On the analysis of accumulation curves*

J. Andrés Christen† Miguel Nakamura‡

Summary

Identifying and counting the total number of biological species observed, when plotted *vs.* a measure of the effort used to record them, gives rise to a *species accumulation curve.* We investigate estimation of the total number of species and other relevant properties of accumulation, by elaborating on a Multinomial model (Nakamura and Peraza, 1998, *Journal of Agricultural, Biological, and Environmental Statistics*, **3**, 17–36) that includes specification of a Beta density for recording probabilities. A unified description including complete and incomplete (aggregated) curves is obtained, and a Bayesian framework is considered to allow for inclusion of the biologist's knowledge with regard to recording probabilities and the total number of species. The Beta distribution is used as a prior, but recording probabilities are not restrained to be Beta distributed. The methods yield either closed analytical expressions for inference or relatively simple numerical procedures. Predictive distributions of future recordings that would eventually lead to an optimal decision-theoretical rule for stopping collection effort, may be easily calculated. Two case studies (species of bats and arachnids) are considered for illustration.

Keywords: BIODIVERSITY; NUMBER OF SPECIES; BAYESIAN INFERENCE; MCMC METHODS.

*Submitted to *Biometrics.*

†Instituto de Matemáticas, Unidad Morelia, U. N. A. M. Nicolás Romero 150, Centro 58000 Morelia, Mich., Mexico jac@matem.unam.mx .

‡Centro de Investigación en Matemáticas, A. C. Apartado Postal 402 36000 Guanajuato, Gto., Mexico nakamura@fractal.cimat.mx .

1 Introduction

Biodiversity studies resort to field work where species of a particular type (bats, butterflies, *etc.*) are identified and counted. The number of species observed is then plotted as a function of a measure of the effort used to record them, constructing a *species accumulation curve.* Measures of effort vary from area of quadrants sampled, biomass processed, number of nets or traps explored, or hours, days or any other period of observation (see Colwell and Coddington, 1994). While referring to effort, we say "number of recording periods, n" and the species accumulation curve becomes a plot of n *vs.* total number of species thus recorded so far, $S(n)$. In this article we study the problem of estimating the total number of species and other relevant properties of accumulation, using $S(n)$. This problem may be stated more generally: consider an urn containing balls that are painted using k different colors (k unknown), and that one wishes to estimate k by randomly drawing balls from the urn. A characteristic feature in using species (colors) accumulation curves for estimating k is that only the presence–absence of new species is recorded, and individuals of species already seen are disregarded (the number of balls identified with color c is not recorded, only whether or not color c has been yet observed). One could argue that more detailed data recording is desirable in order to have better chances of estimating k properly, however, as Colwell and Coddington (1994) point out, in the case of biodiversity studies this "[...]can add a substantial cost in time and effort for a large inventory, compared to collecting presence–absence data[...]".

Studied since the 1940's, the problem of estimating the number of species has spanned a wide range of applications in fields such as computer science, astronomy, social science, graph theory, biology, and others. However, most of this work is based on having access to n_x, the number of species recorded exactly x times. See Bunge and Fitzpatrick (1993) for a recent review. As a matter of fact, n_x's cannot be obtained from a species accumulation curve and thus, proper statistical methods for estimating the number of species and other quantities of interest using accumulation curves are in need.

Soberón and Llorente (1993) studied accumulation from a perspective that is mathematical but does not as yet consider any type of formal statistical inference. Based on a pure birth process justified on biological grounds, they fit the expected value of the stochastic process obtained, $A(n)$, to the observed accumulation curve $S(n)$ using least squares numerical techniques. The asymptotic behavior $\lim_{n\to\infty} A(n)$ is postulated as the total number of species k. This groundbreaking approach puts forth the importance of modeling the relationship between n and $S(n)$, but is lacking in an assessment of the quality (error) involved in estimating k, as well as in providing methodology for identifying the model for transition probabilities of the pure birth process. Moreover, it is especially important in biodiversity studies to assess the likely gain in precision regarding k that could be obtained with additional sampling. That is, it is interesting to consider prospective analysis of the likely amount of new species to be encountered if d additional units of effort were to be expended (see Colwell and Coddington, 1994). One of the main goals in analyzing accumulation is to investigate and to optimally plan sampling campaigns as well as suitable times to discontinue them. A correct assessment here is not possible with purely analytical non-inferential techniques, because there is not a framework set up for calculating probabilities (errors) or confidence bounds for future observations.

Another methodology for the statistical analysis of accumulation curves is developed by Nakamura and Peraza (1998) (henceforth N-P). Describing the behavior of species' recording probabilities through a probability density, and using working hypotheses like equal distribution for recording probabilities and homogeneity in time, they verify that the jumps in the accumulation curve $S(n)$ have a certain Multinomial distribution. Assuming a Beta distribution for recording probabilities, they propose asymptotic inference procedures for k and for future values of $S(n)$. Their methods consider likelihood and profile likelihood techniques using numerical optimization as well as bootstraping for some of the required error assessment. The main qualitative difference with regard to the Soberón and Llorente proposal, is that instead of basing the main feature of accumulation on a transition probability for a pure birth process, it now rests on a density for recording probabilities, which has a better chance of providing physical meaning and biological interpretation.

N-P also included the case when recording is reported after aggregation over several periods and the resulting accumulation curve is defined at possibly different interval lengths. This feature is common in biodiversity studies due to several practical reasons and is therefore an important problem to tackle. Referring to this as *incomplete* recording, they propose missing observation methodology using the EM algorithm for likelihood-based inference.

We here suggest to look at this problem from a point of view that may provide additional insight and potential benefits. The model considered by N-P rests on the postulated Beta distribution for recording probabilities. In their article, they consider robustness issues motivated by the fact that—precisely because of the lack of n_x's explained above—recording probabilities in a given setting will not be generally available for empirical verification, and a theoretical distribution for them is not clearly apparent. The Beta assumption is in a way merely a convenient modelling device. Although not documented in their article, an interesting point to note now is that in both of the examples there analyzed, biologists anticipated what the overall shape the estimated distribution of recording probabilities would turn out to be, and furthermore the Beta family was found to be an adequate and convenient tool for describing the recording patterns of different ecosystems. Lacking empirical or theoretical justifications for recording probabilities, it appears reasonable to attempt to analytically take into consideration the biologist's expertise. The main purpose of the present article is to consider a framework—specifically, that of Bayesian inference—that readily acknowledges the nature of this subjective postulation and incorporates it directly into the inferential procedure. In this perspective, a Beta distribution used as a prior is not as critical since, given the data, the posterior distribution for recording probabilities is not restricted to such class. This Beta prior is only used for practical reasons, providing a flexible–enough class of distributions with which to express relevant prior information in a simple manner.

Based on a setup that uses similar modelling components as in N-P, the Bayesian inference developed in this paper gives rise to several closed-form analytical solutions as well as procedures whose correctness does not rely on asymptotic approximations. When analytical expressions are not available, very straightforward Gibbs sampling is possible that results in reliable numerical stability. Prior information about the likely number of species and recording probabilities may be included in the analysis to improve inferences, with posterior distributions providing full error assessment. Also, predictive distributions of future recordings that would eventually allow for a rigorous decision-theoretical answer to the question of when to terminate the search for new species, may also be easily calculated.

In this article we first elaborate on the model used by N-P, then consider a Bayesian framework to approach the inference problem, and provide examples to illustrate the techniques. Basic notions of the Bayesian procedure for inference will be taken for granted throughout. A few modern, specific tools for inference (such as Gibbs sampling and the simulation of log-concave densities) will be used for numerical aspects of the solutions to be considered, and some familiarity with these will also be assumed.

2 The Multinomial Model

Colwell and Coddington (1994) state "species accumulation curve to refer to a data set for a local species assemblage in an area of habitat that is roughly homogeneous both spatially and temporally." Accordingly, letting k be the total number of different species and P_i be the probability of capturing species i during a single unit of effort, $i = 1, 2, \ldots, k$, we assume that the P_i's remain constant over the sampling period. Due to a combination of factors such as species abundance, detectability, weather conditions, *etc.*, aside from the fact that the sampling site is generally selected at random, these probabilities may be considered random variables. As explained by N-P, because it is the pooled effect of all the P_i's that governs the behavior of accumulation and no attempt is made at keeping track of individual assignments of each P_i to each individual species, we may also act as if the P_i's originated from i.i.d. sampling from some generic distribution. This parent distribution for P_i's is thus used as a descriptor of their relative composition.

The model outline that follows differs slightly from N-P's in that incomplete and complete collection lists are merged into a single setting. Instead of viewing the incomplete list as a case of missing values and then indicating a specific algorithm for inference (expectation-maximization), we focus directly on the locations in time where the list was tallied, whether they occur consecutively or spaced apart. Let m be the number of places, $j_1, j_2, \ldots, j_m$, where accumulation is observed exactly, where $0 = j_0 < j_1 < \cdots < j_m = n$. For $i = 1, 2, \ldots, k$ and $j = 1, \ldots, n$, let $Z_{ij} = 1$ if species i is recorded for the first time on trial j and $Z_{ij} = 0$ otherwise. Also define $Z_{i0} = 1$ if species i is not yet recorded by the n-th trail ($Z_{i0} = 0$ otherwise), and let

$$Y_{il} = \sum_{j=j_{l-1}+1}^{j_l} Z_{ij}, \quad 1 \leq l \leq m.$$

That is, $Y_{il} = 1$ if species i is recorded for the first time within trials $j_{l-1} + 1, j_{l-1} + 2, \ldots, j_l$ and $Y_{il} = 0$ otherwise. We define $Y_{i0} = Z_{i0}$. Let the number of new species recorded within trials $j_{l-1} + 1, j_{l-1} + 2, \ldots, j_l$ be

$$X_l = \sum_{i=1}^{k} Y_{il}$$

and the number of species not yet recorded $X_0 = \sum_{i=1}^{k} Y_{i0}$. The aggregated accumulation curve reported only on the m sampling trials $j_1, j_2, \ldots, j_m$ is $S(l) = S(l-1) + X_l$ and the X_l's are exactly the jumps in $S(l)$.

Using a modified version of N-P's argument we prove in Appendix A that

$$X_1, X_2, \ldots, X_m \mid D, k \sim MN_m(\pi_0, \pi_1, \ldots, \pi_m, k) \tag{1}$$

where the P_i's have distribution $\mathbf{D}$ and

$$\pi_l = E[(1-P_i)^{j_{l-1}}\{1-(1-P_i)^{h_l}\}], \tag{2}$$

$\pi_0 = E\{(1-P_i)^n\}$ and $h_l = j_l - j_{l-1}$.

At this point there is a clear—although somewhat complex—Bayesian way to proceed. A joint prior distribution for $\mathbf{D}$ and k may be defined to find the posterior $f(k \mid X_1, X_2, \ldots, X_m)$. This entails defining a probability distribution on distributions for the P_i's and coping with the complexities of infinite dimensional parameters such as $\mathbf{D}$ referred to in Bayesian nonparametrics (see, for example, Liu, 1996, and references therein). A theoretically interesting and challenging problem is posed, but using the following argument for simplification we convert this into a standard parametric problem and leave the infinite dimensional scenario for future research.

Obviously, only the average behaviors of the P_i's are relevant as far as likelihood based inference is concerned: the determining features of the distribution of P_i are given only by the π_l's. It is intuitively appealing to consider a fictitious random variable p possessing the same distribution as the P_i's, and to consider the model

$$X_1, X_2, \ldots, X_m \mid p, k \sim MN_m(\pi_0, \pi_1, \ldots, \pi_m, k), \tag{3}$$

where the relevant moments now have the form

$$\pi_l = (1-p)^{j_{l-1}}\{1-(1-p)^{h_l}\} \tag{4}$$

and $\pi_0 = (1-p)^n$. Consideration of the posterior distribution of p as an approximation for the common posterior distribution of the P_i's, obviously results in an enormous gain in simplicity.

At first sight this simplification might seem overly restrictive. It is interesting to note, however, that for typical distributions on $[0,1]$ which are likely to be found or considered in the context of species accumulation, we may in fact, produce good approximations between moments (2) and those restricted to be of the form (4). Examples which we consider "unreasonable" distributions $\mathbf{D}$ are those which specify that fairly heterogeneous values of p are equally probable, such as prescribed by the uniform (0, 1), a wide unimodal distribution on (0, 1), or a generic U-shaped density. On the other hand, "reasonable" distributions to consider—of a type which have already proven useful in describing recording probabilities—are fairly concentrated, having relatively light tails (some being symmetric and others heavily skewed).

The question is: given a "reasonable" distribution, is it possible to approximate its moments (2) by moments (4) for some p? Selecting p as the value that minimizes the maximum of absolute differences between the first eight of these moments, we generally find in examples of "unreasonable" distributions that the corresponding moments may be only roughly approximated, whereas for many "reasonable" distributions these sets of moments become virtually indistinguishable from one another. We here use simple, specific densities to illustrate our point: consider Beta densities with parameters (1, 1), (3, 3), (.5, .5), (.6, 75), (24, 70), and (20, 1). The first three of these exemplify what we have above called "unreasonable" distributions for recording, and the last three are "reasonable". The values of p that produce the greatest resemblance between (2) and (4) over $1 \leq l \leq 8$ are .423, .466, .612, .008, .254, and .953, respectively, and the maxima of the eight absolute differences are .0773, .0345, .1124, .0003, .0022, and .0017. Notice that the second set of six quantities should be viewed on the (0, 1) scale, because each difference is at most 1. In this sense, "reasonable" densities produce notoriously smaller discrepancies.

Once the previous simplification is adopted, it is immediate to see that the likelihood is

$$L(\mathrm{X};\, p,\, k) = \frac{k!}{\prod_{l=1}^{m} X_l!(k-u)!}(1-p)^{nk+v}\prod_{l=1}^{m}\{1-(1-p)^{h_l}\}^{X_l} \qquad (5)$$

where $\mathrm{X} = (X_1, X_2, \ldots, X_m)$, $u = \sum_{l=1}^{m} X_l$, $v = -nu + \sum_{l=1}^{m} j_{l-1}X_l$ and $h_l = j_l - j_{l-1}$.

As mentioned earlier, researchers in ecology have a fairly good notion of the general distributional shape for p, as well as a good idea of plausible values of k. Assuming this judgment regarding p gives no bearing on the judgment for k, for the joint prior of p and k we consider that $f(p,k) = g(p)h(k)$, where $p \sim Beta(\alpha, \beta)$ for given α, β and $h(k)$ is an arbitrary probability function with support contained in the non-negative integers.

Notice that the Beta distribution and the parameters α, β are also used in the present formulation, as it is in the N-P viewpoint. However, the role of this Beta distribution is quite different. In the present context, it is used as an instrument for assessing the researcher's prior opinion (via the parameters α, β) regarding the value of the artificial variable p used to approximately describe the moments of the P_i's. In sharp contrast, N-P view the Beta distribution as a postulated device used to explain a structural aspect of the model and the parameters α, β are incorporated into the task of estimation. We will see, in fact, that the posterior distribution for p is rather more general than a Beta distribution.

3 Inference for Equal Period Recording

When $h_l = 1$ the accumulation curve is defined at consecutive periods, as is usually the case when n represents time (days, weeks, *etc.*) and accumulation is recorded systematically at each time period. The likelihood is reduced to

$$L(\mathrm{X};\, p,\, k) = \frac{k!}{\prod_{l=1}^{m} X_l!(k-u)!}(1-p)^{nk+v}p^u.$$

Thus, the joint posterior is

$$f(p,\, k \mid \mathrm{X}) \propto \frac{k!}{(k-u)!}(1-p)^{nk+v+\beta-1}p^{u+\alpha-1}h(k). \tag{6}$$

Identifying the relevant part of $\int f(p,\, k \mid \mathrm{X})\, dp$ with a beta density kernel we obtain the posterior marginal for k, which is

$$f(k \mid \mathrm{X}) \propto \frac{k!}{(k-u)!}h(k)\frac{\Gamma(nk+v+\beta)}{\Gamma(u+\alpha+nk+v+\beta)}, \tag{7}$$

for $k = u, u+1, \ldots$. To obtain the marginal for k we need only to sum over k for normalizing the above expression, producing the sole instrument for inference about k under this Bayesian perspective.

Regarding the joint posterior in (6), notice that

$$f(p \mid \mathrm{X},\, k) \propto f(p,\, k \mid \mathrm{X}) \propto (1-p)^{nk+v+\beta-1}p^{u+\alpha-1},$$

that is, $p \mid \mathrm{X},\, k \sim Be(u+\alpha,\, nk+v+\beta)$. ¿From this we can readily obtain the posterior marginal for p, because $f(p,\, k \mid \mathrm{X}) = f(p \mid \mathrm{X},\, k)\, p(k \mid \mathrm{X})$ and thus

$$f(p \mid \mathrm{X}) = \sum_{k=u}^{\infty} f(p \mid \mathrm{X},\, k)\, f(k \mid \mathrm{X}), \tag{8}$$

where $f(p \mid \mathrm{X},\, k)$ is just the corresponding Beta density and $f(k \mid \mathrm{X})$ is given by (7). Notice this density is a weighted average of Beta densities, showing that the marginal posterior for p is a lot more flexible than the Beta family alone.

3.1 Predicting Future Recordings

For prediction of future recordings, which would be very useful in deciding stopping times for search campaigns, the required instruments for inference—namely, conditional distributions of future values given observed data—cannot be obtained analytically in closed form. MCMC (Gibbs) sampling techniques are here proposed for simulating realizations of future values conditionally on the observed accumulation data. The Gibbs sampling algorithm is based on the "full conditionals", which are the conditional distributions of one (or a block of) parameter(s) given the rest of parameters and the data. Here we are interested in predicting $X_{n+1},\, X_{n+2},\, \ldots$; these unobserved X_i's are simply treated as extra unknown parameters. Full conditionals in this problem are easy to obtain and the corresponding chain is conveniently simple. Letting X be the vector of observed and future X_i's, the full conditionals are

$$p(k \mid \mathrm{X},\, p) \propto \frac{k!}{(k-u)!}(1-p)^{nk}h(k) \tag{9}$$

for $k = u,\, u+1,\, \ldots,$ and

$$p \mid \mathrm{X},\, k \sim Be(u+\alpha,\, nk+v+\beta).$$

(Note that u and v have the same expressions as before but now include the future, simulated, X_i's.) For a future observation X_i, $i = n+1,\, n+2,\, \ldots$ it may be shown (see Appendix B) that

$$p(X_i = x_i \mid \mathrm{X}_{-i},\, k,\, p) = \frac{C^r_{x_i} p^{x_i}\{(1-p)^{n+1-i}\}^{r-x_i}}{\{p+(1-p)^{n+1-i}\}^r} \tag{10}$$

for $x_i = 0,\, 1,\, \ldots,\, r$ where $r = k - \sum_{j \neq i} X_j$ and X_{-i} is the vector of X_j's having removed X_i. Because it is not difficult to sample from either of these distributions, we can implement the Gibbs sampling scheme to obtain samples from future recording effort. For example, fixing a number of days to analyze, say d, we could focus on $\sum_{i=n+1}^{n+d} X_i$ to estimate to number of new species we expect to record in d additional days. One could also fix a number of new species, say M, and estimate the number of additional units of effort required to obtain them.

The Gibbs sampling algorithm begins with arbitrary values for the parameters to be sampled (in this case k, p, X_{n+1}, X_{n+2}, $\ldots$, X_{n+d}) and simulates in turn from the full conditionals updating parameter values. This algorithm forms a Markov Chain that, under broad regularity conditions, has a stationary distribution that equals the joint (posterior) distribution for the parameters. Simulated stationary values may then be taken as realizations of the posterior distribution. Several details should be contemplated when using MCMC techniques (see Gamerman, 1997). However, Gibbs sampling considered in this article performs well and very stable approximations are achieved in runs of no more than 5 minutes on a Pentium processor. In comparison to MCMC sampling proposed recently for very high dimensional problems and models which are a lot more elaborate, it is fair to say that the sampling schemes described here are rather straightforward.

4 Inference for Unequal Period Recording

When h_l is not constant we see, using the prior for k and p as above, that

$$f(p,\, k \mid \mathrm{X}) \propto \frac{k!}{(k-u)!} p^{\alpha-1}(1-p)^{nk+v+\beta-1} \prod_{l=1}^{m} \{1-(1-p)^{h_l}\}^{X_l} h(k).$$

Since easily tractable integrands to marginalize the above joint posterior—as was the case for $h_l \equiv 1$—cannot be obtained here, we call upon Gibbs sampling. It is easy to see that the full conditional for k is also (9). But the full conditional for p is now

$$f(p \mid \mathrm{X},\, k) \propto p^{\alpha-1}(1-p)^{nk+v+\beta-1} \prod_{l=1}^{m} \{1-(1-p)^{h_l}\}^{X_l}$$

which again lacks a convenient, closed form. However, because the second derivative of $\log f(p \mid \mathrm{X},\, k)$ can easily be shown to be negative, this density is log-concave. Methods for simulating from log-concave densities (see Devroye,

1986) can be used. Because the first derivative of $\log f(p \mid \mathrm{X},\, k)$ is available and relatively simple, we decided to use a slightly modified version of the algorithm explained in Gilks and Wild (1992) to sample from the above full conditional for p. A Gibbs sampling scheme may now be implemented. Complete C coding of this and all other algorithms may be obtained from the authors.

4.1 Future Recordings

For future recordings, adopting the same notation as in Section 3.1, we only need now to address the full conditional for a future X_i, because the full conditionals for k and p have just been discussed in the previous Section. It is shown in Appendix B that the full conditional for a future observation X_i is

$$p(X_i = x_i \mid \mathrm{X}_{-i},\, k,\, p) = \frac{C^r_{x_i}\{1-(1-p)^{h_i}\}^{x_i}\{(1-p)^{n-j_{i-1}}\}^{r-x_i}}{[\{1-(1-p)^{h_i}\}+(1-p)^{n-j_{i-1}}]^r} \tag{11}$$

for $x_i = 0,\, 1,\, \ldots\,,\, r$ where $r = k - \sum_{j \neq i} X_j$.

5 Examples

The example of bat accumulation considered in Soberón and Llorente (1993) and again in N-P will be revisited here under the alternate approach. We also consider a new example regarding species of arachnids accumulated in coffee orchards in Southern Mexico. By plotting several of the posterior distributions described in this paper, results will consist mainly of graphical displays. Highest Posterior Density sets (HPD's) and Maximum a posteriori (MAP) estimators may also be given. The main point in this section is to exemplify the type of information the biologist may extract from these analytical tools, and how it is interpreted. A third, contrived, yet worthy example, is presented using the same bat data with artificial aggregation in the accumulation curve, to illustrate and compare the method described in Section 4.

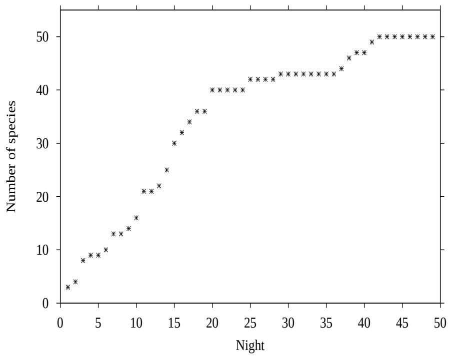
We should stress perhaps that one of the main strengths in the methodology is that it analytically takes into account the expertise of the biologist/ecologist via reasonable priors. Since our main purpose now is demonstration, we here disregarded true interaction and feedback with the biologist and in all cases simply used a Uniform prior on [0, 1] for p (*i.e.*, $Beta(1,1)$) and a uniform prior for k for the integers $1, 2, \ldots, 120$ (in fact, these are acting as vague priors). In a realistic application, of course, the issue of prior elicitation should be addressed with greater care (see Pham-Gia, Turkkan, and Duong, 1992, for example).

5.1 Bats

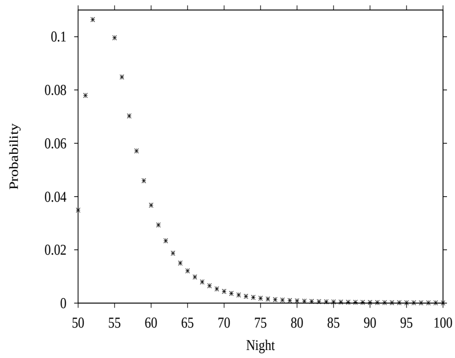
Consider species of bats observed at the Chajul Biological Station in Southeast Mexico (Medellin, 1993). A total of 50 species were observed having set nets over a period of 49 consecutive nights. Figure 1 shows plots for the accumulation curve and for the posterior of k and p using the methods of Section 3. The 90% HPD set for k is [50, 62] and the MAP estimator (the maximum of $f(k \mid \mathrm{X})$) is $k = 54$ with a probability in excess of 0.11; we could then say that it is very likely that there are some 4 to 12 species yet to be recorded. The posterior probability that $k > 76$ is less than 0.01, immediately providing insight into an upper bound for the total number of bat species present.

N-P analyzed this data using numerical implementation of asymptotic profile likelihood methods and obtained a 95% confidence interval for k as [50, 68], and a maximum likelihood point estimator at $k = 54$. Furthermore, their fitted Beta density which accounts for recording probabilities was roughly bell-shaped, concentrated around 0.05, closely agreeing with the posterior obtained here for p. We note the results are qualitatively very similar, despite the broad differences in the models and the methods for assessment of uncertainty that the two procedures for statistical inference provide.

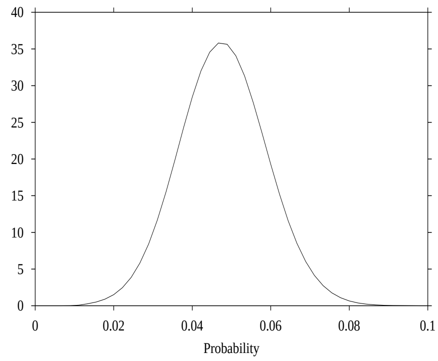
We also performed one example of predictive analysis over ($d =$) 5 extra days of field activity using the methodology described in Section 3.1. From Gibbs sampling we obtained the predictive distribution plotted in Figure 1(d). Note that it is just as likely to encounter one or no new species, with these two numbers accumulating nearly 0.7 of probability. Does this suggest to stop sampling? This decision should depend on the current costs incurred as well as how one measures the quality of the information gained so far. For certain, an analytic approach to optimal stopping of the sampling effort is interesting in itself and needs to be the specific subject for future research.



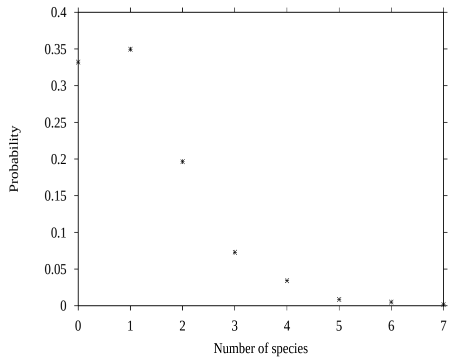
(a)



(b)



(c)



(d)

Figure 1: Bat Example. (a) Accumulation curve, (b) posterior pmf for k, (c) posterior pdf for p and (d) predictive pmf for the number of new species to be observed in 5 more days of capture.

5.2 Arachnids

This example considers species of weaver-spiders observed in coffee orchards in the Soconusco region of Chiapas, Southern Mexico. It involves two coffee orchards, Irlanda (Ir.) and Hamburgo (Ha.); 45 and 46 species were observed over a period of 23 months, respectively[1]. These plantations are adjoining, located approximately at the same altitude (900–990m), so at first sight both accumulations may appear in fact to be replicates of the same stochastic process. In both cases the accumulation curve is recorded every month ($h_l \equiv 1$), again enabling the use of the analytic methods in Section 3.

Figure 2 shows plots for the accumulation curves and for the posteriors of k and p for both data sets. Note that, although basically the same number of species has been observed in both plantations, the growth of the accumulation curves is inherently different. Corresponding 90% HPD intervals for k at Ir. and Ha. are [45, 60] and [46, 51]. Orchard Ir. results in greater uncertainty regarding the parameter k, with greater probability assigned to larger values. Closer graphical examination of the curves suggests that accumulation at Ir. after 23 months is still somewhat active, whereas at Ha. the process is closer to becoming motionless or saturated. Posteriors for p suggest that Ir. has recording probabilities which are slightly larger than those at Ha., so even if both orchards are in the same region a characteristic is suggested which is causing the search for species to be of a different nature. In fact, as the main researcher in biology explains, plantations do have strong distinguishing features. Ir. is organically maintained, includes remains of the original forest for the provision of shade, and coffee shrubs are there planted along constant altitude countour levels. Ha. uses chemical fertilization and weeding, is set on cleared forestland, and shrubs there are planted in straight rows.

Using the procedure described in Section 3.1, we again obtained predictive analysis, assuming the possibility of extending the search for 5 additional months was being considered. Gibbs sampling produced both predictive distributions plotted in Figure 2(d). Here again, the results make it evident that there seem to be more new species to be found sooner at Ir. (where there is, as seen, greater physical complexity) if sampling were to continue for another 5 months. Here, it is likely to observe 0 to 6 new species at Ir. and between 0 and 4 at Ha.

[1] The assumption of homogeneity in time is considered sensible here despite seasonal variation, because seasons only influence the *location* of spider webs on the shrub, and all parts of the shrub were inspected.

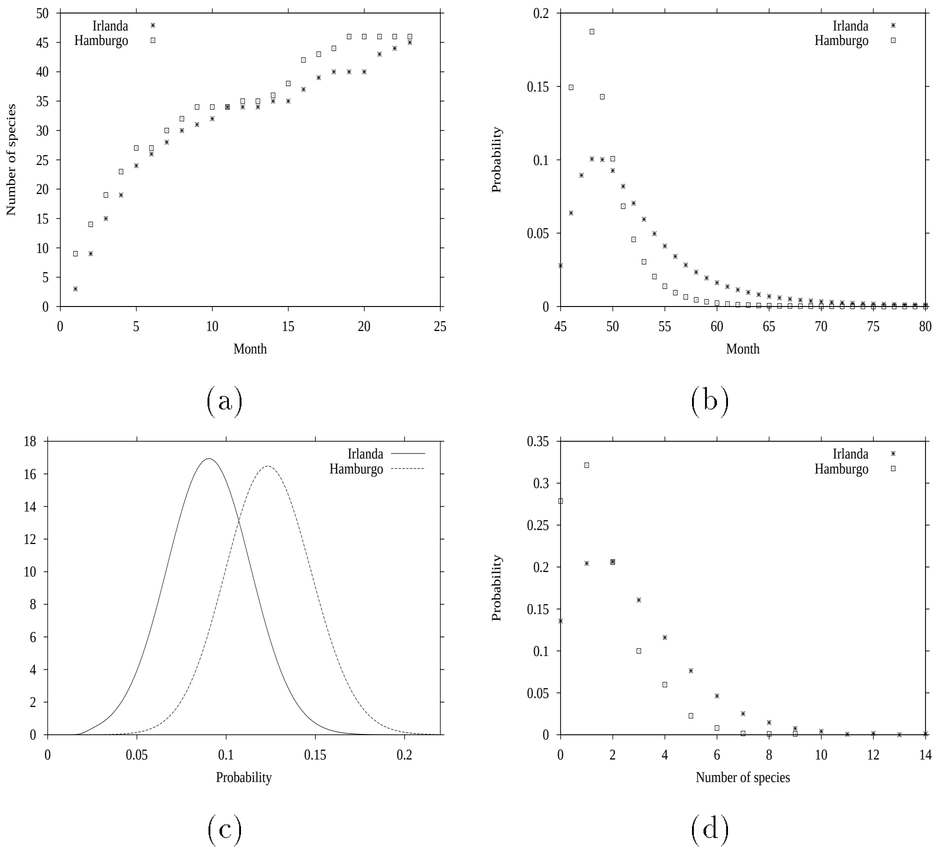


Figure 2: Arachnid Example. (a) Accumulation curves, (b) posterior pmf's for k, (c) posterior pdf's for p and (d) predictive pmf's for the number of new species to be observed in 5 more days of capture.

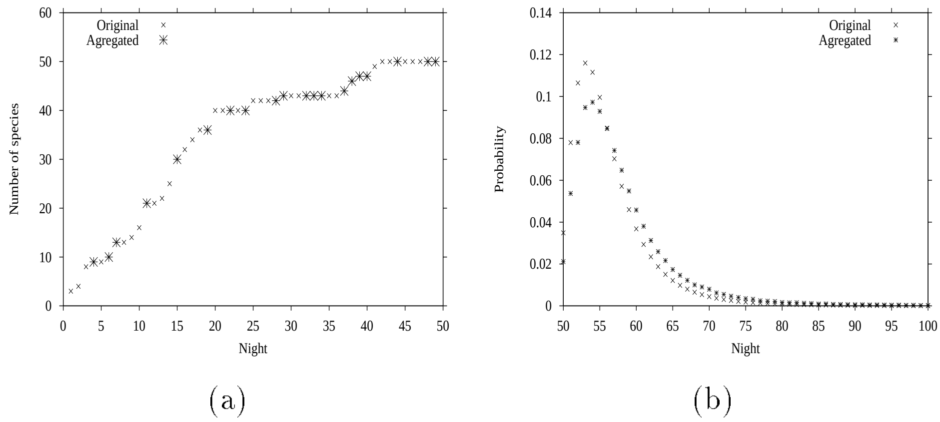


Figure 3: Aggregated Bat Example. (a) Original and aggregated accumulation curves and (b) posterior pmf's for k for the original and aggregated curves.

5.3 Aggregated Bats

To illustrate analysis performed on incomplete lists, we again considered Chajul bats and randomly aggregated species observed over the complete sampling period by selecting the h_l's from 1 to 7 uniformly. We obtained $m = 13$ observational points. The resulting aggregated accumulation is shown in Figure 3(a) along with the original curve. Gibbs sampling for incomplete lists as explained in Section 4 was used. To check for convergence, we considered samples of 10,000 points with various starting values, and found that the same posteriors were obtained in all cases[2]. Figure 3(b) displays the posterior mass function for k superimposed on the original posterior for k. Remarkably, despite the fact that observational points have been reduced by more than a third, both posteriors agree rather closely, with the slightly larger uncertainty arising from the loss of information due to aggregation. The close agreement obtained here would not hold in general; in this example, it holds because even the reduced accumulation curve already seems to have stabilized, pointing to a narrow range of values for k. If a curve were to be analyzed at an earlier stage in which it is still clearly increasing, much more uncertainty regarding k would obviously be anticipated.

[2]As a matter of fact, one can also use this method when $h_l \equiv 1$. We utilized the original Chajul data and, with a sample size of 10,000 points, achieved a perfect match with the posterior obtained using the analytical expressions in Section 3.

6 Discussion

Model specification for accumulation curves considered by N-P was generalized here for incomplete lists. Again basing main features of accumulation on the notion of recording probabilities, and following a device for simplification, Bayesian inference is considered in order to incorporate prior information which biologists may possess regarding the total number of species and the shape of the distribution of recording probabilities.

Ecology and biodiversity agencies are typically not interested in a single accumulation curve but may deal instead with a very large number of different geographical locations and lists of species. If proper statistical inference is to be advanced and performed on each accumulation curve on a case-by-case basis, a simple procedure would certainly be preferred. Implementation of techniques using geographical information systems is beginning to become widespread. Computer intensive and potentially unstable numerical procedures with critical starting points required, for which monitoring by an experienced statistician is often recommended, simply do not lend themselves to be considered for the analysis of large numbers of case studies. We are not suggesting that numerical simplicity should ever completely justify an inferential procedure. Instead, in the present biodiversity setting, a sensible compromise between numerical simplicity and rigorous mathematical models must be reached in a typical application of any technique. The procedures considered here yield either closed, exact, analytical forms or numerical procedures that are straightforward to implement. Gibbs sampling is used to obtain posterior distributions, and convergence is found to be adequate and stable.

The methods here developed consider priors for p in the Beta family. We have reasons to believe that this simple family of priors will be sufficiently flexible for conveying information in most applications. However, if ever deemed necessary, linear combinations of Beta densities may also be considered, with the corresponding posterior distributions following through transparently. The methodology enables inference procedures for the total number of species, and the density for recording probabilities. We consider the elicitation of priors to be an important issue but have not addressed it here in depth.

As mentioned earlier, development of a valid sequential procedure for optimal stopping of the sampling effort is also the subject of future, specific, research. The basic tool for inference relevant to this problem has been studied here, that is, the predictive distribution for future recordings. However, further theoretical and practical aspects must first be addressed in order to derive a formal algorithm. Sequential procedures take into account all possible future outcomes to evaluate the relative costs of stopping and continuing with the sampling program. On the theoretical side, because very likely analytical solutions will not be available, clever, non-brute-force, numerical algorithms will be needed to obtain a stopping rule of practical value, which in turn will depend strongly on the specific loss function considered. Therefore, on the practical side, careful elicitation of adequate loss functions will be needed so as field biologists feel comfortable with the output of "black box" algorithms, telling them when to abandon the search for species. In this respect, we forsee a great amount of interdisciplinary research.

7 Acknowledgments

Dr. Nakamura's work was partially supported by CONACYT grant 0328P-E9506. We thank Dr. Guillermo Ibarra for kind discussion regarding the arachnid examples. This work was completed while A. Christen was visiting CIMAT.

References

[1] Bunge, J. and Fitzpatrick, M. (1993). Estimating the number of species: A Review. *Journal of the American Statistical Association* **88**, 364–373.

[2] Colwell, R. K. and Coddington, J. A. (1994). Estimating terrestrial biodiversity trough extrapolation. *Philosophical Transactions of the Royal Society of London*, Ser. B (Special volume on "The quantification and estimation of organismal biodiversity") **345**, 101–118.

[3] Devroye, L (1986). *Non-uniform random variate generation.* New York: Springer.

[4] Gamerman, P. (1997). *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference.* London: Chapman and Hall.

[5] Gilks, W. R. and Wild, P. (1992). Adaptive rejection sampling for Gibbs sampling. *Applied Statistics* **41**, 337–348.

[6] Liu, J. S. (1996). Nonparametric hierarchical Bayes via sequential imputations. *The Annals of Statistics* **24**, 911–930.

[7] Medellin, R. (1993). Estructura y diversidad de una comunidad de murciélagos en el trópico húmedo mexicano. In *Avances en el estudio de los mamíferos mexicanos*, R. Medellin and G. Ceballos (eds), Asociación Mexicana de Mastozoología A. C., Publicaciones Especiales **I**, 333–354.

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

[9] Pham-Gia, T., Turkkan, N., and Duong, Q. P. (1992). Using the mean deviation in the elicitation of the prior distribution. *Statistics and Probability Letters* **13**, 373–381.

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

A Proof of (1)

Following the lines of N-P (Appendix A) closely, we prove (1) using our present notation. Since $Y_{il} = 1$ if species i is recorded for the first time within trials $j_{l-1}+1, j_{l-1}+2, \ldots, j_l$ and $Y_{il} = 0$ otherwise we define for $1 \le i \le k$ and $1 \le l \le m$

$$\pi_{il} = P(Y_{il} = 1 \mid P_i = p_i) = (1-p_i)^{j_{l-1}}\{1-(1-p_i)^{h_l}\}$$

and

$$\pi_{i0} = P(Y_{i0} = 1 \mid P_i = p_i) = (1-p_i)^n.$$

Let P_i have a common density g, thus $P(Y_{il}=1) = E(\pi_{il}) = \int_0^1 \pi_{il} g(p_i)\, dp_i$.

Note that only one of the variables $Y_{i0}, Y_{i1}, \ldots, Y_{im}$ has the value of one while the rest is zero, therefore their joint pgf is

$$\begin{aligned}\phi_{Y_{i0},Y_{i1},\ldots,Y_{im}}(s_0, s_1, \ldots, s_m) &= E(s_0^{Y_{i0}} s_1^{Y_{i1}} \ldots s_m^{Y_{im}}) \\ &= E(\pi_{i0})s_0 + E(\pi_{i1})s_1 + \cdots + E(\pi_{im})s_m.\end{aligned}$$

Considering now $X_l = \sum_{i=1}^{k} Y_{il}$ and assuming independence among species the joint pgf of these variables is

$$
\begin{aligned}
\phi_{X_0, X_1, \ldots, X_m}(s_0,\, s_1,\, \ldots\,,\, s_m) &= E(s_0^{X_0} s_1^{X_1} \ldots s_m^{X_m}) = \\
&= E(s_0^{\sum_{i=1}^{k} Y_{i0}} s_1^{\sum_{i=1}^{k} Y_{i1}} \ldots s_m^{\sum_{i=1}^{k} Y_{im}}) = \\
&= E(\prod_{l=0}^{m} \prod_{i=1}^{k} s_l^{Y_{il}}) = \\
&= E(\prod_{i=1}^{k} \prod_{l=0}^{m} s_l^{Y_{il}}) = \\
&= \prod_{i=1}^{k} E(\prod_{l=0}^{m} s_l^{Y_{il}}) = \\
&= \prod_{i=1}^{k} (E(\pi_{i0}) s_0 + E(\pi_{i1}) s_1 + \cdots + E(\pi_{im}) s_m).
\end{aligned}
$$

Note now that $E(\pi_{il})$ does not depend on i and therefore we define for $1 \leq l \leq m$

$$
\pi_l = E(\pi_{il}) = E[(1-p_i)^{j_{l-1}} \{1 - (1-p_i)^{h_l}\}]
$$

and $\pi_0 = E\{(1-p_i)^n\}$. Thus we obtain

$$
\phi_{X_0, X_1, \ldots, X_m}(s_0, s_1, \ldots\,, s_m) = (\pi_0 s_0 + \pi_1 s_1 + \cdots + \pi_m s_m)^k
$$

which is identified as the pgf of a multinomial density with index k and $m+1$ classes having probability parameters $\pi_0,\, \pi_1,\, \ldots\,,\, \pi_m$.

B Proof of (11)

Given the likelihood in (5) it is seen that

$$
f(X_i = x_i \mid \mathrm{X}_{-i},\, k,\, p) \propto \frac{1}{x_i!(r - x_i)!} \{1 - (1-p)^{h_i}\}^{x_i} (1-p)^{x_i (j_{i-1} - n)}
$$

for $x_i = 0, 1, \ldots\,, r$, where $r = k - \sum_{l \neq i} x_l$. Multiplying by the constant $r!(1-p)^{-r(j_{i-1}-n)}$ we obtain

$$
f(X_i = x_i \mid \mathrm{X}_{-i},\, k,\, p) \propto C_{x_i}^{r} \{1 - (1-p)^{h_i}\}^{x_i} \{(1-p)^{n - j_{i-1}}\}^{r - x_i}.
$$

The normalizing constant is then found as

$$
\begin{aligned}
K &= \sum_{x_i=0}^{r} C_{x_i}^{r}\{1-(1-p)^{h_i}\}^{x_i}\{(1-p)^{n-j_{i-1}}\}^{r-x_i} \\
&= \{1-(1-p)^{h_i}+(1-p)^{n-j_{i-1}}\}^{r},
\end{aligned}
$$

and therefore (11) is verified. We obtain (10) as a particular case letting $h_i \equiv 1$.