Model III.

- Model I was erroneously chosen less often in the simulations with 49 observations than in those with 25 observations.

We compared the distribution of best fitting models to the experimental data (see Table 9.3) with the distribution of best fitting models to simulations with Model I both for $n = 25$ and $n = 49$. Parameter a was chosen nearest to the estimate of a for all data together. We used $a = -2$ for seed removal. The observed distribution differed significantly from the simulated one ($p = 0.01$ to 0.015, Kruskal-Wallis test).

Discrimination of the models

How different should alternative models be to be accurately discriminated? To answer this question we used the Kullback-Leibler (K-L) distance as a measure for the discrepancy between two models, and between a model and the data (see Section 9.2). The K-L distance of model A to the "real" model (the simulated one) depends on the parameter values of model A. If the real model is nested in model A, the parameters can always be chosen so that the distance is 0. In other cases, the distance will have some positive value, depending on the parameters of model A. The parameters minimising the K-L distance can be determined, and these parameters belong to the version of model A best fitting to the real model. This minimum K-L distance of model A to the real model will be indicated as the K-L distance of model A to the real model. If this distance is small, the difference between model A and the real model is small, and in model selection model A will often be preferred over the real model, i.e. the simulated one. If the distance is calculated in a limited number of data points, the distance will depend on the values of the independent variables (the seed masses) at these data points. The more observations (seeds) are used, the larger the K-L distance between models will become (see Section 9.2).

Figure 9.9 gives an impression of the K-L distance to the real model and the best approximating versions of Models I to IV for the real model being Model V.

In Section 9.3, we showed some versions of the Models II, III and IV that we wanted to distinguish from simpler models (Figure 9.3). For 25 observations, with mass geometrically spaced, the minimum K-L distances of Model I to the three examples of Model II with increasing value for a are 0.530, 1.044 and 0.530, respectively (Figure 9.3a). The minimum KL distances of Model III to Model II with increasing values of a and c are 0.011, 0.009 and 0.008, respectively (Figure 9.3b). The minimum K-L distances of Model IV to Model II with increasing parameter c (decreasing a) are 1.19, 1.88 and 1.88, respectively (Figure 9.3c). Note, however, that another choice of the seed mass distribution may dramatically affect the minimum KL distance. For example, applying the actual used distributions of seed masses in trials with 25 observations to the case of Model IV with the smallest parameter value for *c* (*a* = -7.02, *b* = 4.48 and *c* = 2.74, KL distance in case of geometric spacing 1.19) the mean of the minimum K-L distances for the 43 seed mass distributions is 0.48 (min.: 0.012, max.: 1.52).

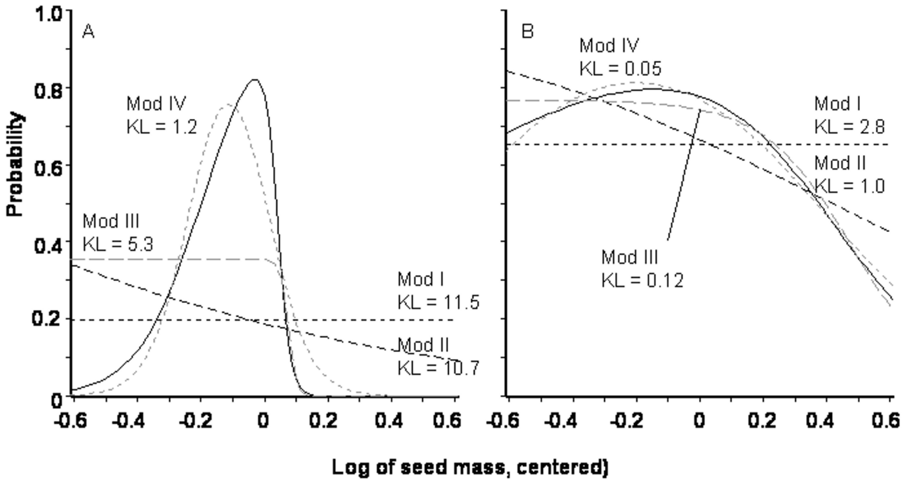


Figure 9.9. Model V (black line) and the best approximating Models I, II, III and IV. The parameter values (*a*,*b*,*c*,*d*) of Model V are in Figure 9.9A: (-2, -10, -2,50) and in Figure 9.9B:(-2, -2, -2,5). The minimum K-L distance of these models to Model V is given for 49 equidistantly spaced log-seed masses. In Figure 9.9B, Model V can be approximated reasonably well by the Models IV, III and II. In Figure 9.9A, only Model IV looks somewhat like Model V, but its distance to Model V is larger than the distance of Model II to Model V in Figure 9.9B.

The probability of the simulated model being identified as the best model increases with its difference from simpler models. For instance, if Model IV is simulated, it is chosen as the best model more often if the minimum K-L distance to the simulated model is larger for the models with less parameters (Models I and II) or with the same number of parameters (Model III). Figure 9.10 shows how the percentage of correct model choices depends on the least of all minimum K-L distances of the simpler models to the simulated model.

Figure 9.11 shows that Model I is chosen as best fitting model more often if the minimum K-L distance of Model I to the simulated model is small. For Model II, the same conclusion holds, provided that the K-L distance of Model I to the simulated model is not small too. If both Model I and Model II have a small K-L distance to the simulated model, Model I is often preferred above Model II, illustrating in fact that parsimonious models are favoured.

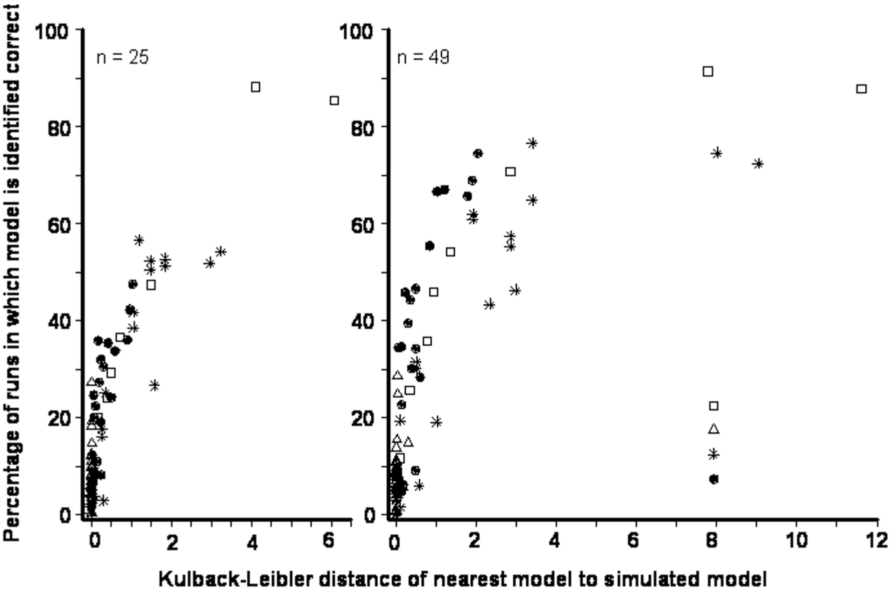


Figure 9.10. The probability of being classified as the correct model as a function of the K-L distance of the nearest simpler model to the simulated model.

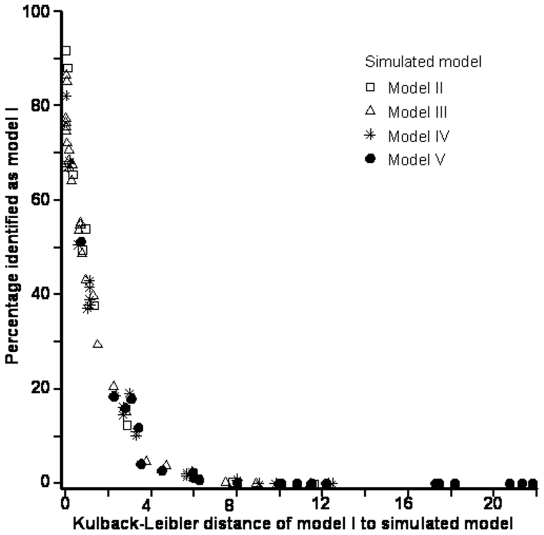


Figure 9.11. The K-L distance of Model I to the simulated model plotted against the percentage of the simulation runs in which Model I is chosen as best model using the AICc criterion. The number of observations in each simulation is 49.

Figure 9.12 shows how the frequency of choosing Model II depends on the K-L distance to Model II. In this figure and the following ones, the K-L distance is square root transformed to obtain an improved illustration of the data with a small K-L distance. The square root transformation is preferred above other possible transformations because the power of a test tends to be proportional to the square root of the number of observations, and the K-L distance is proportional to the number of observations.

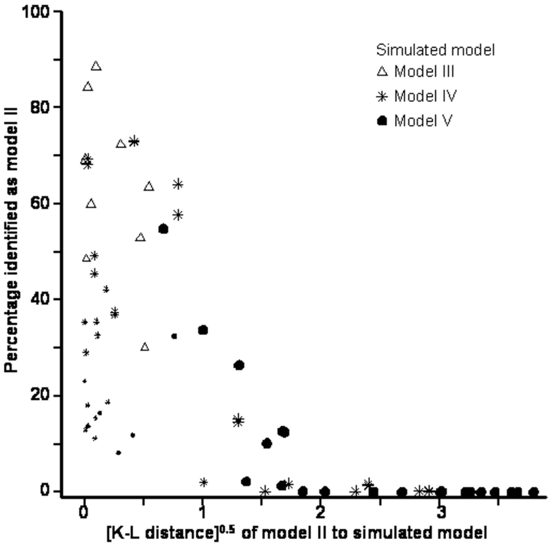


Figure 9.12. Percentage of the simulation runs in which Model II is chosen as best model (AICc criterion) as a function of the square root of the K-L distance of Model II to the simulated model. If close to the simulated model, Model I is often preferred over Model II. Markers scaled by distance of Model I to the simulated model: (1): > 2; (2): between 1 and 2; (3): between 0.5 and 1; and (4): smaller than 0.5). The number of simulated observations is 49.

Relation between K-L distance and model identification

The percentage of the simulation runs erroneously identified as Model I increases with decreasing K-L distance of the real model to Model I. Using 25 observations instead of 49 almost halves the K-L distance of Model I to the real model. The probability of choosing Model I instead of the simulated model depends only on the K-L distance between them, not on the number of simulated observations (Figure 9.13).

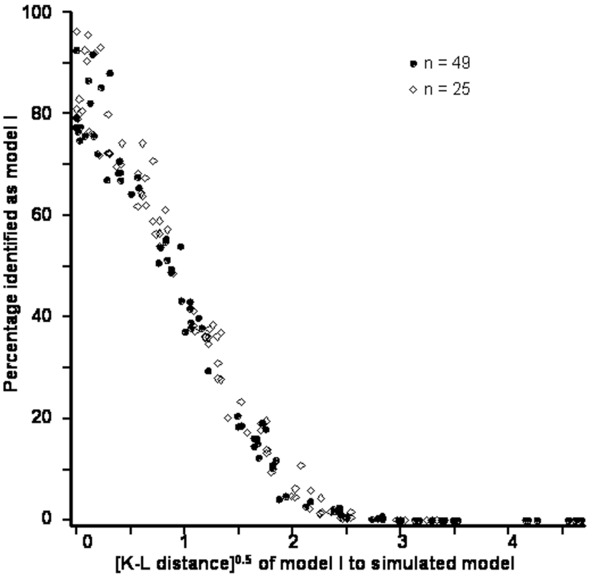


Figure 9.13. Percentage of simulation runs in which Model I is erroneously chosen as best model (AICc criterion) as a function of the square root of the K-L distance. Simulations with either 25 or 49 observations per trial

If the K-L distance for 49 observations is very small (say 0.05), halving the K-L distance will not increase the percentage of best fits of Model I considerably. On the other hand, if the K-L distance is very large, say above 10, halving the K-L distance will only slightly increase the choice for Model I. Thus, for models and parameter values with an intermediate K-L distance, a larger number of observations will reduce the K-L distance proportionally as well as the number of erroneous choices for Model I. Note that this minimum K-L distance between models is the theoretical distance, whereas the AIC calculates the observed difference for a given data set. If the minimum discrepancy to any of the simpler models is 2, the probability of choosing the extended model is about 50%.

The power of the AIC

Let us compare two nested models of which the more extended one is true but does not differ too much from the simpler one. Then twice the difference between the log-likelihoods, T, is asymptotically non-central χ^2 distributed with degrees of freedom $\nu = k_1 - k_2$ and non-centrality parameter λ (Cox and Hinkley, 1974). We will sketch some of the implications of the non-central χ^2

distribution of T for the power of the AIC. If T is $\chi'^2_{\nu,\lambda}$ distributed, the power of the AIC to choose the more extended model (or the probability that $T > 2\nu$) can be calculated. Figure 9.14a shows the power for models with only one parameter difference ($\nu = 1$), assuming $\lambda = 2M$, where M is the minimum K-L distance of the simpler model to the more extended one. This value for λ shows a good fit to the simulation results, and Akaike (1974) has proven that the non-centrality parameter λ can be approximated by $2M$. Note, however, that for the general results, the exact value of the non-centrality parameter is irrelevant. Given $\lambda = 2M$, for $\nu = 1$ a power of about 50% is reached if $M = 1$ and of about 80% if $M = 2.5$. As long as nothing is known about the specific properties of the models, the probability $P(T > 2\nu)$ for $T \sim \chi'^2_{\nu,\lambda}$ with $\lambda = 2M$ can be used as a first impression of the potential power of the AIC.

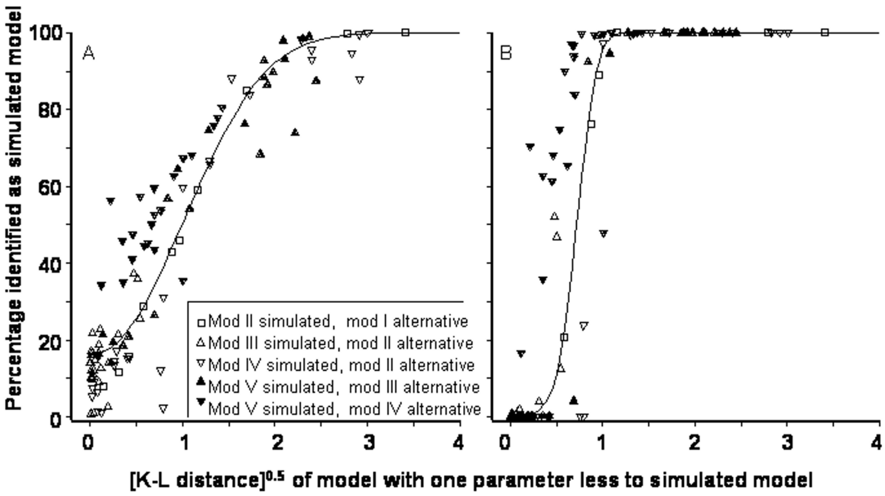


Figure 9.14. Percentage of simulation runs correctly identifying the simulated model (AICc criterion) as a function of the square root of the K-L distance to a single alternative model with one parameter less. Simulations with 49 observations per trial. The best fit is determined for separate runs (a), and for 25 runs combined (b). The drawn black lines show the theoretical prediction of the power if $\lambda = 2M$.

An increase in the number of observations will result in a more or less proportional increase of M, and thus of the non-centrality parameter λ. Sometimes, it will not be feasible to increase the number of observations in one trial with uniform conditions, for instance in the same trial. In this case, carrying out several trials, each with their own conditions, can increase the number of observations. The differing conditions in each trial may necessitate estimating a different set of model parameters for each trial. Let each trial contain n observations and let the number of trials be r. In this case, $T \sim \chi^2_{r\nu,\lambda'}$ with $\lambda' = r\lambda$, and the more extended model is chosen if $T > 2r\nu$. Figure 9.15 shows the relation between the number of trials and the power $\beta = \mathrm{P}(T > 2r\nu \mid T \sim \chi'^2_{r\nu,r\lambda})$ for λ = 0, 0.5, 1, 2, 3 and 4. For λ = 0 (the simpler model is true) and λ = 0.5 the power decreases with an increasing number of trials. Increasing the number of trials leads to an increase in power only for λ = 2, 3 and 4.

To illustrate how the K-L distance is related to the power of the AIC, we composed a set of 25 simulated trials for each simulated model by randomly drawing (with replacement) 25 runs (trials). For the 25 runs combined, the simulated model was tested against a model with one parameter less. This was repeated 1000 times. Figure 9.14b shows the relation between the percentage of correctly identified models and the K-L distance to the simpler model for a single simulation run. The theoretically expected power is also shown.

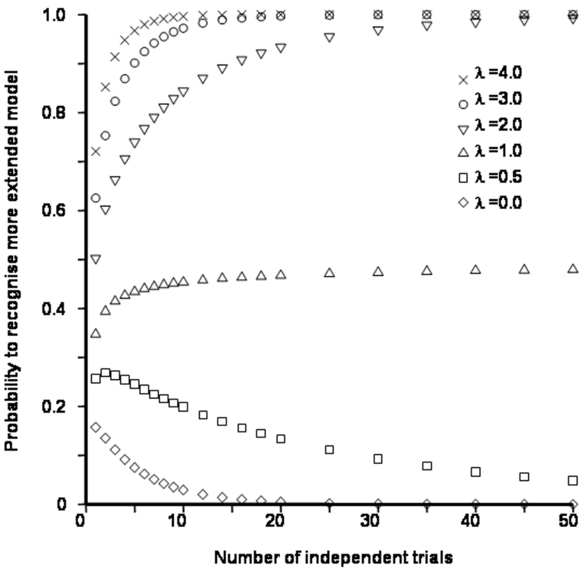


Figure 9.15. Power to correctly select a more extended model rather than a simpler one (AICc criterion) as a function of the number of independent trials. The difference between the maximum of the log-likelihoods for each trial is non-central χ^2 distributed with non-centrality parameter λ and degrees of freedom ν.

9.5 CONCLUSIONS AND DISCUSSION

Experimenters must sometimes find a balance between what is statistically desirable and what is biologically realistic and feasible. One example of such a situation is formed by ecological experiments that consist of many replicate trials, each containing a limited number of observations. This approach may be the only way to obtain sufficient data, or be more informative for answering the ecologically relevant question. However, the variation between trials should be accounted for statistically. We have shown that rigorous application of this principle may make it impossible to distinguish any pattern present in ecological data.

In ANOVA types of problems it is a generally accepted rule that two groups can only be combined if they do not differ significantly. We applied this rule also for combining data in model selection and model parameter estimation. Using fixed effect models, the number of parameters is proportional to the number of fits of the model to separate groups of data sets. So if each trial is considered separately, the number of parameters is proportional to the number of trials. Before combining data in larger groups, it should be tested whether the fit of any of the models to the combined data is better than the best fit of any of the models to the separate data. To identify the best fitting model we used the AICc (a non-specific robust adaption of the AIC to large number of parameters or small data sets) as selection criterion.

If lumping of data is not allowed, we showed that the identifiability of a true model against simpler alternatives only increases with the number of replicate trials if the minimum K-L distance between the two models is sufficiently large (> 1/2) in each trial. That is, using the AIC(c) the power of model identification only increases by increasing the number of separate trials if the power in each trial is sufficiently large. Note that if instead of the AICc the LR test is used, the power would increase with an increasing number of trials. For a large difference in the number of parameters the LR criterion is less conservative than the AIC, leading to the situation that the AIC criterion may prefer the simpler model even though the LR test suggests that the simpler model is unlikely when compared to the extended one. The AIC leads to the model giving the best prediction and the LR test to the most parsimonious model being not too unlikely. The BIC is highly biased against the more extended models, whereas its claimed consistency is only relevant for very large numbers, and therefore totally uninteresting for most biological experiments with a relatively small number of observations.

To test whether data of several trials can be lumped, the fit of fixed and random effect versions of the models at each level of data lumping should be compared. At a given level of data lumping, the parameters of the fixed effect models have a fixed value, whereas the parameters of the random effect model are drawn for each trial from a probability distribution with fixed parameters at that level of lumping. The random effect models are the analogue of random effect models in ANOVA. In our example, the random effect model with data lumped in rich and poor years appears to fit best.

Next to fixed and random effect models, the fit of some mixed effect models might also be considered. For instance, we fitted models with a constant value for the slope parameter and trial-specific values for the other parameter(s). The fit of such mixed models have to be compared with the fit of models with parameters at only one level of data lumping. Note however, that such a mixed model may seem to show a general trend, for instance a relation with a maximum, but that this might imply for some trials a uniformly increasing trend and for others an uniformly decreasing trend, depending on which part of the curve is observed. The supposed general trend would in such a case be based on extrapolation and thus ecologically irrelevant.

Before starting an experiment it is sensible to investigate whether the statistical power is sufficient to answer the research questions. If one of the aims is to distinguish different models, the minimum K-L distance (MKLD) between alternatives can be used as an indication for the power of the AIC(c) decision procedure. To calculate the MKLD to a model, the parameters of that model and the values (or distribution) of the independent variables have to be specified. The parameter values can be based on previous experience, and can also reflect the minimum deviation of the complex model from the simpler one for which the complex model still merits consideration. The values of the independent variable(s) may be chosen to maximize the MKLD to the more complex model. The MKLD should be sufficiently large to be able to distinguish the more complex model from the simpler one. Increasing the number of observations by increasing the number of separate trials will only increase the power if the power in each separate trial is already sufficiently high.

The method described above can be summarized as:

1. Select models describing the hypothetical relationships that are of ecological importance.

2. Choose for each model realistic parameter ranges and decide for which parameter range a model should be distinguishable from the simpler alternatives.

3. Consider the range of independent model variables applicable in each experiment.

4. Calculate the MKLD of simpler models to more extended ones for the realistic parameter values (point 2) and independent variables (point 3).

5. If the MKLD is less than 1/2, models are not distinguishable without data lumping. In advance answer the question whether data lumping might be acceptable. If MKLD is sufficiently larger than 1/2, increasing the number of trials will increase the power of model identification. Calculate the number of replicate trials necessary to reach the desired power.

6. Perform the experiments.

7. Fit models to data of each trial and to the lumped data. Use fixed and random effect models. If necessary, also use mixed effect models.

8. Select the model with the lowest AIC(c), in this way both selecting the model type and the level of lumping. Remember that small differences between AIC(c)'s are not very informative.

Although an extensive body of literature exists on model selection, no guidelines are given on how to deal with data collected in a large set of separate trials as in our example. To our knowledge this is the first attempt to provide guidelines to facilitate the use of model selection methods in ecological applications and incorporate the intended model selection into the experimental set-up.

ACKNOWLEDGEMENTS

We thank Elizabeth van Ast for correction of the English, and Johan Grasman and two anonymous reviewers for critical comments on an earlier version of the manuscript. P.A.J. was supported by the Netherlands Foundation for the Advancement of Tropical Research (NWO-WOTRO grant 84-407).

REFERENCES

Akaike, H. (1974). A new look at the statistical model identification. IEEE Transactions on Automatic Control 19: 716-723.

Borowiak, D. S. (1989). Model Discrimination for Non-Linear Regression Models. Marcel Dekker Inc., New York.

Burnham, K. P. and D. R. Anderson (2002). Model Selection and Inference. A Practical Information-Theoretic Approach. Springer, New York.

Cox, D. R. and D. V. Hinkley (1974). Theoretical Statistics. Chapman and Hall, London.

Hallwachs, W. (1994). The Clumsy Dance between Agoutis and Plants: Scatterhoarding by Costa Rican Dry Forest Agoutis (*Dasyprocta punctata*: Dasyproctidae: Rodentia). PhD thesis, Cornell University, New York.

Hemerik, L. and N. van der Hoeven (2003). Egg distributions of solitary parasitoids revisited. Entomologia Experimentalis et Applicata 107: 81-86.

Hemerik, L., N. van der Hoeven and J. J. M. Van Alphen (2002). Egg distributions and the information a solitary parasitoid has and uses for its oviposition decisions. Acta Biotheoretica 50: 167-188 .

Hilborn, R. and M. Mangel (1997). The Ecological Detective. Confronting Models with Data. Princeton University Press, Princeton.

Huisman, J., H. Olff and L. F. M. Fresco (1993). A hierarchical set of models for species response analysis. Journal of Vegetation Science 4: 37-46.

Hurvich, C. M. and C.-L. Tsai (1989). Regression and time series model selection in small samples. Biometrika 76: 297-307.

Jansen, P. A., M. Bartholomeus, F. Bongers, J. A. Elzinga, J. Den Ouden and S. E. Van Wieren (2002). The role of seed size in dispersal by a scatter-hoarding rodent. pp. 209-225. In: Levey, D., W. R. Silva, and M. Galetti (Eds) Seed Dispersal and Frugivory: Ecology, Evolution and Conservation. CAB International, Wallingford.

Jansen, P. A. (2003). Scatterhoarding and Tree Regeneration. Ecology of Nut Dispersal in a Neotropical Rainforest. PhD thesis, Wageningen University, The Netherlands.

Kullback, S. and R. A. Leibler (1951). On information and sufficiency. Annals of Mathematical Statistics 22: 79-86.

Linhart, H. and W. Zucchini (1986). Model Selection. John Wiley and Sons, New York.

Schwarz, G. (1978). Estimating the dimension of a model. Annals of Statistics 6: 461-464.

Smith, C. C. and O. J. Reichman (1984). The evolution of food caching by birds and mammals. Annual Review of Ecology and Systematics 15: 329-351.

Umbach, D. M. and A. J. Wilcox (1996). A technique for measuring epidemiologically useful features of birthweight distributions. Statistics in Medicine 15: 1333-1348.

Van der Hoeven, N., (in press). A general method to calculate the power of likelihood-ratio based tests to choose between two nested models. Journal of Statistical Planning and Inference.

Vander Wall, S. B. (1990). Food Hoarding in Animals. Chicago University Press, Chicago.

Vander Wall, S. B. (2003). Effects of seed size of wind-dispersed pines (Pinus) on secondary seed dispersal and the caching behavior of rodents. Oikos 100: 25-34.

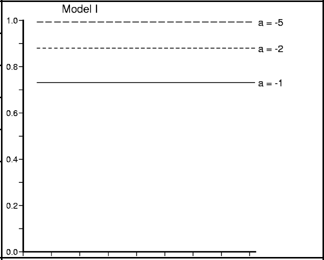
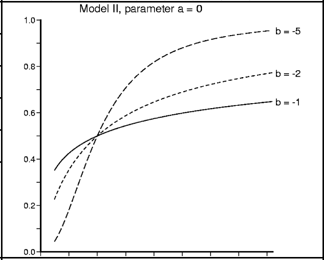
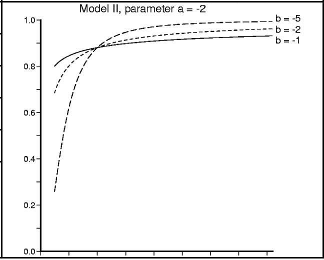
Nelly van der Hoeven,
Department of Theoretical Biology, Leiden University

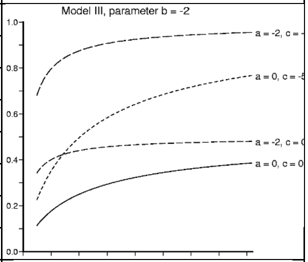
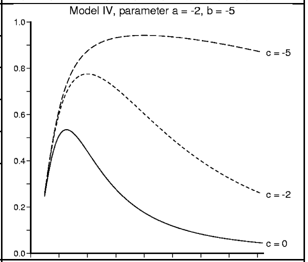
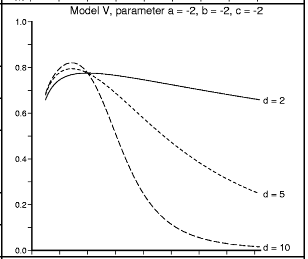
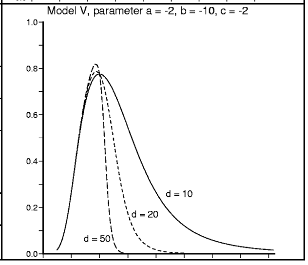
Lia Hemerik
Biometris, Department of Mathematical and Statistical Methods,
Wageningen University

Patrick A. Jansen
Forest Ecology and Forest Management Group, Wageningen University

APPENDIX A

Results of simulations with Models I to V (see equation (9.13)) and different sets of parameter values (1000 runs per set). Numbers indicate the number of runs at which each of the five models was selected as the best-fitting according to the AICc criterion. The best fitting model if all 1000 runs are combined is also given. The models for each parameter combination are illustrated in figures.

Number of observations	Model	Parameters				Model number					Best model, all simulations together	Figures with examples of model
						I	II	III	IV	V		
		a	*b*	*c*	*d*	Number of runs in which model was chosen as best.						
25	I	-1				809	117	45	15	14	1	
25	I	-2				794	158	35	10	3	1	
25	I	-5				963	37	0	0	0	1	
49	I	-1				773	128	68	17	14	1	
49	I	-2				792	136	50	8	14	1	
49	I	-5				925	74	0	1	0	1	
25	II	0	-1			620	242	100	16	22	1	
25	II	0	-2			376	474	95	30	25	2	
25	II	0	-5			4	854	109	19	14	2	
25	II	-2	-1			742	200	29	19	10	1	
25	II	-2	-2			572	366	36	19	7	2	
25	II	-2	-5			45	882	59	10	4	2	
49	II	0	-1			494	358	84	37	27	2	
49	II	0	-2			123	707	92	32	46	2	
49	II	0	-5			0	878	58	28	36	2	
49	II	-2	-1			654	257	46	27	16	1	
49	II	-2	-2			378	541	45	23	13	2	
49	II	-2	-5			3	914	50	27	6	2	

<table>
<tr><th rowspan="3">Number of seeds</th><th rowspan="3">Model</th><th colspan="4" rowspan="2">Parameters</th><th colspan="5">Model number</th><th rowspan="3">Best model, all simulations together</th><th rowspan="3">Figures with examples of model</th></tr>
<tr><th>I</th><th>II</th><th>III</th><th>IV</th><th>V</th></tr>
<tr><th><i>a</i></th><th><i>b</i></th><th><i>c</i></th><th><i>d</i></th><th colspan="5">Number of runs in which model was chosen as best</th></tr>
<tr><td>25</td><td>III</td><td>0</td><td>-2</td><td>0</td><td></td><td>588</td><td>221</td><td>124</td><td>36</td><td>31</td><td>1</td><td rowspan="8">
</td></tr>
<tr><td>25</td><td>III</td><td>0</td><td>-2</td><td>-5</td><td></td><td>360</td><td>488</td><td>111</td><td>28</td><td>13</td><td>2</td></tr>
<tr><td>25</td><td>III</td><td>-2</td><td>-2</td><td>0</td><td></td><td>721</td><td>153</td><td>82</td><td>16</td><td>28</td><td>1</td></tr>
<tr><td>25</td><td>III</td><td>-2</td><td>-2</td><td>-5</td><td></td><td>611</td><td>321</td><td>45</td><td>13</td><td>10</td><td>2</td></tr>
<tr><td>49</td><td>III</td><td>0</td><td>-2</td><td>0</td><td></td><td>432</td><td>355</td><td>138</td><td>31</td><td>44</td><td>2</td></tr>
<tr><td>49</td><td>III</td><td>0</td><td>-2</td><td>-5</td><td></td><td>151</td><td>692</td><td>92</td><td>28</td><td>37</td><td>2</td></tr>
<tr><td>49</td><td>III</td><td>-2</td><td>-2</td><td>0</td><td></td><td>669</td><td>154</td><td>107</td><td>27</td><td>43</td><td>1</td></tr>
<tr><td>49</td><td>III</td><td>-2</td><td>-2</td><td>-5</td><td></td><td>398</td><td>486</td><td>70</td><td>30</td><td>16</td><td>2</td></tr>
<tr><td>25</td><td>IV</td><td>-2</td><td>-5</td><td>0</td><td></td><td>310</td><td>118</td><td>267</td><td>162</td><td>143</td><td>3</td><td rowspan="6">
</td></tr>
<tr><td>25</td><td>IV</td><td>-2</td><td>-5</td><td>-2</td><td></td><td>384</td><td>39</td><td>174</td><td>266</td><td>137</td><td>3</td></tr>
<tr><td>25</td><td>IV</td><td>-2</td><td>-5</td><td>-5</td><td></td><td>139</td><td>601</td><td>110</td><td>97</td><td>53</td><td>2</td></tr>
<tr><td>49</td><td>IV</td><td>-2</td><td>-5</td><td>0</td><td></td><td>109</td><td>146</td><td>298</td><td>301</td><td>146</td><td>3</td></tr>
<tr><td>49</td><td>IV</td><td>-2</td><td>-5</td><td>-2</td><td></td><td>191</td><td>17</td><td>178</td><td>462</td><td>152</td><td>4</td></tr>
<tr><td>49</td><td>IV</td><td>-2</td><td>-5</td><td>-5</td><td></td><td>17</td><td>577</td><td>125</td><td>194</td><td>87</td><td>2</td></tr>
<tr><td>25</td><td>V (IV)</td><td>-2</td><td>-2</td><td>-2</td><td>2</td><td>722</td><td>130</td><td>81</td><td>38</td><td>29</td><td>1</td><td rowspan="6">
</td></tr>
<tr><td>25</td><td>V</td><td>-2</td><td>-2</td><td>-2</td><td>5</td><td>347</td><td>283</td><td>229</td><td>90</td><td>51</td><td>2</td></tr>
<tr><td>25</td><td>V</td><td>-2</td><td>-2</td><td>-2</td><td>10</td><td>16</td><td>299</td><td>422</td><td>178</td><td>85</td><td>3</td></tr>
<tr><td>25</td><td>V (IV)</td><td>-2</td><td>-10</td><td>-2</td><td>10</td><td>59</td><td>0</td><td>39</td><td>542</td><td>360</td><td>4</td></tr>
<tr><td>25</td><td>V</td><td>-2</td><td>-10</td><td>-2</td><td>20</td><td>17</td><td>1</td><td>44</td><td>579</td><td>359</td><td>4</td></tr>
<tr><td>25</td><td>V</td><td>-2</td><td>-10</td><td>-2</td><td>50</td><td>22</td><td>2</td><td>54</td><td>584</td><td>338</td><td>4</td></tr>
<tr><td>49</td><td>V (IV)</td><td>-2</td><td>-2</td><td>-2</td><td>2</td><td>684</td><td>119</td><td>115</td><td>52</td><td>30</td><td>1</td><td rowspan="6">
</td></tr>
<tr><td>49</td><td>V</td><td>-2</td><td>-2</td><td>-2</td><td>5</td><td>160</td><td>337</td><td>305</td><td>150</td><td>48</td><td>3</td></tr>
<tr><td>49</td><td>V</td><td>-2</td><td>-2</td><td>-2</td><td>10</td><td>0</td><td>124</td><td>554</td><td>261</td><td>61</td><td>3</td></tr>
<tr><td>49</td><td>V (IV)</td><td>-2</td><td>-10</td><td>-2</td><td>10</td><td>0</td><td>0</td><td>10</td><td>723</td><td>267</td><td>4</td></tr>
<tr><td>49</td><td>V</td><td>-2</td><td>-10</td><td>-2</td><td>20</td><td>0</td><td>0</td><td>5</td><td>552</td><td>443</td><td>5</td></tr>
<tr><td>49</td><td>V</td><td>-2</td><td>-10</td><td>-2</td><td>50</td><td>0</td><td>0</td><td>11</td><td>318</td><td>671</td><td>5</td></tr>
</table>