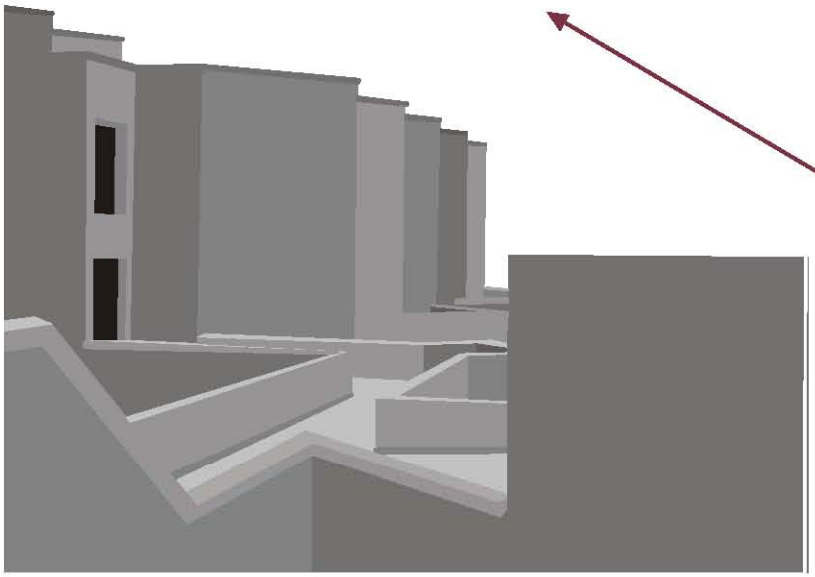

Comunicaciones del CIMAT

ESTIMATION OF SPATIAL SAMPLING EFFORT BASED ON PRESENCE-ONLY DATA OVER A CLASS OF SPECIES

Miguel Nakamura, Daniel Fernández and Jorge Soberón

Comunicación del CIMAT No I-08-11/21-05-2008
(PE /CIMAT)

CIMAT

Estimation of Spatial Sampling Effort Based On Presence-Only Data over a Class of Species

Miguel Nakamura
Centro de Investigación en Matemáticas, A.C.

Daniel Fernández
Centro de Investigación en Matemáticas, A.C.

Jorge Soberón
University of Kansas

May 20, 2008

Abstract

Sampling bias contained in data of biological surveys is very common. Bias is clearly a function of roads, cities, rivers, or other physical features that determine accessibility of collectors, and many data sets of species are presence-only. We set out to estimate spatial sampling bias on a region, based on presence-only data, by explicitly incorporating information on these accessibility factors, and by considering a class of species that may share a common pattern of search. We also resort to the concept of species richness, in order to estimate, indirectly, number of individuals. We construct a probabilistic (multinomial) model that enables standard likelihood inference procedures to be implemented. Simulation scenarios for exploration of the model and experimenting with the estimation procedure are included. Illustrative examples over a region of Mexico with mammals and butterflies are also discussed.

Introduction

Specimens in biological collections constitute one of the largest sources of biological data in the world, in the order of 1.6×10^9 specimens (Chalmers 1996). These collections were almost universally performed in a non-systematic way (Peterson et al. 1998; Soberón et al. 2000). There is seldom any attempt to perform taxonomic collecting following the rules of statistical sampling.

Often the locations where a data point occurs and the intensity of effort are not controlled as in a traditionally designed sampling survey. This gives rise to the question of *sampling effort* over a region, or a bias in geographical space, induced by the way in that biological data is collected. The preceding concept of sampling effort is complicated further by the fact that many data sets of species are *presence-only*. This means that if sampling effort was spent at a location without a species being recorded, that effort is not necessarily tracked.

This paper deals with the issue of quantifying sampling effort for presence-only data. A fundamental notion to be exploited here is that of a *class* of species. This will be a group of species for which taxonomists share some common pattern of search. For example, mammals of certain sizes in a given region are similarly collected, using the same, general methods, reflecting the idiosyncrasies of mammalogists. The idea is that presence of a large ensemble of specimens in a region is an indicator of the sampling effort having been exercised there. This notion is not available when considering presences of a *single* species, such as when using presence-only data in the prediction of ecological niches (Graham et al. 2004; Peterson 2001).

Single-species data is inherently low density over an area, and because absences are not available, sampling effort cannot be precisely inferred. A similar notion of a class of species has been called upon for approaches to better simulate pseudo-absences in niche prediction methodology (Zaniewski et al. 2002) or to construct prediction models for the presence of a single species (Ferrier et al. 2002).

Figure 1 shows two examples of presence-only data regarding broad classes of species—butterflies and mammals—over a region of central Mexico (roughly one-fourth of the country, spanning parts of the States of Coahuila, Aguascalientes, Durango, Guanajuato, Nayarit, Nuevo León, San Luis Potosí, Tamaulipas, Jalisco, and Zacatecas) that will be used to exemplify throughout. A 70 (longitude) by 54 (latitude) system of 5km grid points is defined on the region. The locations of main paved highways and large cities (population at least 40,000) are identified. For butterflies—an example to be fully developed below—we use a medium size database of the 177 acknowledged species and subspecies of the Papilionid and Pierid butterflies of Mexico. This database originally referred to about 55,000 specimens and was obtained from visiting the 25 most important museums in the world that hold Mexican collections. Data from an extensive literature review that probably covers most of the non-digitized holdings was also included. The data has been checked for taxonomic and geographical consistency. Llorente *et al.* (1998)

present a full description of the database. The mammals database is a compilation, provided by the Comisión Nacional de Biodivesidad (CONABIO) of Mexico, of near 300,000 georeferenced specimens of 450 species of terrestrial mammals

It is quite apparent and natural that these presences tend to concentrate geographically around roads and towns (Bojórquez-Tapia et al. 1995; Soberón et al. 2000; Wohlgemuth 1998), and that the larger the town, the larger the concentration. Puerto Vallarta, in the lower-left corner, for example, displays a conspicuous larger amount of presences in its vicinity. Furthermore, there are distinct characteristics between both groups. Mammals appear to be collected in more remote areas than butterflies, and along secondary roads in addition to primary highways.

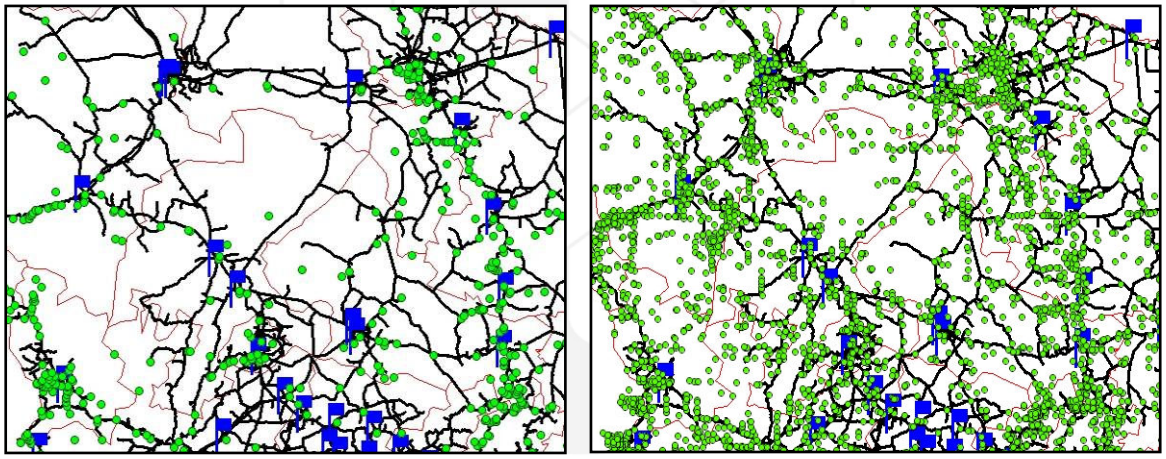

Figure 1: Presence-only data for all butterflies (left) and mammals (right), in central Mexico. Black lines are main paved roads and flags indicate main cities. Soft lines are State boundaries.

Our main motivation in this work resides in the estimation of environmental sampling bias, a concept to be plugged-in directly into methodology for the prediction of a single-species niche as a function of environmental variables. The methodology we bear in mind was developed by Argaez *et al*. (2005); its software implementation is called BioP (www.cimat.mx/software). In this technique, geographical sampling bias is explicitly identified and required as input, with the purpose of constructing a niche using presence-only data for a single species. But the geographical sampling effort may be of interest in itself, for the planning of future explorations, or the assessment of the degree of knowledge of a particular region (Soberón et al. 1996), or for the generation of pseudo-absences in single-species niche predictions as in (Zaniewski et al. 2002).

Two spaces are relevant, geographical space and environmental space. See (Hirzel et al. 2002) for a detailed discussion on the role of these spaces and relationships in the context of areas of distribution. In what follows, we adopt a similar notation, some of it depicted in Figure 2. The geographical region of study is G. We assume a regular square grid over G of resolution d. The nodes in this grid are denoted by g_i. The set $\mathcal{P}$ consists of nodes where cities lie (or more generally, *point* sources of effort concentration such as biological stations, towns, *etc*.), and κ_i is a measure of the size (*e.g.* population) of a city located at node r_i. The set $\mathcal{L}$ consists of nodes that track roads (or more generally, *linear* sources of effort concentration, such as coasts, river beds, trails, *etc*.).

The environmental (multidimensional) space associated with G, will be represented by E. Its distinct elements are $e_1, e_2, \ldots, e_t$. The environment associated with node g is denoted by $e = \varphi(g)$.

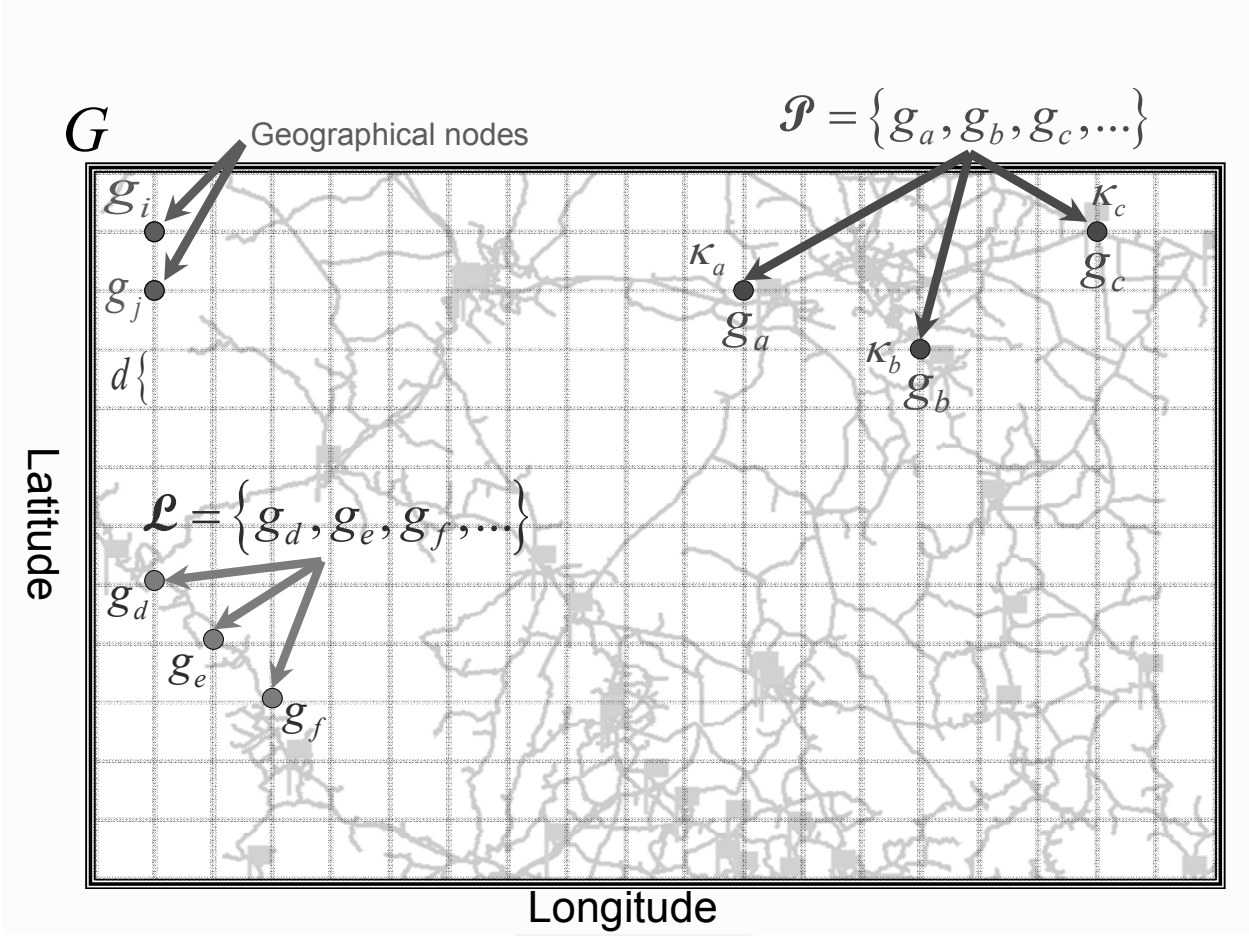

Figure 2: Grid over geographical space G, and notation used to describe linear and point concentrations of sampling effort. Quantities κ_i denote sizes of point sources, such as city population.

Modeling

Data consists of n *observed* locations of presences of all species in the class. Physically, an observed presence record at a site occurs because two parallel processes that are taking place happen concurrently: (a) the site is visited by humans; (b) the class is detected because the class is present. These two processes can be clearly and distinctly identified with two different actors. The presence of the class is prescribed by biological factors, whereas the site visit and subsequent detection of a specimen is essentially due to human activity. The concept of sampling bias is identified directly with process (a) above, but observed data is a combined product of both effects. Our approach begins by regarding both of these processes as random (and independent), so that probabilities become natural tools of description. The key is to relate observed data to geographical bias via a probability model.

We conceive the notion of a *mask* over geographical/environmental space that can be defined from biological considerations. This mask is a subset of geographical space outside of which the occurrence of the class is impossible. For example, if the class is butterflies, it may be reasonable to assume that no butterfly populations can be found above an altitude of 4,000 meters (stray individuals may sometimes occur in very unsuitable places). This would define the mask, as all points in geographical space with a smaller altitude. Notice that this pre-specification of a mask is quite different from specifying a niche for a single species. Class information is cruder, and we assume that such mask can indeed be specified. Notice also that a site outside the mask can in fact be sampled by humans. The relevance of the mask is to specify that presences are restricted over the mask, thus providing a further possible reason for not having recorded a positive observation at a site when visited.

The key idea in our approach is to conceptualize the set of all site visits ever put into effect as a set of individual random trials, some of which have given rise to presence records and others which have not. A two-stage probability tree (Figure 3) may be used to describe the two concurrent processes that result in obtaining a presence record. In the first stage, humans select an environment for inspection at random. In the second stage, detection is produced at the selected environment depending on features inherent to the class of species. Each stage has associated probabilities for the possible outcomes; in the first stage the outcomes are the possible environments $(e_1, e_2, \ldots, e_t)$ and in the second stage, the possible outcomes are simply success-failure, or 0–1 (either the class is detected or it is not). When a selected environment turns out to be outside the mask, then the result at the second stage is obviously a failure. Probabilities in the first stage will be denoted by $S(e_i;\gamma)$ and probabilities at the second stage will be denoted by $D(e_i;\eta)$ and $1-D(e_i;\eta)$. The parameters γ and η are introduced to allow for flexibility in the specification of probabilities. Suggestions for explicit forms of D and S will be given below.

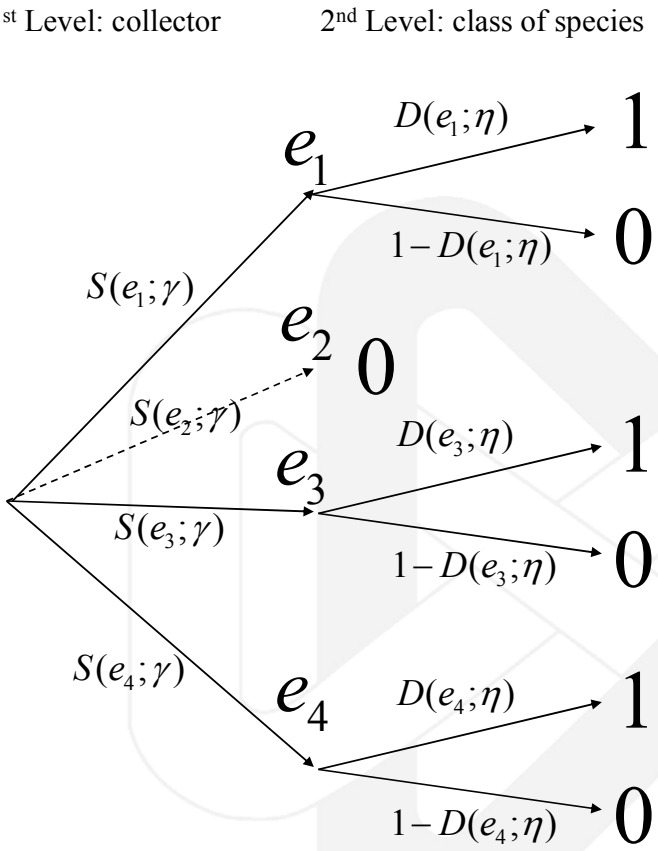

Figure 3: Two-stage probability tree used to conceptualize sampling effort that has given rise to presence records (1) or lack of presence records (0). In the 1st stage, collectors indirectly select an environment to visit, with probabilities that depend on physical accessibility. In the 2nd stage, detection is success-failure, with probabilities depending on characteristics of the class of species as related to the sampled environment. The dotted arrow means an environment that lies outside the mask, which therefore cannot give rise to a presence record.

In order to generate the set of observed presences, a large, unknown number of site visits, N, is presumed to have been expended. Each of these site visits constitutes a trial that navigates through the probability tree described previously. In the end, only n observed presences for the whole class are recorded, and these are distributed at random over environments within the mask. Site visits that concluded in "no presence detected" (either because the class was present but it was not detected, or because the environment was outside the mask) are not identified nor counted. Notice that it is important to keep track of multiple presences at a given site, because these can very well occur due to intense sampling effort and it is sampling effort we wish to

estimate. In single-species niche prediction, multiple presences are often collapsed into a single determination of presence, impairing a legitimate analysis of effort.

The idealized random process described above is in fact a multinomial experiment. That is, N trials are classified at random into one of several bins. One of these bins collects all zeros, $N-n$ of them. The other bins are labeled by the environments within the mask (t of them), and altogether add up to n counts corresponding to the n observed presence records. The probabilities of being classified into one of these latter "mask" bins are $S(e_i;\gamma)\, D(e_i;\eta)$, and the probability of being classified into the "zeros" bin is $1-\sum_{i=1}^{t} S(e_i;\gamma)D(e_i;\eta)$. Notice that although the number of bins may be quite large, all their probabilities are described with only two parameters, γ and η.

Once probabilities have been specified through functions D and S, the resulting probability model associated with the observations is thus multinomial. The parameters of this distribution are γ, η, and N. The main parameter of interest is γ (because $S(e;\gamma)$ represents environmental sampling bias), so that N and η are technically nuisance parameters (in statistical inference jargon). The number of trials, N, is in fact unknown, unlike typical applications of the multinomial model. This will introduce special challenges for fitting the model based on observed data, which will be addressed when this multinomial structure is addressed in the next section.

By context and on empirical grounds, it appears sensible to let probabilities $S(e;\gamma)$ depend on the physical and transportation infrastructure, that is, roads and cities. We resort to a quantification of collector "accessibility", denoted by $C(e;\gamma)$, to describe how easily an environment e can be approached or entered, from nodes in $\mathcal{P}$ and $\mathcal{L}$. There are many ways one might proceed to do this. We use a simple measure based on straight-line distances on the plane, but more elaborate measures could be easily considered, where other characteristics of terrain that affect accessibility could be incorporated such as mountain ranges, rivers, *etc*. Define the "distance" between two nodes as follows:

$$\delta\left(g_i, g_j\right) = \begin{cases} d & \text{if } g_i = g_j \\ \left\| g_i - g_j \right\| & \text{if } g_i \neq g_j \end{cases},$$

where $\|\cdot\|$ denotes Euclidian distance in geographical space. This definition incorporates the discreteness of the grid; if two presences share a pixel, we arbitrarily assign the distance d because all we know is that they are within the corresponding square. Now define a linear combination of weighted sums of "road" and "city" terms, as follows:

$$C^*(g;\gamma) = \gamma \sum_{g_j \in \mathscr{P}} \left(\delta\left(g, g_j\right)\right)^{-1} \kappa_j + \sum_{g_j \in \mathscr{L}} \left(\delta\left(g, g_j\right)\right)^{-1}.$$

In this expression, a geographical point g is examined, and its reciprocal distances to nodes in $\mathscr{P}$ and $\mathscr{L}$ are accumulated. The main ideas are that a geographical node close to a road network and/or a larger city has a larger accessibility index, and that the parameter γ allows for differences in relative weights between $\mathscr{P}$ and $\mathscr{L}$. A similar notion of "environmental accessibility" is obtained by accumulating these geographical accessibilities, by setting

$$C(e;\gamma) = \sum_{\{g_i : \varphi(g_i) = e\}} C^*(g_i;\gamma).$$

Once this measure of accessibility is established, rescaling to probabilities can be obtained by normalization. In this way, we let

$$S(e;\gamma) = \tau_e(\gamma) C(e;\gamma),$$

where $\tau_e(\gamma) = \left(\sum_{z=1}^{t} C(e_z;\gamma)\right)^{-1}$. This is just saying that the probability of inspecting an environment is directly proportional to its accessibility.

Specification of $D(e;\eta)$ is another matter, as its justification must lie in the realm of biology. Here, we require a way to describe the probability of detecting a specimen of the class of species, as a function of the environment. The way we propose to do this is to relate $D(e;\eta)$ to the concept of *abundance*, via species *richness* for the class.

Obtaining the number of individuals of almost any species in a region, with the exception of conspicuous plants, or large animals, is a very difficult practical problem. One way to get an estimation of number of individuals, to order of magnitude, is to resort to the canonical log-normal approximation to the distribution of species abundance which has a long tradition in ecology (May 1975). Both theory and evidence suggest that, for many taxonomical groups, the number of individuals per species follow a log-normal distribution. Moreover, Preston (1962) proposed the "canonical hypothesis" that allows to fix one of the two parameters of the lognormal reducing it to a single parameter distribution. This allows calculation of the number of

individuals in a community as a function of the number of species. Equations B.1 and B.2 in (May 1975) applied to each g allow the estimation of the number $A(g)$ of individuals in a community composed of $R(g)$ species, as long as $R(g) \gg 1$ (in practice, at least 10 species) and the Canonical Hypothesis holds. Using the definition $\text{erf}(x) = 2\pi^{1/2} \int_0^x \exp(-t^2)dt$, May gives

$$R(g) \simeq [2\pi^{1/2} / \ln(2)]\Delta_g \exp(\Delta_g{}^2) \qquad (1.1)$$

and

$$A(g) \propto I(g) = [\pi^{1/2} / \ln(2)]\Delta_g \exp(4\Delta_g{}^2)\,\text{erf}(2\Delta_g), \qquad (1.2)$$

where the parameter Δ_g, which is related to the mode in the distribution of species abundance, can be used to solve implicitly for $A(g)$ as a function of $R(g)$. May (Equation B.3) suggests the further approximation $[\pi^{1/2} / \ln(2)]\Delta_g \exp(4\Delta_g{}^2)$, but we used (1.2) in the calculations below.

Species richness, $R(g)$, is information that can be estimated using a variety of ways. For example, by regressing well sampled localities against environmental variables (Iverson and Prasad 1998; Lobo and Martin-Piera 2002; Wohlgemuth 1998). Soberón and Llorente (unpublished) obtained such a relation for the butterflies of Mexico, at a grid resolution of about 10km × 10km:

$$R(g) = 100 + 0.0089Pre(g) - 0.0097Alt(g) - 2.70Lat(g) - 0.824T_{\min}(g)$$

where *Pre(g)* = annual precipitation in mm, *Alt(g)* = average elevation in meters above sea level, *Lat(g)* = coordinates of North Latitude and $T_{\min}(g)$ = Minimum monthly average temperature. This regression has an $r^2 = 0.68$

The essential point is that we assume $R(g)$ over G to be a known piece of information, and that abundance, $A(g)$ can be deduced accordingly. This will provide the crucial connection needed between a biological concept and $D(e;\eta)$, which will ultimately enable us to say something about effort based on presence-only records. The deduction is, for each g, first solve (1.1) numerically for Δ_g. Then substitute this value in (1.2), to compute $I(g)$. A notion proportional to abundance at the environment e is then obtained as

$$A(e) = \sum_{\{g_i : \varphi(g_i) = e\}} I(g_i).$$

Accepting this as a working approximation, we next postulate a form for $D(e;\eta)$ as a function of $A(e)$. Striving for a parsimonious formulation (a one-dimensional parameter), we assume the probability of detecting a member of the class is equal to one if abundance exceeds a certain threshold. Below that threshold, the relationship is assumed to be proportional. This assumption amounts to specifying

$$D(e;\eta) = \begin{cases} \eta A(e) & \text{if } A(e) \le \dfrac{1}{\eta} \\ 1 & \text{if } A(e) > \dfrac{1}{\eta} \end{cases},$$

so that $1/\eta$ represents the threshold and η the proportionality constant.

All things assembled result in a three-parameter multinomial model, that explains presence records for the class, using information contained in $\mathcal{P}$, $\mathcal{L}$, and $R(g)$ over G. To grasp the significance of this probability model and the role of its parameters, we illustrate simulated presence records for the class "butterflies" over the region depicted in Figure 1. Environmental layers used are temperatures (mean, mean max, mean min, absolute min, absolute max), as well as precipitation, altitude, humidity, climate type, and soil type. The environmental mask is taken to be the region itself, that is, the class butterflies is assumed to be possible everywhere.

Three parameter settings that modulate the relative weight ascribed to roads and cities and detection as a function of richness are illustrated in Figure 4, all with $N = 750$: Setting #1 has $\gamma = 5/1000$, $\eta = 0.0025$; Setting #2 has $\gamma = 5000$, $\eta = 4.258\text{E}-6$; Setting #3 has $\gamma = 5/1000$, $\eta = 4.258\text{E}-6$. Simulation means that 750 artificial trials are made to run through the two-stage probability tree with specified probabilities $S(e_i;\gamma)$ and $D(e_i;\eta)$, and whenever a success results in the second stage, a new dot on the map is produced. There are less than 750 points on each of these maps (550 for Setting #1, 252 for Setting #2, 261 for Setting #3), due to simulated failures that amount to visits with no detection.

Setting #1 is a situation where roads and cities play a decisive role in the recording of presences. Good accessibility gives rise to presences, despite the fact that species richness is relatively smaller. In contrast, presences in Setting #2 are mainly driven by the species richness, and accessibility plays little role. Setting #3 is somewhat intermediate; it is driven by richness but accessibility still produces a few isolated presences in less-rich areas. Despite our simplistic

measures of accessibility, it is quite remarkable that the general character of these clusters of simulated points are able to reflect the general features displayed in Figure 1 (particularly Setting #1), in that points are distributed at random over the region, but clustered more around towns and roads and where butterflies abound.

What are realistic values of γ, η, and N for butterflies and mammals, and are there distinguishing differences amongst collectors? This entails parameter estimation based on observed data, to be addressed in the next section.

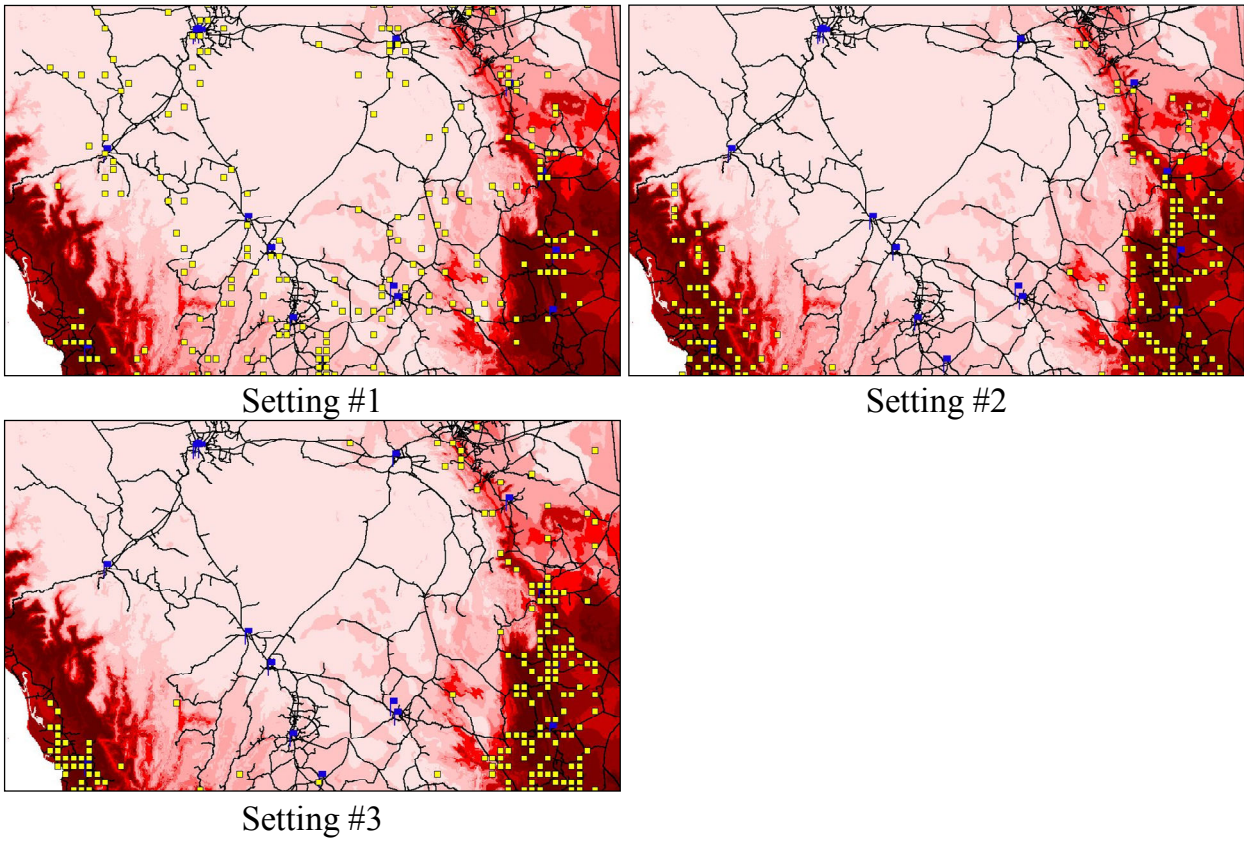

Figure 4: Simulated presence records for class butterflies, using the stated multinomial probability model, under parameter Settings #1, #2, and #3. The shades of red in the background correspond to actual species richness over the region. Dots are simulated presences taking into account the accessibility structure as a function of pictured roads and cities, as well as the relationship of detection with species richness.

Fitting the model

As described above, the intrinsic probability model is multinomial. We conceived N trials giving rise to counts over $t+1$ bins, where the total number of counts over the first t bins is n. Let $X_1, X_2, \ldots, X_t, X_{t+1}$ denote the observed counts. The number of (unobserved) zeros is $X_{t+1} = N - \sum_{i=1}^{t} X_i = N - n$. Based on a multinomial density, one can then immediately write down the likelihood function of observed counts as a function of the parameters γ, η, and N:

$$L(\gamma, \eta, N) = P[X_1, \ldots, X_{t+1}; \gamma, \eta, N] =$$

$$\frac{N!}{\prod_{i=1}^{t+1} X_i!}\left[1 - \sum_{i=1}^{t}\{S(e_i;\gamma)D(e_i;\eta)\}\right]^{X_{t+1}} \prod_{i=1}^{t}\{S(e_i;\gamma)D(e_i;\eta)\}^{X_i}.$$

This in itself provides a building block for statistical inference, particularly parameter estimation. Maximum likelihood estimates are obtained by maximizing the function L.

Because the main parameter of interest is γ, inference under presence of nuisance parameters (N and η) is relevant. One extremely useful tool for this is the *profile likelihood*, defined as

$$L_P(\gamma) = L(\gamma, \hat{\eta}(\gamma), \hat{N}(\gamma)),$$

where the values $\hat{\eta}(\gamma)$ and $\hat{N}(\gamma)$ maximize $L(\gamma, \eta, N)$. This last maximization must be achieved numerically for each fixed value of γ.

The point estimate $\hat{\gamma}$ is obtained by maximizing $L_P(\gamma)$. But most importantly, instruments for quantifying uncertainty can also be found directly by considering this function. If x is a proportion, an $x \times 100\,\%$ *likelihood interval* is defined to be the set of all values of γ such that $L_P(\gamma)/L_P(\hat{\gamma}) \geq x$. This in itself is meaningful for statistical inference, and using the standard approximation of $-2\log(L_P(\gamma)/L_P(\hat{\gamma}))$ via the χ^2 distribution with one degree of freedom, it is easy to show that the value $x = 0.14$ in a likelihood interval gives rise to an approximate 95% confidence interval for the one dimensional parameter γ. See Sprott(2000) and Kalbfleish (1985) for further details. The function $R_P(\gamma) = L_P(\gamma)/L_P(\hat{\gamma})$ is called the *relative profile likelihood*.

An interesting experiment is to explore if the arbitrarily specified values of parameters can be recovered by using our simulated data sets depicted in Figure 4. Using these simulated data points and the profile likelihood approach for parameter estimation we obtain estimated values $(\hat{\gamma},\hat{\eta},\hat{N}) = (0.0043, 0.0026, 913)$ for Setting #1 and $(\hat{\gamma},\hat{\eta},\hat{N}) = (0.0037, 4.1\text{E}-6, 1007)$ for Setting #3. Relative profile likelihood plots for γ under each setting are presented in Figure 5. The true values of γ are located well within the 0.14 likelihood intervals, confirming that estimation is not only feasible but surprisingly precise for the parameter of interest. Note, however, that errors in the estimates for N are relatively less precise.

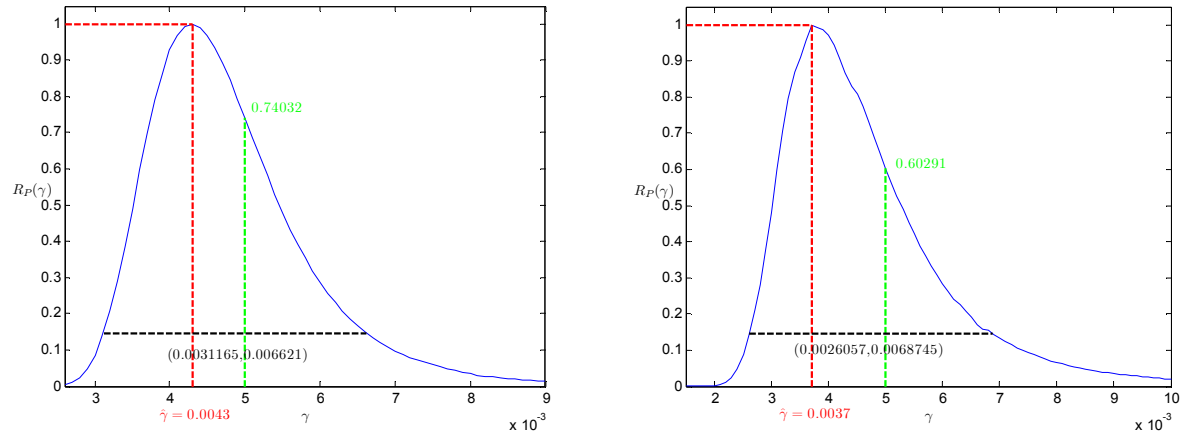

Figure 5: Relative profile likelihood plots for main parameter of interest, γ, for Setting #1 (left) and Setting #3 (right). The corresponding 0.14 likelihood intervals (approximate 95% confidence intervals), also plotted, are (0.0031, 0.0066) for Setting #1 and (0.0026, 0.0069) for Setting #3. The maxima of the curves correspond to the maximum likelihood estimate based on simulated data shown in Figure 4. The true value of γ, used in simulations, is also shown.

We now present results of this estimation process over our study region, for butterflies and mammals, using data of Figure 1. The point estimates of the parameters are $(\hat{\gamma},\hat{\eta},\hat{N}) = (5.8\text{E}-6, 0.0416, 3375)$ for butterflies, and $(\hat{\gamma},\hat{\eta},\hat{N}) = (1.05\text{E}-3, 2.1\text{E}-6, 1781)$ for mammals. Corresponding 0.14 likelihood intervals for γ are (4.0E-6, 8.2E-6) for butterflies and (8.2E-4, 14.6E-4) for mammals. Effects of the resulting probabilities $S(e;\hat{\gamma})$ are shown in Figure 6, in geographical space. A node g is here plotted using a standardized color intensity-scale based on $C^*(g;\hat{\gamma}) / \sum C^*(g_i;\hat{\gamma})$. It is interesting to note the very contrasting differences in sampling effort that are apparent between these two groups. Butterflies display a hotspot towards

the center of the region whereas effort for mammals is heavily concentrated along mountain ranges. It is indeed true that the patterns of collections of butterflies and mammals in Mexico are widely different, with the former concentrated along highways and a few field stations and large cities (Soberón et al. 2000), and the latter much more regularly dispersed over the surface of the country (Medellin, pers. com.). The method we present here allows precise quantification of such contrasting patterns.

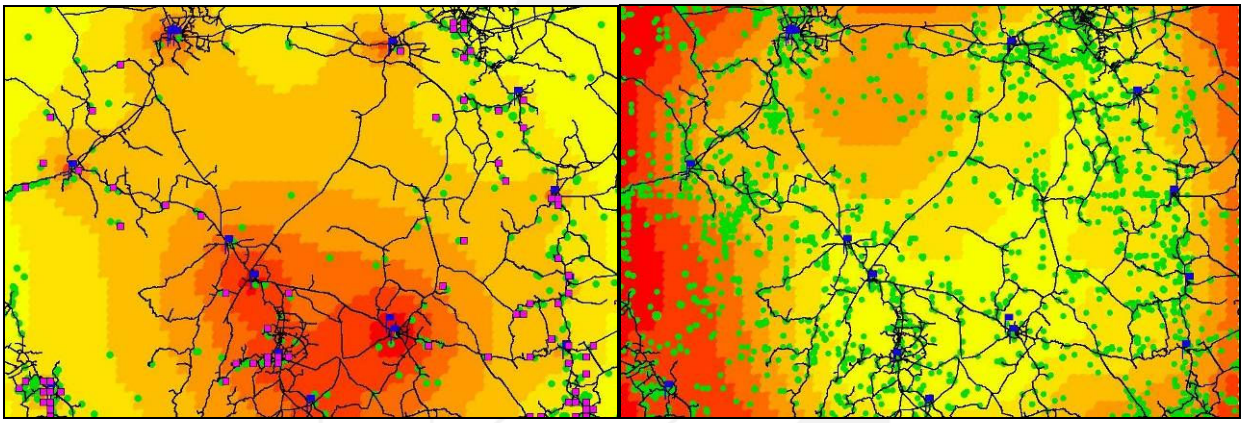

Figure 6: Estimated effort, $S(\varphi(g);\hat{\gamma})$, in geographical space. Butterflies at left, and mammals at right. Green dots are presences of the corresponding class. In readying an example in niche prediction to be discussed below, observed presences of the single butterfly species *Zerene cesonia cesonia* are superimposed on the left panel, using purple squares.

Discussion

At present, the number of "presence-only" registers of species that biologists can access is very large, and increasing. The Global Biodiversity Information Facility (GBIF) alone provides access to more than 10^8 records, most of them georeferenced (www.gbif.org) (Edwards 2004). However, generally speaking, this type of data has very significant sampling problems, since as we said, effort is not randomly distributed in space, and it is very uneven, with log-normal distributions of number of records per sample site (Graham et al. 2004; Iverson and Prasad 1998; Mora et al. 2008; Soberón et al. 2007; Soberón et al. 2000), and the treatment of. Conventional methods for describing and analyzing such kinds of data are not enough (Pearce and Boyce 2006) and new approaches are needed, if biologist are going to make sense of this large and significant mass of data.

One avenue of development is to acknowledge explicitly the particularities of taxonomic data and develop: 1) ways to quantify the bias, and 2) procedures to extract as much information as possible from such biased but very abundant data. In this note we have shown how two sources of bias can be quantified by using ancillary data. $S(e;\hat{\gamma})$ is estimated by accessibility to roads and cities, since the "highway effect" is a well known problem of taxonomical collections. The detectability function, $D(e;\eta)$ is estimated by using another large set of data, namely the entire collection of butterflies, which represent the accumulated effort of many decades of scientists working over all the states of Mexico.

What effect does sampling bias have on niche prediction for an individual species, based on presence-only data? BioP (Argaez et al. 2005) recognizes the role of non-homogeneous sampling effort and explicitly takes it into account. Figure 7 displays two different end results for predicted niches for sp. *Zerene cesonia cesonia* (whose presence records are shown in the left panel of Figure 6), one completely disregarding sampling effort (that is, assuming it has been uniform in geographical space) and another using the estimated effort for class butterflies developed previously. Although similar, there are a few notable, local, differences. When effort is explicitly taken into account, greater probability of presence is granted in both the NE and SW corners. But interpretation of the left panel in Figure 6 indicates that this conclusion follows different reasons. In the SW corner, an important number of presences of the species are found, but the sampling intensity for the class is relatively higher. In the NE corner, only a few presences occur nearby, but with less expended effort. What BioP is doing is to re-weigh empirical evidence according to sampling effort.

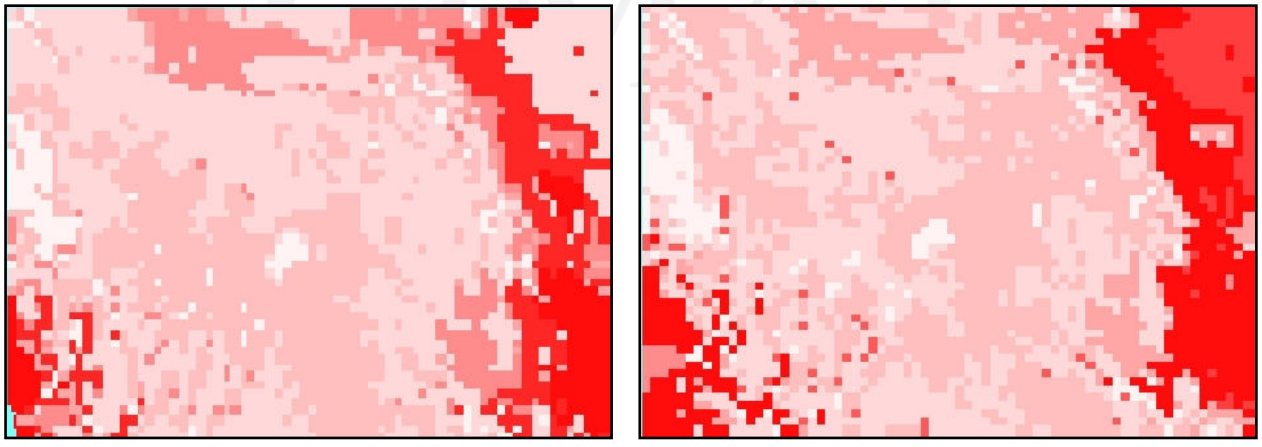

Figure 7: Example of BioP predicted distributions for *Zerene cesonia cesonia* over study region with sampling effort assumed uniform in geographical space (left), and using estimated sampling effort (right).

All computations regarding sampling effort and likelihood methods were performed using Matlab and did not present critical issues. Numerical maximizations were achieved using the Nelder-Mead simplex search algorithm, which does not require any derivatives (fminsearch function in Matlab) and were well-behaved for the three-dimensional parameter. Regarding the discrete parameter N, embedding into a continuous interpolating function was achieved via the relationship $\ln(N!) = \ln(\Gamma(N+1))$. It was setting up multiple data sources via GIS that was more problematic. Estimation of effort is a specific module to be soon included within the BioP software.

Acknowledgments

The authors wish to thank Raúl Jiménez at CONABIO for supplying mammal data and richness. Maarten Bladt contributed with discussion and Christian O. Gómez helped with BioP implementations.

References

Argaez, J., A. Christen, M. Nakamura, and J. Soberón. 2005. Prediction of potential areas of species distributions based on presence-only data. Environmental and Ecological Statistics 12:27-44.

Bojórquez-Tapia, L. A., I. Azuara, E. Ezcurra, and O. Flores-Villela. 1995. Identifying conservation priorities in Mexico through geographic information systems and modeling. Ecological Applications 5:215-231.

Chalmers, N., R. 1996. Monitoring and inventorying biodiversity: collections, data and training, Pages 171-179 *in* F. di Castri, and T. Younes, eds. Biodiversity, Science and Development: Towards a New Partnership. Wallingford, CAB International.

Edwards, J. 2004. Research and societal benefits of the global biodiversity information facility. BioScience 54:485-486.

Ferrier, S., G. Watson, J. Pearce, and M. Drielsma. 2002. Extended statistical approaches to modelling spatial pattern in biodiversity in northeast New South Wales. I. Species-level modelling. Biodiversity and Conservation 11:2275-2307.

Graham, C., S. Ferrier, F. Huettman, C. Moritz, and A. T. Peterson. 2004. New developments in museum-based informatics and applications in biodiversity analysis. Trends in Ecology and Evolution 19:497-503.

Hirzel, A. H., J. Hausser, D. Chessel, and N. Perrin. 2002. Ecological-niche factor analysis: how to compute habitat-suitability maps without absence data? Ecology 83:2027-2036.

Iverson, L. R., and A. Prasad. 1998. Estimating regional plant biodiversity with GIS modeling. Diversity and Distributions 4:49-61.

Kalbfleish, J. G. 1985, Probability and Statistical Inference. New York, Springer Verlag.

Lobo, J. M., and F. Martin-Piera. 2002. Searching for a predictive model for species richness of Iberian dung beetle based on spatial and environmental variables. Conservation Biology 16:158-173.

May, R. M. 1975. Patterns of species abundance and diversity, Pages 81-120 *in* M. L. Cody, and J. M. Diamond, eds. Ecology and Evolution of Communities. Cambridge, Massachusetts, Belknap Press.

Mora, C., D. P. Tittensor, and R. A. Myers. 2008. The completeness of taxonomic inventories for describing the global diversity and distribution of marine fishes. Proceedings of the Royal Society of London, B. 275:149-155.

Pearce, J., and M. S. Boyce. 2006. Modelling distribution and abundance with presence-only data. Journal of Applied Ecology 43:405-412.

Peterson, A. T. 2001. Predicting species' geographic distributions based on ecological niche modeling. Condor 103:599-605.

Peterson, A. T., A. Navarro-Siguenza, and H. Benitez-Diaz. 1998. The need for continued scientific collecting; a geographic analysis of Mexican bird specimens. Ibis 140:288-294.

Preston, F. W. 1962. The canonical distribution of commonness and rarity. Ecology 43:185-215,410-432.

Soberón, J., R. Jiménez, J. Golubov, and P. Koleff. 2007. Assessing completeness of biodiversity databases at different spatial scales. Ecography 30:152-160.

Soberón, J., J. Llorente, and H. Benítez. 1996. An international view of national biological surveys. Annals of the Missouri Botanical Garden 83:562-573.

Soberón, J., J. Llorente, and L. Oñate. 2000. The use of specimen label databases for conservation purposes: An example using Mexican Papilionid and Pierid butterflies. Biodiversity and Conservation 9:1441-1466.

Sprott, D. A. 2000, Statistical Inference in Science. New York, Springer Verlag.

Wohlgemuth, T. 1998. Modeling floristic species richness on a regional scale: A case study in Switzerland. Biodiversity and Conservation 7:159-1977.

Zaniewski, E. A., A. Lehman, and J. M. C. Overton. 2002. Predicting species spatial distributions using presence-only data: a case study of native New Zealand ferns. Ecological Modelling 157:261-280.