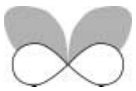
Journal of Biogeography, **30**, 1719–1727



Geographical sampling bias and its implications for conservation priorities in Africa

Sushma Reddy[1]* and Liliana M. Dávalos[1,2] [1]*Division of Vertebrate Zoology, American Museum of Natural History, New York, USA; and Department of Ecology, Evolution, and Environmental Biology, Columbia University, New York, USA and* [2]*Center for Biodiversity and Conservation, American Museum of Natural History, New York, USA*

Abstract

Aim To design and apply statistical tests for measuring sampling bias in the raw data used to the determine priority areas for conservation, and to discuss their impact on conservation analyses for the region.

Location Sub-Saharan Africa.

Methods An extensive data set comprising 78,083 vouchered locality records for 1068 passerine birds in sub-Saharan Africa has been assembled. Using geographical information systems, we designed and applied two tests to determine if sampling of these taxa was biased. First, we detected possible biases because of accessibility by measuring the proximity of each record to cities, rivers and roads. Second, we quantified the intensity of sampling of each species inside and surrounding proposed conservation priority areas and compared it with sampling intensity in non-priority areas. We applied statistical tests to determine if the distribution of these sampling records deviated significantly from random distributions.

Results The analyses show that the location and intensity of collecting have historically been heavily influenced by accessibility. Sampling localities show dense, significant aggregation around city limits, and along rivers and roads. When examining the collecting sites of each individual species, the pattern of sampling has been significantly concentrated within and immediately surrounding areas now designated as conservation priorities.

Main conclusions Assessment of patterns of species richness and endemicity at the scale useful for establishing conservation priorities, below the continental level, undoubtedly reflects biases in taxonomic sampling. This is especially problematic for priorities established using the criterion of complementarity because the estimated spatial costs of this approach are highly sensitive to sampling artefacts. Hence such conservation priorities should be interpreted with caution proportional to the bias found. We argue that conservation priority setting analyses require (1) statistical tests to detect these biases, and (2) data treatment to reflect species distribution rather than patterns of collecting effort.

Keywords
Africa, biodiversity hotspots, birds, conservation priority setting, locality records, sampling

INTRODUCTION

Analyses of species distributions to identify areas of priority for conservation of biological diversity have become a standard approach to reconcile species preservation goals with the limited resources available for protecting and managing natural habitats. Conservation biologists have thus become intent on searching for the most efficient way to represent the greatest number of species in the fewest areas (Reid, 1998; Myers *et al.*, 2000). Prominent among criteria for setting conservation priorities is complementarity. Complementarity explicitly describes the degree to which an

*Correspondence: Sushma Reddy, Division of Vertebrate Zoology, American Museum of Natural History, Central Park West at 79 Street, New York, NY 10024-5192, USA. E-mail: sushma@amnh.org

area contributes taxa otherwise not represented to a set of areas targeted for conservation (Williams *et al.*, 1996). This criterion generates a minimum set of areas whose protection will conserve a maximum number of taxa, by definition, the most area-efficient approach to conservation planning (Williams, 1996). Because the search for such areas is implemented through heuristic algorithms that reasonably approximate the goal of including all species, a near-minimum set of areas is obtained (Williams *et al.*, 1996). Given the goal of maximizing species coverage, species represented by a single locality immediately add their corresponding area to the near-minimum set of conservation priorities, making complementarity analyses highly sensitive to sampling artefacts (Faith, 2002).

Priority setting exercises employing the complementarity criterion utilize species ranges, often in grid-based spatial data bases, in their analyses (see Williams *et al.*, 1996; Williams, 1996). Species ranges, however, are abstractions of where specimens were actually collected, often considering ecological continuity or its surrogates to extrapolate from known localities to unsampled areas (Brown *et al.*, 1996). The data available for generating species ranges, and hence for conservation analyses, are necessarily incomplete (Kodric-Brown & Brown, 1993; Winker, 1996). In the face of incomplete data, minimum requirements for conservation analyses have been suggested. These include measures of precision, accuracy, and sampling bias pertaining to the spatial and temporal consistency of record collection (Williams *et al.*, 2002).

Ideally, to establish conservation priority areas, sampling effort should be uniform so that all recorded variations in distribution and abundance patterns are real and not the result of variation in sampling efforts (Williams *et al.*, 2002). Significant differences in the probability of detecting species between areas will complicate distinguishing areas that are truly high in species richness and endemism from those that are simply sampled more intensely and therefore seem unique (Nelson *et al.*, 1990). The evaluation of spatial sampling biases is indispensable to design conservation strategies and interpret their robustness and reliability. Previous statistical analyses have demonstrated the geographical biases inherent to locality records (Nelson *et al.*, 1990; Freitag *et al.*, 1998; Peterson *et al.*, 1998; Parnell *et al.*, 2003). But until now there has been no explicit measure of the geographical sampling bias or demonstration of the effect, if any, of these spatial biases on conservation priority analyses.

Designing tests for spatial sampling bias

The raw data of geographical distributions, point localities of where specimens were collected are necessary to assess biases in sampling. The locality data allow certainty as to where species have been sampled within their ranges and hence can be quantitatively tested for deviations from random or even distribution of collecting localities. Such quantitative analyses cannot be performed on distributional ranges. Locality records also provide some measure of sampling effort within the range of each species.

The first test we designed examined whether sampling points were biased by accessibility. That is, are sampling points closer to areas of human habitation and means of transportation? If so, that would mean that sampling in this region is not even or random, but skewed towards areas that are more accessible to collectors. Although this test can detect if sampling points are biased, it cannot establish whether or not the conservation priority areas are influenced by this bias. For instance, many sampling points may be clustered around rivers, because these happen to also be the areas supporting greater biodiversity, independent of human observation.

We designed another method that allows us to test for sampling biases as they relate to conservation priority areas, while accounting for the fact that different numbers of species exist in different regions of a continent. In other words, how might geographical conservation priority areas, a set of areas designated to having the greatest cumulative species richness, be affected by sampling bias? Simply comparing the distribution of point localities of all species might be able to pinpoint areas that are better sampled than others, however there would be no way of distinguishing if this is because of the different number of species that exist in different areas. The second test we designed was able to control for this by evaluating the sampling within the range of each species. If there were no bias in sampling, then we would expect species to be sampled throughout their ranges in a manner statistically indistinguishable from random.

When comparing sampling inside and outside conservation priority areas we controlled the area of the range of each species by choosing those species that were roughly equally distributed (ratio of $1 : 1 \pm 0.25$) inside and outside the conservation priority areas. In this way, the difference in the number of sampling points inside and outside conservation priority areas cannot be attributed to the size of the range of the species. We also extended this comparison with all species in the data set by adjusting for differences in area.

An African example

Several roughly congruent sets of priority areas for conservation for sub-Saharan Africa have been proposed based on varying criteria (da Fonseca *et al.*, 2000). To test for possible biases in the identification of conservation priorities caused by biased sampling, we chose those areas obtained by explicitly analysing species distributions, independent from anthropogenical threat (da Fonseca *et al.*, 2000; Balmford *et al.*, 2001, Fig. 1). These areas, proposed as the 'Blueprint for Conservation in Africa', maximize the complementarity among the ranges of *c.* 4000 species of birds, mammals, snakes, and amphibians (da Fonseca *et al.*, 2000; Balmford *et al.*, 2001; Brooks *et al.*, 2001, Fig. 1).

In order to evaluate the extent of sampling bias inherent in species distributions and its impact on conservation priority areas, we used a large data set of point localities for sub-Saharan African passerines extracted from the Hall & Moreau (1970) atlas. This highly referenced atlas has been used to verify or establish the ranges of African passerine

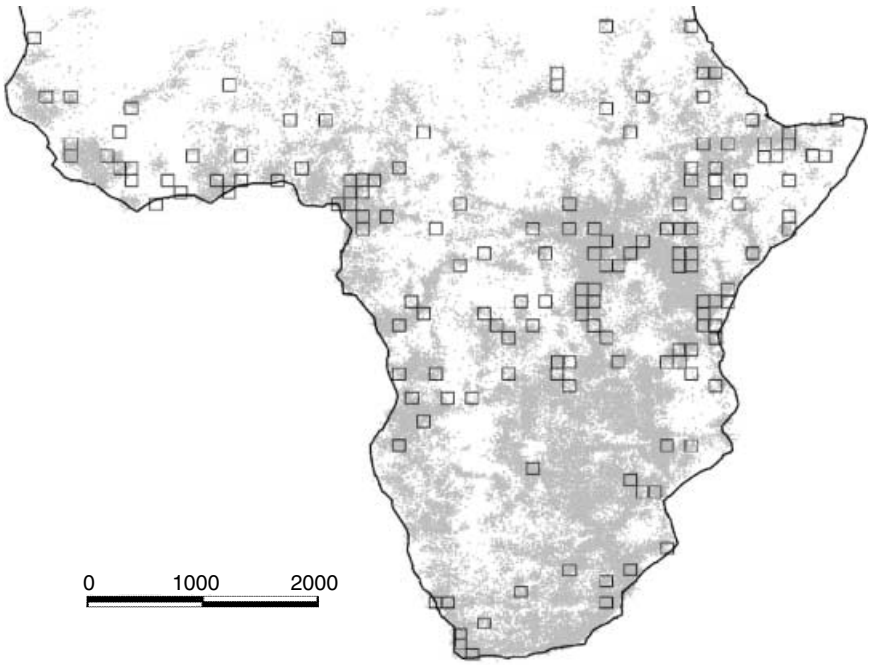


Figure 1 Map of sub-Saharan Africa showing locality points, and conservation priority areas. Point localities are represented by the grey dots. The black square outlines are the 200 conservation priority areas for sub-Saharan Africa designated using the species ranges of 4000 birds, mammals, snakes, and amphibians (da Fonseca *et al.*, 2000; Balmford *et al.*, 2001; Brooks *et al.*, 2001). Note that conservation priority areas do not necessarily correspond to areas with high density of point localities across all species (see Introduction: Designing tests for spatial sampling bias).

birds (e.g. Keith *et al.*, 1992; Urban *et al.*, 1997; Fry *et al.*, 2000). The Atlas contains over 25% of the species included in recently published conservation priority analyses (da Fonseca *et al.*, 2000; Balmford *et al.*, 2001; Brooks *et al.*, 2001), and was used to verify more recent unvouchered observations (J. Fjeldsa & C. Rahbek, pers. comm.). Therefore, our data set is a substantial and representative sample of the primary data used in those analyses. Furthermore, birds comprise the best sampled of the four vertebrate groups incorporated in the Blueprint for Africa data set (Brooks *et al.*, 2001).

In this paper, we evaluate the spatial bias of raw locality data used to define species distributions. First, we compare the distribution of these collecting localities with respect to rivers, cities and roads in sub-Saharan Africa. Next, we compare sampling density inside and outside conservation priority areas and their surroundings. Finally, we used the patterns of spatial sampling biases discovered to establish their effect on conservation priority settings.

MATERIALS AND METHODS

Locality data

We used a large data set of 78,083 point localities for 1068 species and subspecies of passerine birds digitized from a published atlas (Hall & Moreau, 1970, Fig. 1). This data set incorporates locality information from museum specimens collected since the 1800s to 1970. Each point corresponds to at least one specimen collected, and some to many. Of the total, 3504 points were geographically unique localities, because many collecting points overlap on a single locality.

The Hall & Moreau (1970) data set is one of the largest compilations of specimen collection data for sub-Saharan Africa. Nevertheless, this data set has its limitations, mainly that it concentrates on former British colonies and that it does not include specimens collected after 1970. These data still allow us to examine both geographical and historical biases. The point localities from this atlas were used to infer species ranges and subsequently in conservation priority setting exercises. Thus, the underlying biases in these primary data are carried into subsequent analyses.

Conservation priority areas

We analysed the areas that the Blueprint for Conservation in Africa (da Fonseca *et al.*, 2000; Balmford *et al.*, 2001; Brooks *et al.*, 2001) proposed as maximizing the complementarity among the ranges of *c.* 4000 species of terrestrial vertebrates. These 200 1-degree squares represent the top areas from which 97.5% of all species in the analyses have been recorded (da Fonseca *et al.*, 2000). Of these, 155 1-degree squares contain the entire distribution of one or more species and are therefore irreplaceable, while 45 could be replaced by alternative 1-degree squares (da Fonseca *et al.*, 2000). We treated both types of priority areas equally, since the majority of conservation priority areas corresponded to irreplaceable 1-degree squares, and the variation in geographical location of the remaining 45 cells in repeated prioritization analyses is small (cf. da Fonseca *et al.*, 2000; Balmford *et al.*, 2001).

Species ranges

Traditional range maps are often just simplified boundary lines, somewhat arbitrarily drawn around points of

observation and associated vegetation types. In contrast, quantitative models for predicting species distributions use explicit geophysical input parameters and produce repeatable results (Peterson *et al.*, 2000). To determine the ranges of the 1068 species included in the analysis, we used a bioclimatic model (N. Caithness and S. Reddy, unpublished data; Caithness 1995). This model uses a principal components method to predict the distributions of these species across Africa.

This bioclimatic model was implemented using Matlab v.5.10 (The Mathworks, 1996) and uses quarter-degree by quarter-degree squares as its units. Forty environmental variables consisting of minimum and maximum monthly temperature, mean annual minimum and maximum temperature, monthly rainfall, mean annual rainfall, and elevation were extracted from a climate data set (Hutchinson *et al.*, 1995) and incorporated into the model. The environmental variables at the localities where a species was collected were used as a 'training set' in a principal components analysis to find the variation in environmental factors or tolerance of a species. This variation is then used to extrapolate over all other areas with similar environmental conditions, and as such defines areas that potentially meet the autecological requirements of the species, and so represent its potential distribution. However, species are also constrained by historical events and do not always occur in all places that suit their environmental tolerance (Anderson *et al.*, 2002). For this reason, the model has another step in which the spread of point localities for each species is calculated using a probability density function. The results of both potential distribution and density of observations at each grid cell are then multiplied in order to eliminate disjunct or marginal areas that are probable over-predictions with no nearby localities. The result is the final predicted range of the species, or an extrapolation of where a species is most likely to be found taking into account where the species has already been observed to occur and its autecology.

Several models for predicting species distributions have been developed (e.g. BIOCLIM, Busby 1991; GAP, Scott & Jennings, 1998). Our model is similar to these other models in that it predicts the potential distribution of a species using environmental information, but differs in that it also incorporates a historical element. Species ranges were checked against current references (Keith *et al.*, 1992; Urban *et al.*, 1997; Fry *et al.*, 2000) to confirm the accuracy of the model's predicted ranges.

Analyses of spatial bias in sampling density

Test 1: bias due to accessibility

We obtained data sets of rivers, roads, and cities (> 50,000 inhabitants) of sub-Saharan Africa from ESRIs Digital Chart of the World (ESRI, 2000). Here we are assuming that the course of most rivers has remained unchanged and that roads and cities were at least paths and settlements during the past two centuries. We used an equal-area cylindrical projection to plot these geographical data in ArcView v. 3.2 (ESRI, 1999; Fig. 2). Next, we generated the same number of random points as unique localities ($n = 3504$) within sub-Saharan Africa using the 'random' script (Lead, 2001) in ArcView. We calculated the distance of each unique point locality and each random point to the nearest city, river, and road, using the 'assign data by spatial location, nearest' command in the 'geoprocessing wizard' function of ArcView.

We compared the distributions of distances from point localities and distances from random points from each set of

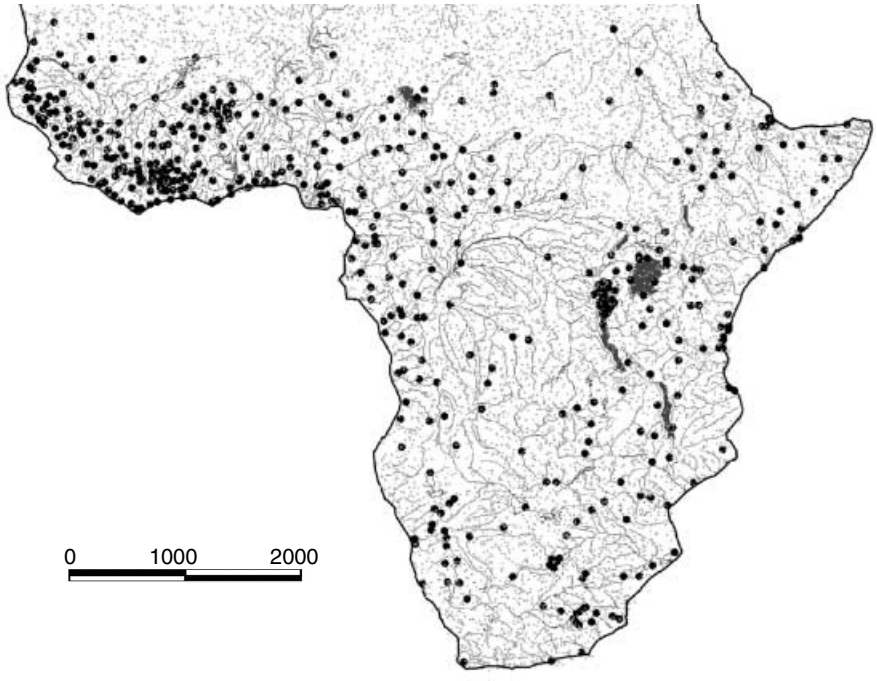


Figure 2 Map of sub-Saharan African showing cities, rivers (dark lines), and randomly generated points. Random points are represented by the grey dots, cities by black dots. Compare the distribution of rivers and cities with the pattern of historical collecting in Fig. 1.

geographical feature using two statistical tests. We applied the Kolmogorov–Smirnov two-sample test, which is a non-parametric test designed to assess whether the distributions of two samples are identical (Sokal & Rohlf, 1995, d.f. = 2). We also used the Mann–Whitney *U*-test to determine if there is a significant difference in location of these sets of ranked distributions (Sokal & Rohlf, 1995).

Test 2: bias in geographical priority areas

The first step was to find species with ranges that have roughly half their ranges inside conservation priority areas. However, conservation priority areas are much smaller than the average distribution of a passerine species and only two species were roughly equally distributed within and outside these areas. To increase sample size and broaden this analysis, we extended the priority areas to the 1-degree squares surrounding them (generally a ninefold increase in area), adding 123 species to this analysis (Fig. 3).

Using the distributional ranges generated for each species, we determined the proportion of the range that lies inside and outside extended priority areas. We then calculated the number of sampling localities in each portion of the range of each species by using the 'assign data by spatial location, inside' command in the 'geoprocessing wizard' function of ArcView. For each species, we plotted the number of sampling localities inside vs. outside priority areas. This is a simple way to visualize how much sampling has taken place per species per unit area. We performed standard major axis regressions in order to fit a line to the points in these plots (Sokal & Rohlf, 1995). We chose this method, a model II regression, because both variables were independently calculated and therefore both subject to error (Sokal & Rohlf, 1995). A test of the null hypothesis (no association) is not valid for model II regressions (Sokal & Rohlf, 1995). However, we determined if the 95% confidence interval of the slope or trendline was significantly different from the line of equivalence (slope = 1). If species were evenly sampled throughout the proportion of their ranges inside and outside conservation priority areas, or the bias in sampling was distributed independent of the determined conservation priority areas, then the points of this graph should be scattered around the line of equivalence.

To investigate the generality of this comparison, we designed equal-area comparisons for all species ($n = 1068$) in our data set. For each species, we took the total number of points in priority areas and divided it by the area of the species' range that was occupied within these areas. We then compared this to the total number of sampling localities outside of priority areas divided by the area of the range of the species that occurs outside of priority areas. We did the same for extended priority areas.

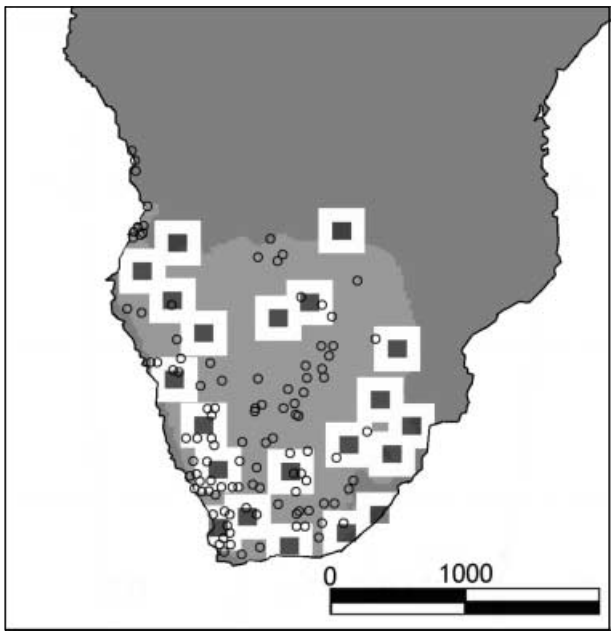


Figure 3 An example of a species range (*Eremopterix verticalis*: Alaudidae) estimated using the bioclimatic model of Caithness and Reddy (see Methods: Species ranges). The circles represent the point localities of where this species was collected. The grey shaded area represents the predicted range of this species and is similar to range depicted in Fry *et al.* (2000). The 1-degree square grey boxes inside the white boxes represent the conservation priority areas (da Fonseca *et al.*, 2000; Balmford *et al.*, 2001; Brooks *et al.*, 2001), while the white boxes surrounding these are the extended conservation priority areas (see Methods: Conservation priority areas). The range of this species is roughly equal inside and outside the extended conservation priority areas and was used in comparison 1 of test 2 (see Methods: Analyses of spatial bias in sampling density).

RESULTS

In all cases, the distance of sampling localities to the nearest river, city, and road is significantly different and closer than a random distribution (Fig. 4). There is a higher proportion of sampling localities within 1 km of cities and rivers, and far fewer points more than 1 km from cities and rivers, than expected by chance alone. Assuming that the current distribution of rivers, cities, and roads reflects accessibility and human settlement at the time specimens were collected, sampling in sub-Saharan Africa is skewed towards these features.

A standard major axis regression fitted to the density of collection within and outside extended conservation priority areas is significantly different from the line of equivalence (Fig. 5a). The regression analysis for species equally distributed inside and outside priority areas shows a strong association ($r^2 = 0.74$; $y = 0.15 + 0.667x$) such that species are overwhelmingly more sampled in extended priority areas than outside. Of 125 species compared, 103 of them were better sampled in extended priority areas. The 99% confidence interval of the slope of the standard major axis regression does not include the line of equivalence, rejecting the null hypothesis that these lines cannot be distinguished.

Standard major axis regression lines were also calculated to the densities of collection adjusted for the proportion of range area within and outside conservation priority areas, as

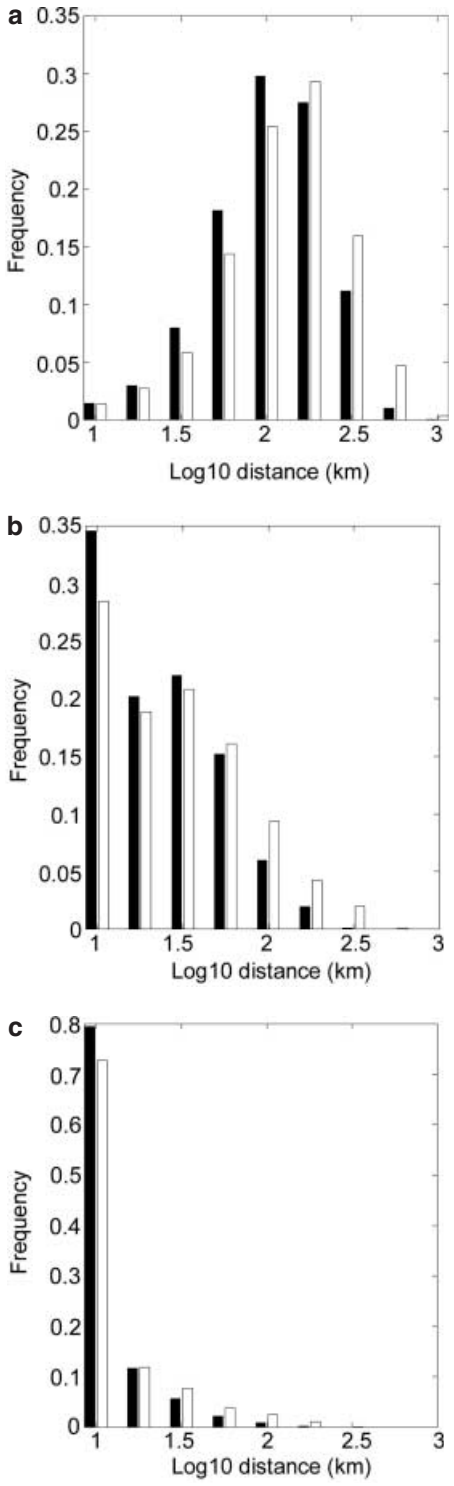


Figure 4 Distributions of distances to accessible features. Frequency distribution of distance of point localities (black) and random points (white) to the nearest (a) city, (b) river, and (c) road. In each case, the frequency distributions of distances to features from point localities and random points were significantly different ($P < 0.0001$), with distances from point localities significantly closer to features ($P < 0.0001$).

well as within and outside extended conservation priorities (Fig. 5b). Both these analyses showed a similar association such that sampling of each species has been significantly concentrated within priority areas and extended priority areas (priority areas, $r^2 = 0.58$, $y = 0.024 + 0.612x$; extended priority areas, $r^2 = 0.45$, $y = 0.001 + 0.629x$). Although we extended the priority areas in order to increase the sample size of the first comparison, the association found for sampling in the priority areas and extended priority areas indicates that the artefact of extending the areas did not strongly affect the result of the comparison. The 95% confidence intervals of the slope of standard major axis regressions fitted to these two data sets do not include the line of equivalence.

DISCUSSION

This study demonstrates the importance of developing and using statistical tools to analyse geographical data. Uses of species range data, such as priority setting of highly diverse areas, are important and constructive for conservation and biogeography studies. Nevertheless, these analyses must be interpreted according to the limitations of the data used. In light of our results, we recommend investigators, whenever possible, test their data for significant biases in sampling. The tests we designed were suited to these data, but additional methods to examine the extent of bias need to be developed. Data found to be significantly biased in sampling must be corrected using rarefaction, modelling, extrapolation and other such methods. Nonetheless, these systematic biases will not be completely overcome unless more sampling of poorly studied areas is undertaken.

Using the two simple tests, we designed, we show that there is a strong pattern of bias in sampling for passerine birds. Our simple distance analysis shows a significant skew towards accessibility in collection data, a characteristic of specimen collection in particular, and geographical distribution data in general, found by previous studies (Nelson *et al.*, 1990; Peterson *et al.*, 1998; Parnell *et al.*, 2003). This phenomenon holds despite birds being widely recognized as one of the most abundantly sampled taxonomic groups (Williams *et al.*, 1996; Brooks *et al.*, 2001). Sampling was also biased towards areas now designated as conservation priorities. While these priority areas may be especially rich in diversity, they also comprise a disproportionate amount of sampling effort.

Our comparisons show that sampling for each species has been significantly concentrated within and around priority areas (Fig. 5b). These comparisons assume equal probability of detecting a species throughout its predicted range, consistent with complementarity analyses of spatial data bases (Williams, 1996) and therefore with the priority areas we examined. All other things being equal (i.e. vagility, population size, and population density) the probability of sampling an individual organism is a function of the size of its range, so that the greater the area sampled within the range, the greater the number of records should be. The pattern of clustering of species records within priority areas

indicates that greater sampling effort lead to a higher probability of detecting species therein. This is because as more observations from a locality are accumulated, the probability of observing any one species increases. Higher sampling intensity could therefore explain the detection of more species – whether endemic or widespread, common or rare – in currently designated conservation priority areas.

Conversely, low sampling intensity underestimates the number of species present in non-priority areas, particularly small bodied, low density, or hard-to-detect species (Williams *et al.*, 2002). Because most species in a community are rare (Preston, 1948; Nelson *et al.*, 1990), the species richness and uniqueness in composition (endemism) of poorly sampled areas are essentially unknown. Setting biodiversity priorities means comparing areas with one another, and valid comparisons cannot be made unless the same relationship between sample and observed richness can be assumed to hold for all areas being compared (Williams *et al.*, 2002). This is ostensibly not the case for the data we examined. One possible explanation for the bias we found in sampling density is that areas known to have more species tend to attract more observers. That is, as an area's reputation for having high diversity spread, more investigators returned to the same site where many species were already known to exist.

Recent analyses have shown vertebrate diversity to be positively correlated to high human population density (Balmford *et al.*, 2001). Comparisons among different taxonomic groups included in the Blueprint showed that this positive correlation was higher for more abundantly sampled groups. This was interpreted as suggesting a limited role for sampling bias (Balmford *et al.*, 2001, p. 2617). But such comparisons among taxonomic groups do not evaluate for geographical sampling bias. Abundance, or *quantity* of records, is a different characteristic of sampling from spatial skewness, or distributional *quality* of records (Sokal & Rohlf, 1995). In light of our results, this correlation can also be attributed to sampling being biased towards populated areas. Therefore, an alternative explanation would then be that collectors stay within a short distance of inhabited areas, or research facilities (see Nelson *et al.*, 1990), producing the observed pattern. Furthermore, analyses showing biodiversity to be positively correlated to anthropogenic factors such as high human population and habitat modifications (Cincotta *et al.*, 2000; Balmford *et al.*, 2001) might actually be the result of bias in sampling in areas that are more accessible.

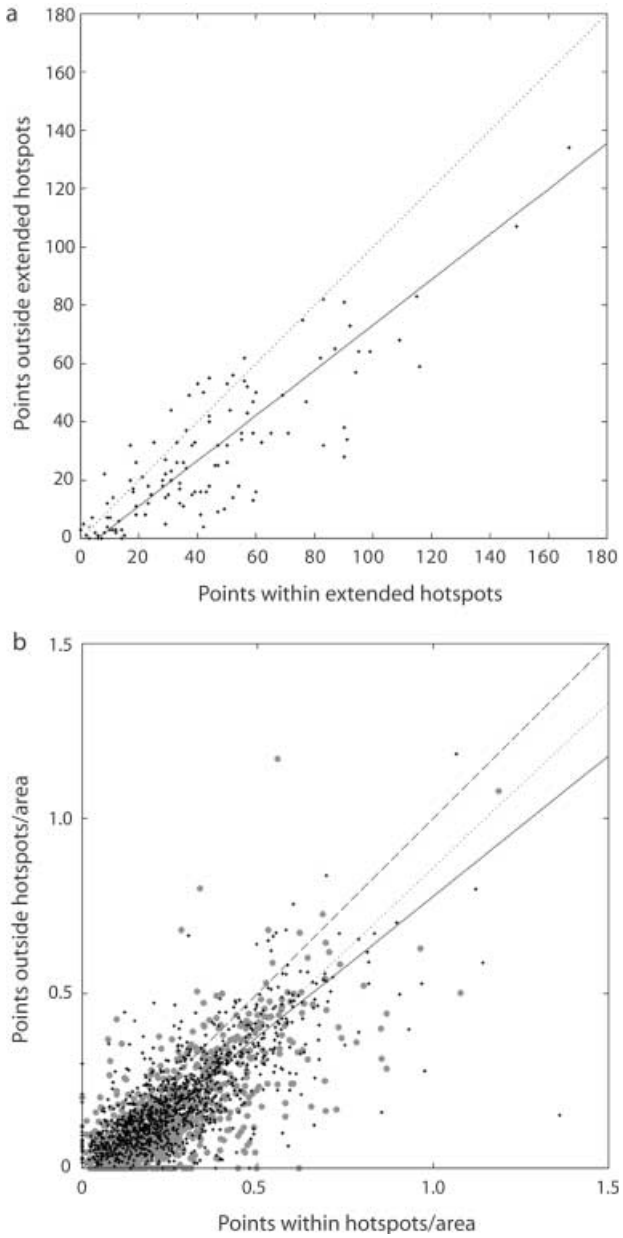


Figure 5 Comparisons of sampling intensity within and outside conservation priority areas, while controlling for area. (a) For each species with its range roughly equally distributed (ratio of $1 : 1 \pm 0.25$) within and outside extended priority areas, the number of sampling points within extended priority areas is shown on the *x*-axis and number of sampling points outside the extended priority areas on the *y*-axis (black points). The two grey points on the bottom left show the comparisons for the two species that are roughly equally distributed within and outside the priority areas (not extended). If sampling inside and outside of the extended priority areas were random or even, the comparison points would be scattered around the line of equivalence (dotted line; slope = 1). The 99% confidence interval of the standard major axis regression fitted to these points (solid line) does not include this line, indicating that these two lines are significantly different. (b) For every African passerine species ($n = 1068$), each black and each grey point represents the number of sampling localities inside and outside priority areas divided by the proportion of the species' range in each of these regions, an equal-area ratio of sampling localities. Black points are comparisons within and outside priority areas and grey points are comparisons within and outside extended priority areas. The 95% confidence intervals of the slope of standard major axis regressions fitted to these two data sets (priority area comparison shown as solid line, $r^2 = 0.58$; extended priority area comparison shown as dotted line, $r^2 = 0.45$) do not include the line of equivalence (dashed line).

Our data set does not include records from the last three decades of bird sampling in Africa. It is possible that the last 30 years of sampling included in the Blueprint analyses have corrected the sampling biases we found in this study. This would require a systematic sampling effort to target inaccessible and uninhabited suitable habitats such that the geographical coverage of primary data increases. This cannot be easily evaluated from the data used by the Blueprint because these analyses used a grid of species ranges instead of locality records (da Fonseca *et al.*, 2000; Balmford *et al.*, 2001; Brooks *et al.*, 2001).

Sampling error, such as that detected in the primary data obtained over more than 150 years is often carried into subsequent analyses for inferring ranges (Brown *et al.*, 1996). Hence, analyses that utilize species range data rather than point localities directly are still subject to similar biases. Our analysis using birds, an abundantly sampled group with large average ranges (Brooks *et al.*, 2001), showed a strong pattern of uncorrected sampling bias in the primary data. The implications for taxa with fewer records across smaller ranges, such as herpetofauna and invertebrates, are even more severe (Faith, 2002).

Critiques of conservation tools to designate global priority areas (e.g. Veech, 2000; Bates & Demos, 2001; Jepson & Canney, 2001) serve to improve conservation efforts overall. Here we evaluate the data available for conservation planning by analysing the shortcomings of the primary data on which more localized regional analyses are based. Too often such efforts are undertaken without consideration to the limitations of the data.

Testing for limitations or gaps in knowledge, such as sampling bias, can pinpoint geographical areas where further research is needed. The real danger of designating these conservation priorities is that often they are interpreted as the only sites that need to be conserved, while the importance of other sites is essentially ignored (Bates & Demos, 2001). No doubt many of the designated priority areas have truly high species richness and endemism. Yet unless the significant differences in sampling effort are accounted for, these patterns are suspect.

Statistical analyses for determining the extent of sampling bias are a necessary step if conservation prioritization is to become robust and reliable (Williams *et al.*, 2002). Methods to correct for sampling biases in diversity studies are only now being developed (Duckworth, 1997; Ponder *et al.*, 2001; Funk & Richardson, 2002; Williams *et al.*, 2002), and need to be applied and tested in prioritization exercises. These methods include rarefaction to homogenize sampling in original locality data, the use of range modelling for species included in conservation prioritization analyses, and the use of extrapolation and other richness estimators in poorly sampled areas (Colwell & Coddington, 1994; Duckworth, 1997; Peterson *et al.*, 2001; Ponder *et al.*, 2001; Funk & Richardson, 2002; Williams *et al.*, 2002).

Furthermore, more extensive surveys in inaccessible, relatively uninhabited areas may also lead to new discoveries and range extensions that could redefine some priority areas. Since priority areas are better sampled, they should be considered with caution in light of these sampling limitations and used to promote extensive surveys in poorly sampled areas (da Fonseca *et al.*, 2000). This study suggests that a major priority for conservation lies not just in the areas currently designated as priorities, but also in promoting scientific knowledge of lesser-sampled areas.

ACKNOWLEDGMENTS

This material is based upon work supported by the Division of Vertebrate Zoology and a NASA grant to the Center for Biodiversity and Conservation at the American Museum of Natural History, and the Center for Environmental Research and Conservation at Columbia University; we thank R. Anderson, F.K. Barker, P. Brito, J. Cracraft, A. Jarvis, P. Makovicky, F. Michelangeli, and M. Weksler for data, statistical advice, and technical support.

REFERENCES

Anderson, R.P., Peterson, A.P., & Gómez-Laverde, M. (2002) Using niche-based GIS modeling to test geographic predictions of competitive exclusion and competitive release in South American pocket mice. *Oikos*, **98**, 3–16.

Balmford, A., Moore, J.L., Brooks, T., Burgess, N., Hansen, L.A., Williams, P. & Rahbek, C. (2001) Conservation conflicts across Africa. *Science*, **291**, 2616–2619.

Bates, J.M. & Demos, T.C. (2001) Do we need to devalue Amazonia and other large tropical forests? *Diversity & Distributions*, **7**, 249–255.

Brooks, T., Balmford, A., Burgess, N., Fjeldså, J., Hansen, L.A., Moore, J., Rahbek, C. & Williams, P. (2001) Toward a blueprint for conservation in Africa. *Bioscience*, **51**, 613–724.

Brown, J.H., Stevens, G.C. & Kaufman, D.M. (1996) The geographic range: size, shape, boundaries, and internal structure. *Annual Review of Ecology and Systematics*, **27**, 597–623.

Busby, J.R. (1991) BIOCLIM – A bioclimate analysis and prediction system. *Nature conservation: cost effective biological surveys and data analysis* (ed. by C.R. Margules and M.P. Austin), pp. 64–68. Commonwealth Scientific and Industrial Research Organization, Melbourne, Australia.

Caithness, N. (1995) *Pattern, process and the evolution of the African antelope (Mammalia: Bovidae)*. PhD dissertation, University of Witwatersrand, Johannesburg.

Cincotta, R.P., Wisnewski, J. & Engelman, R. (2000) Human population in the biodiversity hotspots. *Nature*, **404**, 990–992.

Colwell, R.K. & Coddington, J.A. (1994) Estimating terrestrial biodiversity through extrapolation. *Philosophical Transactions of the Royal Society of London, Series B*, **345**, 101–118.

Duckworth, J.W. (1997) Correcting avian richness estimates for unequal sample effort in atlas studies. *Ibis*, **139**, 189–192.

ESRI (1999) *ArcView version 3.2*. Environmental Systems Research Institute, Inc., Redlands.

ESRI (2000) *Digital chart of the world*. Environmental Systems Research Institute, Inc., Redlands.

Faith (2002) Those complementarity analyses do not reveal extent of conservation conflict in Africa. Science dEbate http://www.sciencemag.org/cgi/eletters/293/5535/1591#381

da Fonseca, G.A.B., Balmford, A., Bibby, C., Boitani, L., Corsi, F., Brooks, T., Gascon, C., Olivieri, S., Mittermeier, R.A., Burgess, N., Dinerstein, E., Olson, D., Hannah, L., Lovett, J., Moyer, D., Rahbek, C., Stuart, S., Williams, P. (2000) …following Africa's lead in setting conservation priorities. *Nature*, **405**, 393–394.

Freitag, S., Hobson, C., Biggs, H.C. & Van Jaarsveld, A.S. (1998) Testing for potential survey bias: the effect of roads, urban areas and nature reserves on a southern African mammal data set. *Animal Conservation*, **1**, 119–127.

Fry, C.H., Keith, S., Urban, E.K. (2000) *The birds of Africa*, Vol. VI. Academic Press, London.

Funk, V.A. & Richardson, K.S. (2002) Systematic data in biodiversity studies: use it or lose it. *Systematic Biology*, **51**, 303–316.

Hall, B.P. & Moreau, R.E. (1970) *An atlas of speciation in African passerine birds*. Trustees of the British Museum (Natural History), London.

Hutchinson, M.F., Nix, H.A., McMahon, J.P. & Ord, K.D., (1995) *Topographic and Climatic database for Africa*, v.1 (CD-ROM). Centre for Resource and Environmental Studies, Australian National University, Canberra.

Jepson, P. & Canney, S. (2001) Biodiversity hotspots: hot for what? *Global Ecology & Biogeography*, **10**, 224–227.

Keith, S., Urban, E.K. & Fry, C.H. (1992) *The birds of Africa*, Vol. IV. Academic Press, London.

Kodric-Brown, A. & Brown, J.H. (1993) Incomplete data sets in community ecology and biogeography: a cautionary tale. *Ecological Monographs*, **3**, 736–742.

Lead, S. (2001) Generate randomly-distributed points. ArcView GIS script AS10955.zip. http://arcscripts.esri.com/

Myers, N., Mittermeier, R.A., Mittermeier, C.G., da Fonseca, G.A.B. & Kent, J. (2000) Biodiversity hotspots for conservation priorities. *Nature*, **403**, 853–858.

Nelson, B.W., Ferreira, C.A.C., da Silva, M.F. & Kawasaki, M.L. (1990) Endemism centres, refugia and botanical collection density in Brazilian Amazonia. *Nature*, **345**, 714–716.

Parnell, J.A.N., Simpson, D.A., Moat, J., Kirkup, D.W., Chantaranothai, P., Boyce, P.C., Bygrave, P., Dransfield, S., Jebb, M.H.P., Macklin, J., Meade, C., Middleton, D.J., Muasya, A.M., Prajaksood, A., Pendry, C.A., Pooma, R., Suddee, S. & Wilkin, P. (2003) Plant collecting spread and densities: their potential impact on biogeographical studies in Thailand. *Journal of Biogeography*, **30**, 193–209.

Peterson, A.T., Navarro-Sigüenza, A.G. & Benítez-Díaz, H. (1998) The need for continued scientific collecting: a geographic analysis of Mexican bird specimens. *Ibis*, **140**, 288–294.

Peterson, A.T., Egbert, S.L., Sánchez-Cordero, V. & Price, K.P. (2000) Geographic analysis of conservation priority: endemic birds and mammals in Veracruz, Mexico. *Biological Conservation*, **93**, 85–94.

Peterson, A.T., Sánchez-Cordero, V., Soberón, J., Bartley, J., Buddemeier, R.W. & Navarro-Sigüenza, A.G. (2001) Effects of global climate change on geographic distributions of Mexican Cracidae. *Ecological Modelling*, **144**, 21–30.

Ponder, W.F., Carter, G.A., Flemons, P. & Chapman, R.R. (2001) Evaluation of museum collection data for use in biodiversity assessment. *Conservation Biology*, **15**, 648–657.

Preston, F.W. (1948) The commonness, and rarity, of species. *Ecology*, **29**, 254–283.

Reid, W.V. (1998) Biodiversity hotspots. *Trends in Ecology and Evolution*, **13**, 275–280.

Scott, J.M. & Jennings, M.D. (1998) Large-area mapping of biodiversity. *Annals of the Missouri Botanical Garden*, **85**, 34–47.

Sokal, R.R. & Rohlf, F.J. (1995) *Biometry: the principles and practice of statistics in biological research*, 3rd edn. W.H. Freeman & Co., New York.

The Mathworks, Inc. (1996) *Matlab version 5*. The Mathworks Inc., Natick.

Urban, E.K., Fry, C.H. & Keith, S. (1997) *The birds of Africa*, Vol. V. Academic Press, London.

Veech, J.A. (2000) Choice of species-area function affects identification of hotspots. *Conservation Biology*, **14**, 140–147.

Williams, P.H. (1996) *Worldmap v. IV Windows: software and user document 4.1*. Natural History Museum, London.

Williams, P., Gibbons, D., Margules, C., Rebelo, A., Humphries, C. & Pressey, R. (1996) A comparison of richness hotspots, rarity hotspots and complementary areas for conserving diversity using British birds. *Conservation Biology*, **10**, 155–174.

Williams, P.H., Margules, C.R. & Hilbert, D.W. (2002) Data requirements and data sources for biodiversity priority area selection. *Journal of Bioscience*, **27(Suppl. 2)**, 327–338.

Winker, K. (1996) The crumbling infrastructure of biodiversity: the avian example. *Conservation Biology*, **10**, 703–707.

BIOSKETCHES

Sushma Reddy is a graduate fellow at the Division of Vertebrate Zoology–Ornithology of the American Museum of Natural History, and the Department of Ecology, Evolution, and Environmental Biology at Columbia University in New York. Her work focuses on the historical biogeography of southern Asia, molecular systematics, quantitative methods for predicting species ranges, and patterns of species distributions and diversity.

Liliana M. Dávalos is an international graduate fellow at the Division of Vertebrate Zoology–Mammalogy of the American Museum of Natural History, and the Department of Ecology, Evolution, and Environmental Biology at Columbia University in New York. Her interest in the bats, birds, forests, and people of the Neotropics spans the efficiency and robustness of biotic surveys, the historical biogeography of the West Indies, and the development of policy for biological conservation.