Estimating species accumulation functions using birth processses

Eloísa Díaz-Francés* and J.K. Lindsey†

*Department of Probability and Statistics, CIMAT, Guanajuato, Mexico

†Biostatistics, Limburgs Universitair Centrum, Diepenbeek, Belgium

Abstract

Soberón and Llorente (1993) present a number of different birth models for the processes by which species are accumulated in studies to estimate the total number of different species in an area. They use least squares to estimate the accumulation curve, and hence the asymptotic number of species. This approach suffers from at least two difficulties: the variance of the number of accumulated species depends on the mean and the successive values of accumulated numbers are not independent, calling for an autoregressive process.

An alternative is to fit the birth model directly to the numbers of new species, perhaps as a nonhomogeneous Poisson process. Another possibility is to use a multinomial distribution. Likelihood procedures are then used to provide intervals of precision for the estimates of total numbers of species.

These approaches are compared in their application to estimation of the total number of species of bats around the Chajul Biological Station in southern Mexico.

KEYWORDS: Accumulation function, biodiversity, birth process, collecting function, likelihood function.

1 Introduction

With the increasing concern about the reduction in biodiversity of animal and plant species on the planet, effective methods of estimating numbers of different species are becoming increasingly important. A species accumulation function, $S(t)$, relates the total number of species discovered to the effort, t, expended to find them. In most situations, one might expect the number of new species per unit effort to decrease as the number already discovered grows, a form of birth process. In other words, the accumulation function is expected to reach an asymptote. One goal is often to estimate this asymptote, that is, the total number of different species in some geographical area.

Biologists have presented a number of different models for the processes by which species are accumulated in such studies (Soberón and Llorente, 1993). These may depend on the way in which the study has been conducted and on how close they are to enumerating all possible species. These

models are developed in terms of a collecting function, $\lambda(j,t)$, that describes the probability of adding new species to an accumulating list as it depends on the number, j, already on the list and on the effort expended. (Statisticians call this an intensity or risk function.) For a given model, the corresponding accumulation function is obtained from the collecting function by solving differential equations. In the literature, these have been solved for the first two moments of the function which is then fitted to the data using least squares.

Several statistical problems arise in using such an approach. The functions are approximated by their first two moments. This may not be sufficiently accurate for the small number of species involved in many studies. Relatively small numbers of discrete counts are being approximated by a continuous normal distribution. The accumulating counts can be expected to be highly correlated. Only the new species after each effort expended are providing new information so that the information contained in the earlier species recorded is being reused many times, perhaps yielding a false measure of unduly high precision.

2 Collecting function models

In a birth model, the collecting (intensity) function will depend in some way on the number of different species previously recorded. The first model suggested by Soberón and Llorente (1993) was the exponential model,

$$\lambda(j) = \alpha - \beta j, \qquad j = 0, 1, \ldots, \lfloor \alpha/\beta \rfloor$$

Because $\lambda(j) \geq 0$, new species will only be observed as long as $j \leq \alpha/\beta$. If this equation is rewritten as

$$\lambda(j) = \beta(N - j)$$

where $N = \alpha/\beta$, we see that the collecting function is proportional to the number of so far unrecorded species. The accumulation function for this model is

$$S(t) = \frac{\alpha}{\beta}\left(1 - \mathrm{e}^{-\beta t}\right)$$

with the asymptote given by α/β and the variance going to zero as t approaches infinity, so that $N = \lfloor \alpha/\beta \rfloor + 1$ is the estimated total number of species.

They called the second of their models logarithmic:

$$\lambda(j) = \alpha \mathrm{e}^{-\beta j}$$

The accumulation function of this model does not reach an asymptote as the collecting function only becomes zero with infinite accumulation. They believed that this could correspond to the accumulation function,

$$S(t) = \frac{1}{\delta}\log(1 + \delta\alpha t)$$

where $\delta = 1 - \mathrm{e}^{-\beta}$. However, Díaz-Francés and Gorostiza (2000) have shown that the latter function does not correspond to the former, taking a very complex form. These authors introduce instead a nonhomogeneous birth process,

$$\lambda(j,t) = \frac{\alpha}{1+\delta\alpha t} + \frac{\gamma}{\delta}\log(1+\delta\alpha t) - \gamma t$$

corresponding to this accumulation function. Again, this does not reach an asymptote.

Soberón and Llorente (1993) finally suggested one nonhomogeneous birth process, the Michaëlis-Menton or Clench (1979) model,

$$\lambda(j,t) = \alpha + \alpha\left[\frac{\beta t}{1+\beta t}\right]^2 - 2\beta j$$

with accumulation function

$$S(t) = \frac{\alpha t}{1+\beta t}$$

Again, $N = \lfloor \alpha/\beta \rfloor + 1$ is the estimated total number of species. The names of these models are obviously derived from their accumulation functions.

Nakamura and Peraza (1998) develop a type of nonhomogeneous Poisson process whereby the collecting function involves a beta distribution,

$$\lambda(t) = \frac{B(\alpha+1, \beta+t-1)}{B(\alpha,\beta)}N$$

where $B(\cdot,\cdot)$ is the beta function. A simplification of this would be to have a constant probability of capture for all new species, a homogeneous Poisson process.

Note that a process may be nonhomogeneous in time if, for example, the biologists making the observations acquire experience over time.

3 Approaches to estimating accumulation functions

Let us first consider the usual method of estimating the accumulation function, by applying least squares to the accumulating numbers, as a function of effort, with mean given by some $S(t)$. A complication with this procedure is that the variance is not constant but is a function of the mean accumulation function, containing only parameters in that function. Hence, there is a very strict relationship between the mean and variance (Díaz-Francés and Gorostiza, 2000) that may not hold empirically. As well, the observations being modelled are the accumulated counts so that successive counts can be expected to be highly correlated, calling for an autoregressive process.

The high autocorrelation among successive accumulated counts indicates that successive differences might be more appropriately modelled. Indeed, these are just the numbers of new species recorded after each effort and their mean is described by some collecting function, $\lambda(j,t)$. This will be the mean of a nonhomogeneous Poisson process that can easily be fitted by Poisson regression (Lindsey, 1995). The probability of the observed data will then be

$$\Pr(y_1,\ldots,y_m) = \prod_{t=1}^{m} \frac{\mathrm{e}^{-\lambda(j,t)\Delta t}[\lambda(j,t)\Delta t]^{y_t}}{y_t!}$$

where y_t is the observed number of new species recorded after total effort, t, and obtained with Δt new effort expenditure since the previous recorded count. In this way, the mean-variance relationship is automatically accounted for and need not be explicitly modelled.

A third approach is to extend the multinomial model of Nakamura and Peraza (1998) to encompass these collecting functions (and any other that might be thought useful). The probability of the observed data will now be

$$\Pr(y_1, \ldots, y_m) = \frac{N!}{\prod_i y_i!(N - \sum_i y_i)!} \prod_{t=1}^{m} \left(\frac{\lambda(j,t)\Delta t}{N} \right)^{y_t} \left(1 - \sum_i \frac{\lambda(j,i)\Delta i}{N} \right)^{N - \sum_i y_i}$$

Note that the parameter, N, now refers directly to the total number of species. It is no longer estimated from the parameters of the collecting function.

One interpretation of this multinomial model is that it conditions the Poisson model on there being a fixed, but unknown, total number of species, N. All of the parameters in this model can also be estimated relatively easily, although standard software is not available.

Two important advantages arise from using not only the appropriate collecting function for a particular situation, but also an appropriate stochastic model. This will provide intervals of precision for the parameters, and specifically for the total number of species when this parameter is present in the model. It will also allow comparative evaluation of the goodness of fit to help in judging which model might provide the better predictions of total species numbers.

4 Examples

4.1 Chajul bats

One of the data sets analyzed by Soberón and Llorente (1993), and reconsidered by both Nakamura and Peraza (1998) and Díaz-Francés and Gorostiza (2000), involves a list of bat species recorded at the Chajul Biological Station in the Lacandon rain forest in southern Mexico, captured using mist nets at several locations near the station. A total of 50 species was recorded with 49 nights of effort. For these data, the biologists involved believe that the exponential model should be appropriate.

From Table 1, we can see that the estimated numbers can vary greatly depending on the model employed. However, the likelihood function, perhaps appropriately penalized for the number of estimated parameters, provides a means to let us compare the predictive ability of the various models, at least for predicting the observed counts. However, this is only true for models using the same counts, either cumulated (the normal models) or new species (the Poisson and multinomial models). Thus, the exponential model estimated using independent normal distributions has a penalized negative log likelihood (AIC) of 114.3 whereas that involving an autoregression (AR) has 82.2, strongly indicating the presence of serial dependence. The results are summarized in Table 2. The exponential model appears to provide a good fit to these data, as compared to the

Table 1: Point estimates of total species numbers for the Chajul bat study using various procedures.

	Stochastic model			
Collecting function	Indep. Normal	Normal AR	Poisson	Multinomial
Exponential	54.5	50.8	53.9	51.4
Clench	83.0	83.9	80.3	50.0
Beta			55.7	53.7

Table 2: Fits of models for the Chajul bat study using various procedures as measured by the negative log likelihood (AIC). The values in the first two columns are not comparable to those in the last two columns.

	Stochastic model			
Collecting function	Indep. Normal	Normal AR	Poisson	Multinomial
Exponential	114.3	82.2	64.7	63.2
Clench	122.2	83.3	65.3	63.1
Beta			67.6	66.3

others. Further information is provided by the profile likelihoods, indicating the precision of the estimates of the total number of species. Those for the multinomial models are shown in Figure 1. Although the exponential and Clench models have very similar fit, the shapes of the likelihood curves are very different. The fitted curves of the accumulation function for these two models are plotted in Figure 2.

4.2 Pakitza butterflies

A second data set analyzed by all of the same authors involves a list of butterflies obtained by 200 person-hours of collecting during September, 1989, in the Pakitza Biological Station in the Parque Nacional Manú, Madre de Dios, Peru. The biologists argue that the logarithmic model should be most adequate for extrapolation because of the size of the area covered, the complexity of the species, the fact that the list was still far from being complete, and the yearly fluctuations undergone by many tropical butterflies.

The fits of various models are displayed in Table 3. Those using the normal distribution have been fitted with constant variance as those with variance depending on the mean were much worse. Notice that an autoregression is not necessary for these data. Those models with an asymptote either did not converge or yielded very large estimates for the asymptote.

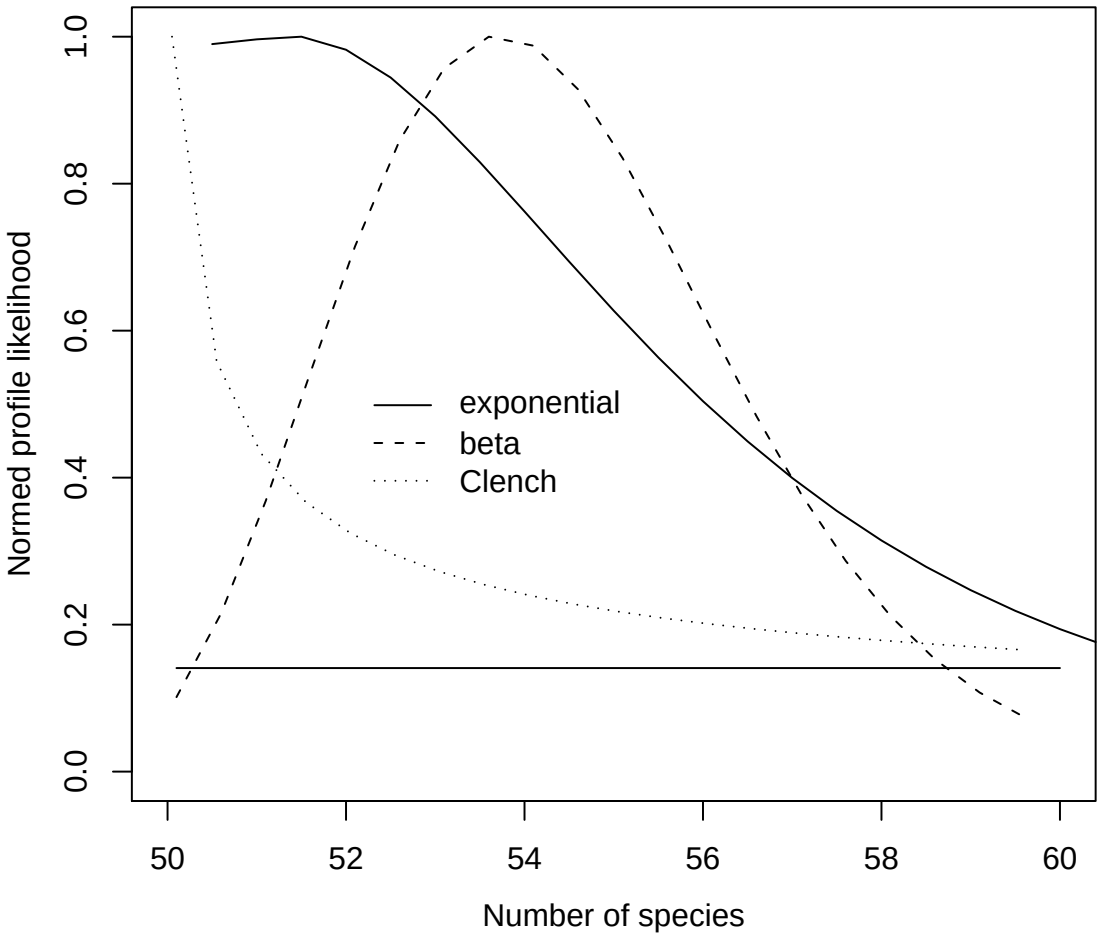

Figure 1: Profile likelihoods for the total number of species from the three multinomial models fitted to Chajul bat data. The solid horizontal line indicates the 95% confidence interval.

Table 3: Fits of models for the Pakitza butterfly study using various procedures as measured by the negative log likelihood (AIC). The values in the first two columns are not comparable to those in the last two columns.

	Stochastic model			
Collecting function	Indep. Normal	Normal AR	Poisson	Multinomial
Exponential	62.7	63.4	—	160.4
Logarithmic	114.4	115.4	74.1	52.1
Clench	59.7	60.7	64.7	153.5
Beta			64.7	121.2

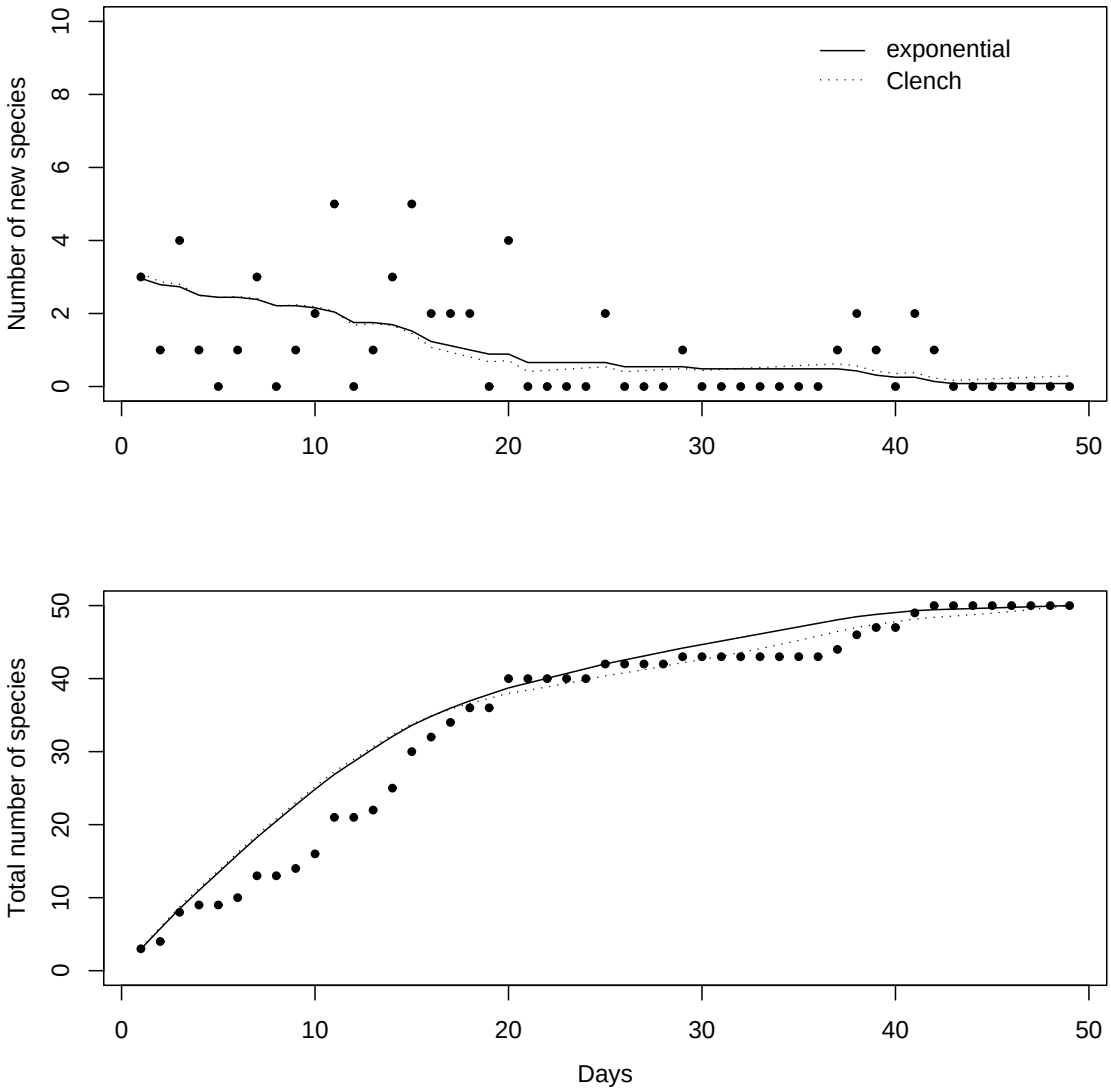

Figure 2: Fitted collecting (top) and accumulation (bottom) curves from the exponential and Clench multinomial models for the Chajul bat data.

Table 4: Point estimates of total species numbers for the Chiapas spider study using various procedures. (First panel: Hamburgo; second panel: Irlanda)

	Stochastic model			
Collecting function	Indep. Normal	Normal AR	Poisson	Multinomial
Exponential	64.2	44.5	46.9	46.0
Clench	59.3	57.1	27.7	46.1
Beta			184.0	58.5
Exponential	43.9	46.0	49.3	47.4
Clench	58.6	62.9	28.5	45.1
Beta			87.8	55.8

Table 5: Fits of models for the Chiapas spider study using various procedures as measured by the negative log likelihood (AIC). The values in the first two columns are not comparable to those in the last two columns. (First panel: Hamburgo; second panel: Irlanda)

	Stochastic model			
Collecting function	Indep. Normal	Normal AR	Poisson	Multinomial
Exponential	55.1	41.3	36.4	34.3
Clench	49.0	38.7	35.7	35.9
Beta			37.7	37.3
Exponential	47.5	39.2	36.5	35.2
Clench	45.0	39.1	36.1	34.1
Beta			37.7	37.1

4.3 Hamburgo and Irlanda spiders

Species of weaver spiders were counted in two coffee orchards, Hamburgo and Irlanda, in the Soconusco region of Chiapas, southern Mexico. These are neighbouring plantations at an elevation of about 900–990m. Irlanda is organically maintained with shade provided by what remains of the original forest; the coffee shrubs were planted along constant contours of altitude. On the other hand, Hamburgo is located in completely cleared land, has the coffee planted in straight rows, and uses chemical fertilization and weeding.

The fits of various models are displayed in Table 5. The profile likelihoods for the multinomial models are shown in Figures 3 and 4 and the accumulation curves are plotted in Figures 5 and 6.

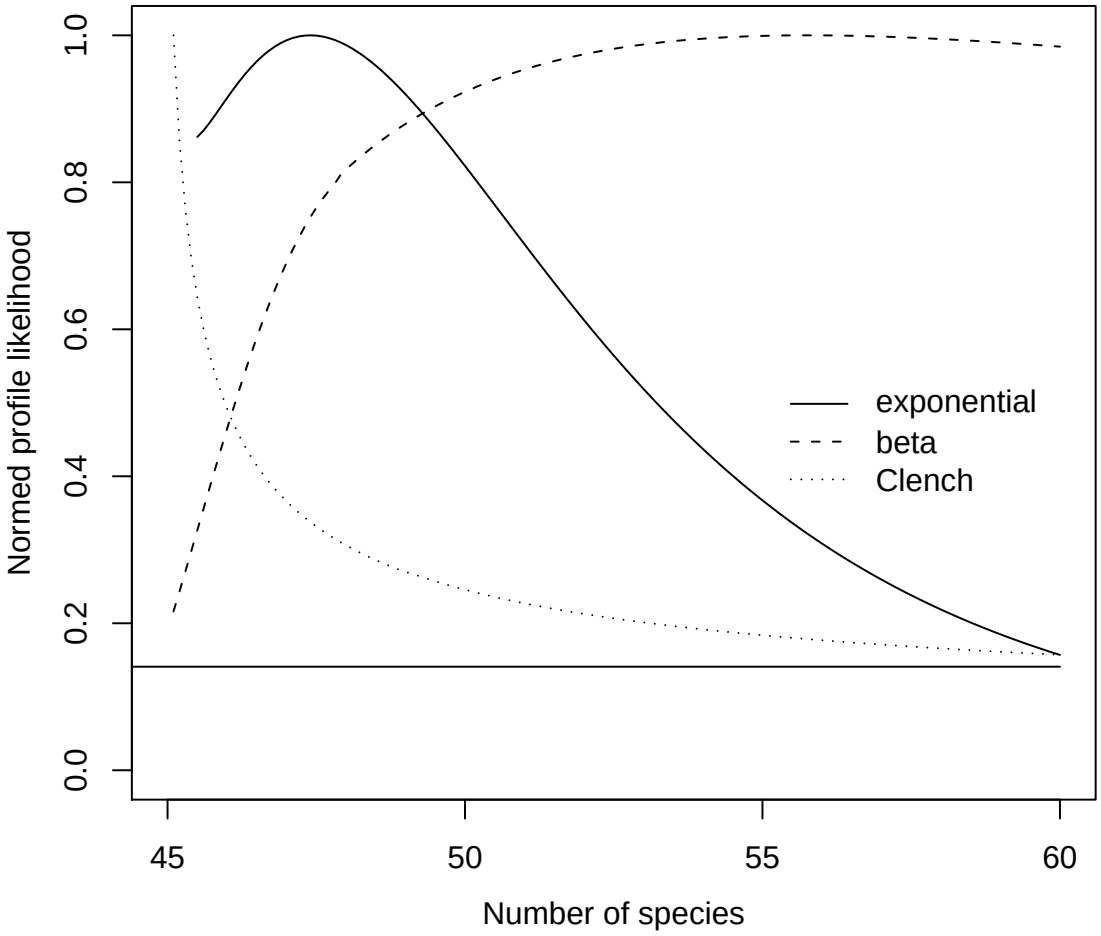

Figure 3: Profile likelihoods for the total number of species from the three multinomial models fitted to Irlanda spider data. The solid horizontal line indicates the 95% confidence interval.

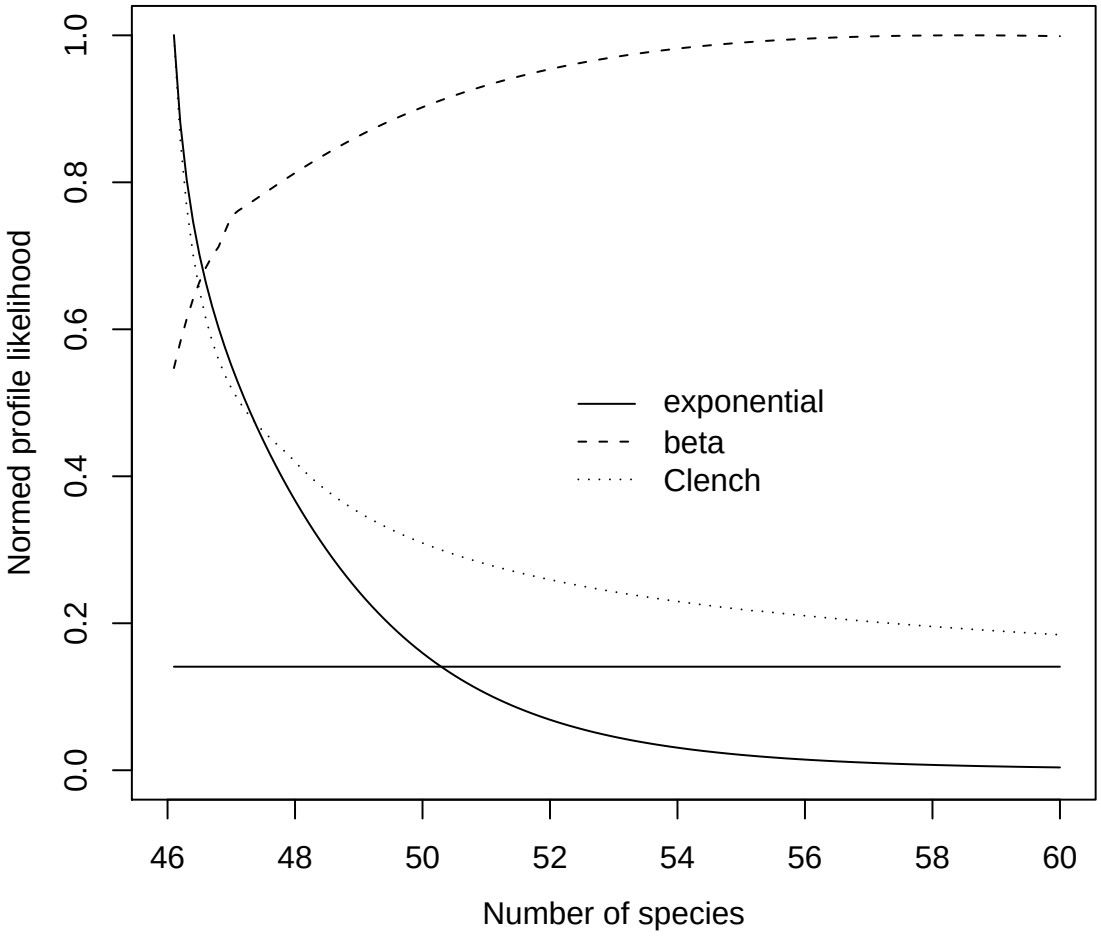

Figure 4: Profile likelihoods for the total number of species from the three multinomial models fitted to Hamburgo spider data. The solid horizontal line indicates the 95% confidence interval.

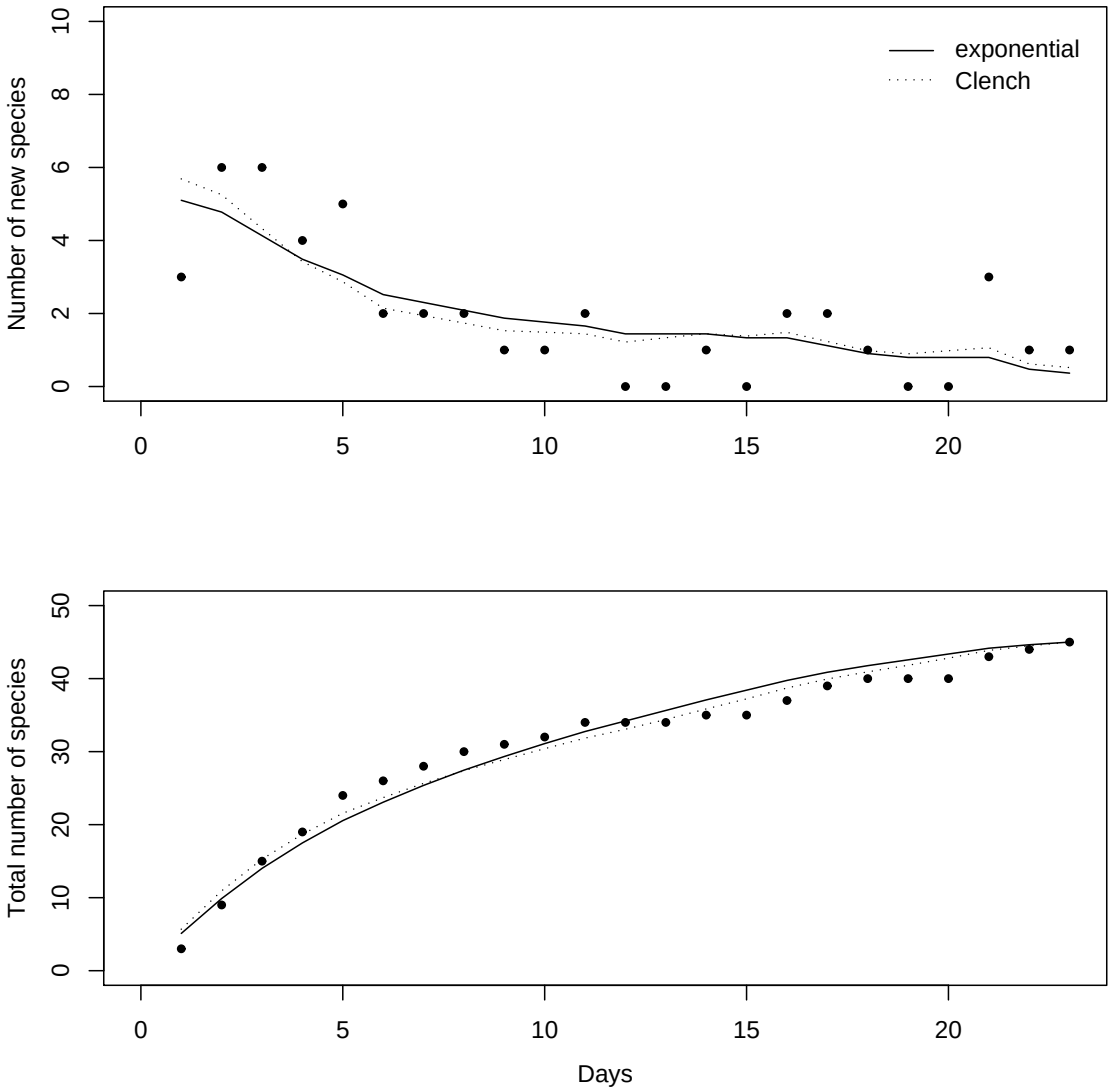

Figure 5: Fitted collecting (top) and accumulation (bottom) curves from the exponential and Clench multinomial models for the Irlanda spider data.

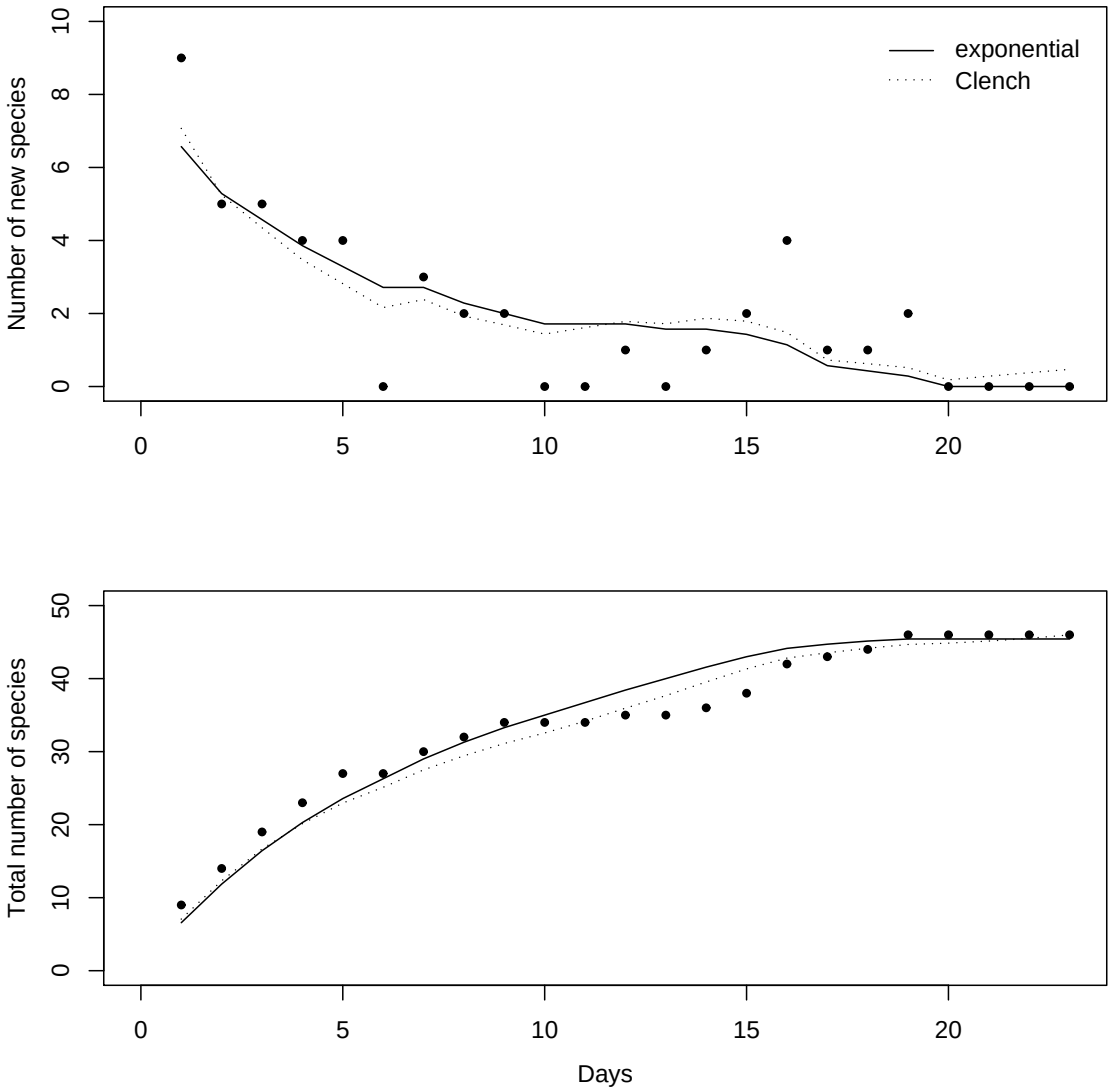

Figure 6: Fitted collecting (top) and accumulation (bottom) curves from the exponential and Clench multinomial models for the Hamburgo spider data.

5 Discussion

Traditional methods of estimating total numbers of species have used least squares to fit the accumulation curve. This approach is made difficult by the complex dependence of the variance on the mean in the birth process underlying these models and by the dependence among successive accumulated values that necessitates assuming an autoregressive process.

Estimation is simplified by working directly with the counts of new species and hence estimating the collecting function instead of the accumulation curve. Once the parameters are estimated, the accumulation curve can be plotted and the estimate of its asymptote obtained.

Traditional birth models provide an estimate of the mean count at the asymptote. Hence, it may be preferable to construct the birth models in terms of a multinomial distribution with fixed but unknown total rather than by using a nonhomogeneous birth process. As we have seen, these approaches can provide quite different results.

The AIC allows one to choose among the models, given the data, but, as always in prediction, the choice is rather risky, especially when few points are available on the curve to be extrapolated. Once a suitable model has been chosen, profile likelihood curves provide a useful of obtaining an interval of precision around the estimate of the total number of species.

Acknowledgments We thank Robert Gentleman and Ross Ihaka and the core group for developing the R software with which the analyses of the examples were done using the functions `gnlr` in the library `gnlm`, `gar` in the library `repeated`, and `elliptic` in the library `growth`. These libraries are available at `www.luc.ac.be/~jlindsey/rcode.html`.

References

[1] Clench, H. (1979) How to make regional lists of butterflies. Some thoughts. *Journal of the Lepidopterists' Society* **33**, 216–231.

[2] Díaz-Francés, E. and Gorostiza, L. (2000) Inferential estimation of species accumulation functions with pure birth processes and likelihood based methods. Submitted to *Journal of Agricultural, Biological, and Environmental Statistics.*

[3] Lindsey, J.K. (1995) Fitting parametric counting processes by using log-linear models. *Applied Statistics* **44**, 201–212.

[4] Nakamura, M. and Peraza, F. (1998) Species accumulation for beta distributed recording probabilities. *Journal of Agricultural, Biological, and Environmental Statistics* **3**, 17–36.

[5] Soberón, J. and Llorente, J. (1993) The use of species accumulation functions for the prediction of species richness. *Conservation Biology* **7**, 480–488.