Reserve Selection in Regions with Poor Biological Data

KEVIN J. GASTON* AND ANA S. L. RODRIGUES

Biodiversity and Macroecology Group, Department of Animal & Plant Sciences, University of Sheffield, Sheffield S10 2TN, United Kingdom

Abstract: New approaches to the identification of priority areas for conservation are gaining popularity for their efficiency in maximizing species representation. However, their dependence on detailed distributional data severely hinders their application to regions where such information is limited, although these are commonly also the regions where conservation planning and action are most urgently required. We used exemplar data on the distribution of southern African birds to assess how sampling effort affects the performance of reserve networks selected by methods based on complementarity. We derived four scenarios of data availability from the initial data, resulting from different levels of sampling effort: abundance data, presence/absence data, low sampling effort, and absence of data. Reserve selection based on data obtained with low sampling effort can be highly effective in the representation of species, with a good relative performance also in terms of representation of species in peaks of abundance. This is because although the data on low sampling effort represent far fewer records than the original data, the records retained are biased toward the selection of peaks of abundance, even for the restricted-range species. Although the best results were naturally obtained from the most effort-intensive data set (with abundance data), these results suggest that methods based on complementarity are potentially valuable tools for reserve selection in regions for which biological data are poor.
Introduction

Most conservation planning involving the selection of reserves is based on one or only a few species (Simberloff 1988), often the most charismatic ones and/or those considered umbrella, indicator, flagship, or keystone species (Caro & O’Doherty 1999). However, reserve networks selected in this way may be ineffective for the conservation of other, nontarget, species (Pimm & Gittleman 1992; Andelman & Fagan 2000; Williams et al. 2000).

Complementarity-based methods for reserve selection have been proposed in response to the acknowledgement that the resources available for conservation are limited and should therefore be employed in ways that maximize the diversity of biological features, such as populations, species, and land systems, that are benefited (Pressey & Nicholls 1989; Pressey et al. 1993; Scott et al. 1993). These methods are becoming increasingly popular as tools for conservation planning (Pressey et al. 1993; Dobson et al. 1997; Ando et al. 1998; Howard et al. 1998; Margules & Pressey 2000), and the complementarity principle is now commonly applied in practical reserve-selection exercises, including the U.S. Gap Analysis Program (Scott et al. 1993; Kiester et al. 1996).

The efficiency of complementarity-based methods is achieved by identifying sets of sites that are highly complementary in their biological composition. To identify the sets, these methods rely on high-quality information on the spatial distribution of all species of concern. Regions with such data are often in developed countries, however, where reserve networks are already well established. The scarcity of distributional data restricts the possibility of applying these approaches to regions where they would be more relevant, such as poor tropical countries with high biodiversity levels and incipient reserve systems (Pimm 2000). Ideally, investment should be made in obtaining distributional data for these countries (Balmford & Gaston 1999), but many lack the time, personnel, and financial resources to do so (Ehrlich 1992), and increasing rates of habitat destruction make reserve selection and conservation action an urgent task.

Complementarity-based methods have also emphasized the representation of species in reserve networks, but this does not ensure the persistence of species over time (Margules et al. 1994; Virolainen et al. 1999; Rodrigues et al. 2000a, 2000b) because they may be represented in sites that are inadequate for their long-term viability (Gaston et al. 2001). The abundance pattern of most species over their geographic range is characterized by the existence of many sites of low abundance and just a few peaks where abundance can be orders of magnitude higher (Gaston 1994; Brown et al. 1995). The long-term effectiveness of reserve networks in retaining species can be improved, on average, by targeting these peaks of abundance for inclusion (Rodrigues et al. 2000a), but to locate them for each species of concern would require a still higher investment in sampling effort from those countries with poor biological data.

Restricted-range species, commonly considered priority targets for conservation investment because of their higher vulnerability to extinction (Stattersfield et al. 1998; Myers et al. 2000), present an even greater challenge. Because there is a widespread positive relationship between occupancy and local abundance, restricted-range species also tend to be locally rare (Gaston et al. 2000). Therefore, a substantial sampling effort is typically required even to detect these species, let alone to obtain data on their relative abundance across their range.

Ideally, complementary reserve selection should be based on the best possible data, obtained through sampling effort sufficient to detect the rarest species and to obtain information on the location of species’ peaks of abundance. It remains unclear, however, how the quality of the networks selected changes with the quality of the data used and, in particular, how well networks based on data obtained from low sampling effort perform.

Freitag and van Jaarsveld (1998) evaluated the sensitivity of selection procedures for conservation areas to survey intensity, survey extent, and taxonomic diversity by randomly deleting records, grid cells, and species from their data set on mammals in South Africa. In this study we are concerned about the effects of survey intensity (sampling effort) across all cells. Freitag and van Jaarsveld’s (1998) study has a number of weaknesses in this regard. First, the approach to data degradation by deleting records at random does not simulate a realistic situation of lower sampling effort across all cells, because it assumes that all records have the same probability of occurrence (for example, a record showing that a species has been seen in 90% of visits or in just 1% of them). As a result, a species’ range in the degraded data is made of a random selection of grid cells from the original range. In a situation of low sampling effort, the most natural outcome would be a reduced range such that the species tends to have been recorded in those sites where it is more abundant. Second, the performance of the reserves obtained using the degraded data was assessed mainly by measures of spatial congruence between sites selected. However, there is often a large number of possible solutions to each problem of complementary reserve selection (e.g., Arthur et al. 1997), and they often differ considerably in terms of the identity of the sites selected (Rodrigues & Gaston 2002). Third, viability considerations were not taken into account in the evaluation of the performance of reserve networks obtained from degraded data sets.

Here, we used exemplar data on the distribution and reporting rates of bird species in South Africa and Lesotho (Harrison et al. 1997) to assess how the effectiveness (Rodrigues et al. 1999) and efficiency (Pressey et al. 1998) of reserve selection with poor biological data.
& Nicholls 1989) of reserve networks obtained by comple-
mentary reserve selection are affected by the inten-
sity of sampling effort.

We are concerned about the variation in the sampling
effort across all candidate sites for reserve selection. We
did not address situations in which a fraction of sites
have been well studied and others have not been stud-
ied at all. For these, the solution probably involves some
form of data interpolation techniques, and the main is-
sure then becomes evaluation of the effectiveness of
such techniques.

Data and Methods

The Southern African Bird Atlas Project (Harrison et al.
1997) has provided the most comprehensive informa-
tion currently available on the distribution of birds in
southern Africa. Several previous researchers have used
this data set as a basis for planning studies (e.g., Allan et
al. 1997; Reyers et al. 2000; Fairbanks et al. 2001). Data
were collected mainly between 1987 and 1992 at a spa-
tial resolution of a quarter-degree grid for Lesotho,
Namibia, South Africa, Swaziland, and Zimbabwe and of
a half-degree grid for Botswana. Observers visiting each
cell recorded the presence of identified species on
checklists, and breeding and nonbreeding records were
considered equivalent. A variable number of checklists
has been submitted for each cell (μ = 39, σ = 88.0). A
total of 909 species was recorded. Based on these data,
reporting rates were calculated for each species in each
cell as the proportion of checklists submitted for that
cell on which the species was recorded. We used the
data for South Africa and Lesotho (1858 cells), excluding
marine, vagrant, marginal, and introduced or escaped
species from the analysis (total of 651 species).

For each species, peaks of reporting rates were de-
defined as cells with reporting rates of ≥80% of the max-
imum value observed for that species, and we assumed
that these peaks of reporting rates corresponded to
peaks of abundance. This assumption was based on the
positive association between abundance and reporting
rates found by Robertson et al. (1995), and it is expected
to be robust because it refers only to relative abundance
among species (thereby avoiding bias related to the spe-
cies’ different levels of conspicuousness). These peaks of
abundance correspond on average to 5.8% of the total
number of records for each species.

We considered the efficiency of a reserve network to
be inversely related to the percentage of total area that it
occupies (a first approximation to its relative cost; Pres-
ssey & Nicholls 1989). We evaluated effectiveness in spe-
cies representation using four different measures. First,
we determined the percentage of the overall number of
species represented (out of 651) in each reserve net-
work. This is the most traditional measure of effective-
ness (Rodrigues et al. 1999), but it may be misleading if
the species are represented only at sites that are inade-
quate to ensure their persistence. As a second measure
of effectiveness, we determined the percentage of spe-
cies represented in at least one of their respective peaks
of abundance. The third and fourth measures of effec-
tiveness refer to the representation of restricted-range
species, hereafter considered the 25% of species with
the smallest ranges in the study area (those present in
<8.8% of the 1858 grid cells; Gaston 1994). We deter-
mined the percentage of restricted-range species repre-
sented at least once and the percentage of these repre-
sented in at least one of their peaks of abundance.

Four scenarios of data availability resulting from differ-
ent levels of sampling effort were derived from the ini-
tial data set on the reporting rates of each of the 651 spe-
cies in each of the 1858 grid cells: abundance, presence/
absence, low sampling effort, and absence of data.

In the first scenario, the location of the peaks of abun-
dance for each species was known. Reserve networks of
variable total area that maximized the number of species
that could be represented in at least one of their peaks
of abundance were obtained.

In the second scenario, we used data on the presence/
absence of each species in each cell to obtain reserve
networks of variable total area that maximize the num-
ber of species represented at least once. This is the most
common scenario in the literature, with most reserve-
selection exercises being based on presence/absence
data, which can potentially be obtained with less sam-
pling effort than abundance data (Rodrigues et al. 2000b).
For both the abundance and presence/absence
scenarios, we assumed for the purpose of these analyses
that the data correspond to “the truth” of species distri-
bution and location of peaks of abundance.

In the third scenario, we used the original information
on reporting rates to simulate a situation in which very
low sampling effort—two visits—was applied across all
cells. A visit to a given cell was reproduced by randomly
re-sampling that cell such that the probability of each
species being observed during the visit was given by the
reporting rate for that species in the cell (i.e., each spe-
cies has an associated binomial distribution of parametes
β(1, p), where p is the reporting rate for that spe-
cies in the cell). Ten replicates of this scenario were
conducted. The presence/absence matrices retained, on
average, across the entire grid 95.9% of the initial 651
species, but the species richness sampled in each cell
was much reduced (42.6% of the original species rich-
ness, on average). This reduction in species richness
was the result of species with lower reporting rates in
each cell being missed when low sampling effort was
applied (Fig. 1). These matrices were therefore highly
biased toward retaining information about those species
that were locally more abundant. These low-sampling-
effort presence/absence matrices were then used to ob-
tain reserve networks of variable area that maximized the number of species represented.

In the fourth scenario, we assumed a total absence of information on the distribution of species in the study area. This was simulated by randomly selecting sets of cells of variable total area. The problem of finding the maximum number of species represented without exceeding a given area is a maximal-covering location problem (Church et al. 1996), represented as

\[
\text{maximize } \sum_{i=1}^{m} y_i
\]

and subject to

\[
\sum_{j=1}^{n} a_{ij} x_j \geq y_i, \quad i = 1, 2, \ldots, m
\]

\[
\sum_{j=1}^{n} x_j \leq k,
\]

\[
x_j \in \{0,1\} \quad j = 1, 2, \ldots, n
\]

\[
y_i \in \{0,1\} \quad i = 1, 2, \ldots, m,
\]

where \(m\) is the total number of species, \(n\) is the total number of sites, \(k\) is the maximum number of sites to be selected, \(y_i\) is 1 if species \(i\) is represented in at least one of the sites selected and 0 otherwise, \(x_j\) is 1 if site \(j\) has been selected and 0 otherwise, and \(a_{ij}\) is 1 if species \(i\) is present in site \(j\) and 0 otherwise (in the abundance-data scenario, being present refers to having a peak of abundance at site \(j\)). For each one of the first three scenarios, maximal-covering location problems were solved for values of \(k\) ranging between 1 and 80 (or until the minimum number of sites needed to represent all species was reached if \(\leq 80\)). All problems were solved optimally with CPLEX software (ILOG 1999).

Results

A significantly larger reserve network was required if the purpose was to obtain networks in which species are represented at their peak of abundance rather than simply anywhere within their ranges (Fig. 2). Indeed, we found that the area needed was more than six times larger if the purpose was to represent species at their peak of abundance (6.4% of the total area instead of 1.0%).

As expected, the reserve networks obtained in the absence-of-data scenario always performed worst in terms of their efficiency and effectiveness (Fig. 2). But whereas relatively high percentages of species could be represented by selecting sites at random (Fig. 2a), it was much more difficult to represent them at their peak of abundance (Fig. 2b), especially for the restricted-range species (Fig. 2c & 2d). For example, 79% of all species were represented in randomly selected networks occupying 1% of the total area (Fig. 2a), but only 13% of all species were represented at their peak of abundance (Fig. 2b). Regarding the restricted-range species, 34% of these were represented at least once in a network of the same size (Fig. 2c), but only 2% were represented at their peak of abundance (Fig. 2d).

The networks obtained in the scenario with presence/absence data were the most effective in representing species at least once within a given area (100% of all species represented within 1% of the total area [Fig. 2a & 2c]), but their performance declined dramatically if the purpose was to represent species at their peak of abundance (21.7% of all species [Fig. 2b] and 21.5% of restricted-range species [Fig. 2d] represented in 1% of the total area). The networks obtained in the abundance-data scenario (with knowledge of the location of these peaks) were not as effective in terms of simple species representation (97.7% of all species [Fig. 2a] and 77.3% of restricted-range species [Fig. 2c] represented within 1% of the area), but, as expected, they were the ones that performed better in representing species at their peak of abundance (70% of all species [Fig. 2b] and 39.3% of restricted-range species [Fig. 2d] represented in 1% of the area).
The most surprising results came from the performance of the networks obtained in the scenario of low sampling effort. They always performed much better than a random selection of sites, indicating that the results obtained cannot be explained by accidental species representation with the selection of increasingly large areas. In terms of species representation, either for all species (Fig. 2a) or for just the restricted-range species (Fig. 2c), the networks obtained in this scenario performed better than the networks obtained in the abundance-data scenario (95% and 85%, respectively, in 1% of the area). They were not as good in representing species at their peak of abundance (41% of all species [Fig. 2b] and 30% of restricted-range species [Fig. 2d] represented in 1% of the area), but in terms of these measures of effectiveness they performed much better than the networks obtained in the scenario with presence/absence data.

**Discussion**

The result that a significantly larger reserve network was required if the purpose was to obtain networks that improve the probability of maintaining species over time, (by representing them at the peak of their abundance), rather than simply those in which they are represented (Fig. 2), agrees with previous results (e.g., Araújo & Will-
of the units we considered (650 km²) may not be sufficient for the maintenance of viable populations (e.g., Mattson & Reid 1991; Nicholls et al. 1996; Manne et al. 1999). Nevertheless, we expect that the results for the different scenarios we explored will generalize to more demanding levels of representation of occurrence and of peak of abundance.

The performance of networks based on low sampling effort was never optimal in reaching the maximum values of species representation that could be obtained for a given area, either in terms of species representation (which, by definition, was obtained by the networks based on presence/absence data; Fig 2a, & 2c) or of representation of species at their peak of abundance (which, by definition, was obtained by the networks based on abundance data; Fig 2b & 2d). But it was generally good, well above the results of a random selection and sometimes close to optimal. This suggests that the data on the basis of which these networks were chosen retained useful information for identifying areas that are highly complementary in terms of species representation and in terms of their representation of peak of abundance.

The high effectiveness of these networks in terms of species representation (Fig. 2a & 2c) was likely a direct consequence of the effectiveness of the low sampling effort in recording at least some information on the distribution of most of the species in the original data set. Indeed, despite an average reduction of 39% in the total number of records, the large majority (on average 95.9%) of all species was recorded in at least one site by the low sampling effort. Their recorded mean range size was much reduced (on average, to 28.6% of the original), but complementary networks obtained from the low-sampling-effort data still represented these species because some information existed regarding their location in the study area. The decrease in the efficiency of these networks compared with the ones obtained with the original presence/absence data was a result expected from the increase in the species’ rarity in the data set (e.g., Lombard et al. 1995; Willis et al. 1996; Pressey et al. 1999; Rodrigues & Gaston 2001).

In terms of absolute numbers, the performance of the networks obtained with data resulting from low sampling effort was considerably worse when the purpose was to represent species at their peak of abundance (Fig. 2b & 2d) rather than to simply represent species (Fig. 2a & 2c). However, their performance is actually quite good compared with the optimal values that can be obtained (i.e., the values achieved by the networks based on abundance data) and the values obtained by random selection of the same percentage of total area. Despite it being a more reduced presence/absence data set than the original one, it performed better than the more complete original data (Fig. 2b & 2d).

In the latter, most records corresponded to sites where species exist at very low abundance (Fig. 3), having been detected only after considerable sampling effort. These sites are likely to be less adequate in maintaining species over time and are therefore relatively uninteresting with regard to the selection of priority areas for conservation. However, these records contributed much to improving the efficiency (i.e., to reduce the area needed) of complementary sets in representing all species. In an extreme situation, after enough sampling effort has been applied, the recorded range of most species will have been substantially expanded by the detection of vagrant individuals. The minimum reserve networks needed to represent all species will then be tiny (very efficient) but highly ineffective in ensuring species’ persistence over time (Rodrigues et al. 2000a & 2b).

In the low-sampling-effort scenario, the probability of a species being recorded in a given cell was directly related to its reporting rate in that cell. Although this means that locally rarer species are likely to be missed, it

![Figure 3. Distribution of the values of reporting rates, with the rank of each site (in descending order of reporting rates) in the original presence/absence scenario (shallower curve) and in the low-sampling-effort scenario (steeper curve). The values are for the Ovambo Sparrowhawk (Accipiter ovampensis, randomly selected) but are representative of the overall result. Although the number of records was substantially reduced in the low-sampling-effort data (from 624 to 120), this was mainly a result of the selective deletion of records with low reporting rates, and most of the peaks of abundance (13 out of 16; larger circles) were retained.](image-url)
also results in a bias toward the inclusion of records that correspond to peaks of abundance (Fig. 3). Indeed, the reduction in the number of data records in the low-sampling scenario (as referred to above, each species’ recorded range was on mean reduced to 28.6% of the original) was made chiefly by the deletion of low local-abundance records, whereas most (81.3%, on average) of the peaks of abundance were retained. Therefore, there was a higher probability that reserve networks selected with the low-sampling-effort data included those sites within species ranges where they were locally more abundant.

This result does not mean that data become less adequate for reserve selection when higher sampling effort is invested. Without high sampling effort, it is unlikely that the rarest species—the ones requiring higher conservation investment—would ever be detected. But it does demonstrate that, in complementary reserve selection, using too much indiscriminate data is not necessarily better than using fewer data. In most sampling schemes designed to obtain presence/absence data across a given region, it should be possible to extract additional information on the location of the peaks of abundance of most species if information on the sampling effort involved in obtaining each record is retained (as by Harrison et al. 1997).

All the species missing from the low-sampling-data matrices were restricted-range species. However, compared with the optimal values that can be obtained and with values obtained by random selection of the same percentage of total area, the relative performance of reserve networks based on these data was actually better in terms of representation of the restricted-range species at their peak of abundance (Fig. 2d) than in the representation of all species (Fig. 2b). This is because the bias in the data obtained by low sampling effort toward the selection of records that corresponded to peak of abundance was stronger for the restricted-range species. Indeed, although the mean percentage of these records in the low sampling matrices was 21.7% for restricted-range species, the corresponding value for the remaining species was only 8.4%. This means that when a complementary data set is selected on the basis of presence/absence data obtained by low sampling effort, there is a higher probability of capturing the peaks of abundance of the restricted-range species than of the more common species.

Naturally, the best results in terms of representing species with viability concerns were obtained by using the data set with more complete information about the location of the peak of abundance for each species. However, these results demonstrate that even distributional data obtained through low sampling effort may be valuable for the application of complementary approaches for the selection of priority areas for conservation. Although more empirical testing is needed, including that for data on other taxonomic groups with finer-scale distributions, our results suggest that these approaches can also be valuable tools in regions with poor biological data. They may be used as an initial coarse approach to the selection of a coherent network of reserves, based on data from low-intensity sampling schemes such as Conservation International’s Rapid Assessment Program (Conservation International 2001). This is good news at a time when the pressure on natural resources requires “quick and dirty” methods of evaluating natural ecosystems and designing reserves to protect them (Ehrlich 1992).

Acknowledgments

K.J.G. is a Royal Society University Research Fellow. A.S.R. is supported by a Portuguese FCT/PRAXIS studentship under the Sub-programme Science and Technology of the 2nd Community Support Framework. We are grateful to L. Underhill and the Avian Demography Unit of the University of Cape Town for providing access to the data and to the numerous volunteers that contributed to its collection. A. Brewer, O. Petchey, R. I. Vane-Wright, and T. Webb kindly commented on the manuscript.

Literature Cited


ILOG. 1999. CPLEX 6.5. ILOG, Gentilly, France.


