doi: 10.1111/j.1461-0248.2007.01025.x

LETTER

Biased data reduce efficiency and effectiveness of conservation reserve networks

Joanna Grand,[1]* Michael P. Cummings,[2] Tony G. Rebelo,[3] Taylor H. Ricketts[4] and Maile C. Neel[1]

[1]*Department of Plant Science and Landscape Architecture, University of Maryland, College Park, MD 20742, USA*

[2]*Center for Bioinformatics and Computational Biology, University of Maryland, College Park, MD 20742, USA*

[3]*South African National Biodiversity Institute, Private Bag X101, Pretoria 0001, South Africa*

[4]*Conservation Science Program, World Wildlife Fund – U.S., 1250 24th Street NW, Washington, DC 20037, USA*

**Correspondence: E-mail: jgrand@umd.edu*

Abstract

Complementarity-based reserve selection algorithms efficiently prioritize sites for biodiversity conservation, but they are data-intensive and most regions lack accurate distribution maps for the majority of species. We explored implications of basing conservation planning decisions on incomplete and biased data using occurrence records of the plant family Proteaceae in South Africa. Treating this high-quality database as 'complete', we introduced three realistic sampling biases characteristic of biodiversity databases: a detectability sampling bias and two forms of roads sampling bias. We then compared reserve networks constructed using complete, biased, and randomly sampled data. All forms of biased sampling performed worse than both the complete data set and equal-effort random sampling. Biased sampling failed to detect a median of 1–5% of species, and resulted in reserve networks that were 9–17% larger than those designed with complete data. Spatial congruence and the correlation of irreplaceability scores between reserve networks selected with biased and complete data were low. Thus, reserve networks based on biased data require more area to protect fewer species and identify different locations than those selected with randomly sampled or complete data.

Keywords

Bias, biodiversity conservation, complementarity, efficiency, MARXAN, rarity, reserve networks, reserve selection algorithms, species detection.

Ecology Letters (2007) 10: 364–374

INTRODUCTION

A major conservation objective is to maintain biodiversity by promoting long-term persistence of species in native ecosystems. Because habitat destruction and degradation are leading causes of biodiversity loss (e.g. Harrison *et al.* 1984; Wilcove *et al.* 1998), much attention focuses on establishing reserve networks to slow rates of habitat loss and fragmentation (Meir *et al.* 2004; Ricketts *et al.* 2005). Human demands for space and natural resources necessitate selecting reserves efficiently. Complementarity-based reserve selection algorithms allow design of reserve networks that achieve quantitatively defined objectives at minimum cost to other land uses (Possingham *et al.* 2000). Because these approaches ensure that areas selected for inclusion in a reserve network complement those already selected (Justus & Sarkar 2002), they represent all target features in the smallest number of sites (minimum set). They thus provide feasible options for establishing reserve networks in regions under intense pressure from competing land uses and are used extensively by conservation organizations.

Although reserve selection algorithms represent significant improvement over ad-hoc, opportunistic selection methods (Pressey *et al.* 1993; Pressey & Tully 1994), they are data-intensive, and most regions lack the economic resources necessary to generate accurate distribution maps for the majority of species. In addition to being incomplete, data sets available for conservation planning can be biased in various ways (Possingham *et al.* 2000). Examples of data bias include over-representation of charismatic or easily detectable species or sites in close proximity to field stations, protected areas, or roads (Possingham *et al.* 2000). Conservation planners typically must use available data to make decisions despite their potential limitations.

Data quality is likely to affect the outcome of conservation plans, but the nature and severity of potential effects are poorly understood. Understanding the biological and economic consequences of data limitations can both improve the choice of conservation priorities based on such data and guide efforts to collect additional data. Use of limited or biased data may decrease representation of taxonomic diversity, decrease reserve network efficiency due to reduced species congruence, decrease correlation with ideal reserve networks, and result in unreliable and ineffective determination of irreplaceability values (Freitag *et al.* 1996, 1998; Freitag & Van Jaarsveld 1998; Polasky *et al.* 2000; Gaston & Rodrigues 2003; Gladstone & Davis 2003). A more complete understanding of the independent and joint effects of data availability and bias on a variety of taxa in different regions may have significant implications for survey strategy and design (Freitag & Van Jaarsveld 1998) and the results of reserve selection algorithms (Possingham *et al.* 2000).

Effects of incomplete or biased data may impact the ultimate representation of some species within reserve networks more than others (e.g. rare vs. common species). Previous studies have attempted to quantify the impact of high levels of endemism or rarity on the results of reserve selection algorithms (Pressey *et al.* 1999; Virolainen *et al.* 1999; Rodrigues & Gaston 2001, 2002); however, these studies used ambiguous definitions of rarity. For example, rare species were defined as endemics with restricted ranges (Rodrigues & Gaston 2001), or simply based on the frequency of occurrence in the data set (Pressey *et al.* 1999). This lack of precision obscures the details of species biology that result in differential extinction vulnerability (Rabinowitz *et al.* 1986). Effective species protection is not possible unless we understand the variable effects of rarity type on conservation potential and extinction risk. Here we used the rarity definition of Rabinowitz *et al.* (1986) which is based on three characteristics: geographic distribution, habitat specificity and local population size.

The purpose of this research was to examine the sensitivity of the outcome of a complementarity-based reserve selection algorithm to variation in sampling effort and bias. We used the Proteaceae (Angiospermae: Rosidae) of the Cape Floristic Region of South Africa as the model system for this study, as it is to our knowledge one of the most complete species distribution data sets available at a point locality resolution (Lombard *et al.* 2003). The Cape Floristic Region is one of the world's richest biodiversity hotspots, and the Proteaceae is the best known vascular plant family in the region (Rebelo & Siegfried 1992). We assumed this data set represented 'complete' knowledge of Proteaceae distributions, and degraded it by subsampling it randomly and in ways that reflect realistic biases typically found in species distribution data. We then examined the numbers and rarity classes of species detected in these subsamples, and compared the reserve networks generated from the biased subsamples to those generated from the complete Proteaceae database and the randomly subsampled data, in terms of efficiency, spatial congruence and irreplaceability. Exceptional on-the-ground conservation planning has already been conducted in the Cape Floristic Region (Rebelo & Siegfried 1992; Cowling *et al.* 2003a,b; Rouget *et al.* 2005) using these and other data. This research is not intended to re-examine those efforts, but to explore the broader question of the impact of biased data on conservation planning in general.

METHODS

Data collection

The Protea Atlas Project database contains a total of 673 taxa, most of which are endemic to the Cape Floristic Region (Forshaw 1998). The data were collected by 478 volunteers over 10 years and include over 2.5×10^5 species occurrence records. The database is considered to be a nearly complete inventory of Proteaceae in the Region. Atlassers recorded the abundance of all taxa observed within *c.* 20 ha surrounding sample point locations which were mapped by longitude and latitude.

Using ArcGIS 9.0 (Environmental Systems Research Institute 1999–2004) geoprocessing tools, we divided the study area into 2 km^2 planning units, which we considered to be an appropriate size for regional conservation planning as it is sufficiently large to be biologically reasonable for conservation and not so large that the data become imprecise. At the 2 km scale, many planning units did not contain any survey points. To simulate a truly complete species distribution data set, we excluded unsurveyed planning units from analysis and 9016 planning units were retained as potential reserves.

We filtered the database to include only the highest priority taxa for conservation planning. We removed 11 taxa that were either alien, extinct, or of questionable taxonomic status; 227 putative hybrids; 16 taxa that were either planted or possibly planted; and 44 taxa that did not occur within the boundaries of the Cape Floristic Region as defined by Goldblatt & Manning (2000) and CAPE (Cowling *et al.* 2003a,b). The final data set contained 375 species and subspecies.

Rarity classification

Rabinowitz *et al.* (1986) classifications were conducted by A.G. Rebelo based on expert opinion. Taxa were classified as having either widespread (W) or localized (L) geographic distributions, broad (B) or restricted (R) habitat specificity, and dense (D) or sparse (S) populations. The majority of taxa were classified as either localized with restricted habitat specificity and sparse populations (LRS), widespread with restricted habitat specificity and dense populations (WRD), or widespread with broad habitat specificity and dense populations (WBD = common). No species were classified as localized with broad habitat specificity and sparse populations (LBS).

Subsampling

We implemented three forms of biased subsampling; record-based roads-biased (RB1), site-based roads-biased (RB2) and abundance-biased (AB). Both roads-biased subsampling schemes represented spatial biases in which sites far from roads were poorly sampled, and abundance-biased subsampling represented a non-spatial, detectability bias in which sparse populations were poorly sampled. We generated 3000 biased subsamples; 1000 replicates of each subsampling type. Because we were interested in simulating realistically biased sampling schemes, we did not attempt to control for differences in sampling effort among bias types (although RB1 and RB2 were similar in intensity). Therefore, apparent differences between roads-biased and abundance-biased sampling schemes must be interpreted with caution due to the confounding of bias and sampling effort.

Roads-biased subsamples

The data used for roads-biased subsampling contained eight classes: arterials, freeways, main roads, national routes, roads under construction, secondary roads (mainly gravel), other roads (gravel or 4 × 4 tracks and farm roads), and track footpaths. We eliminated all road classes that we considered unlikely to be used for sampling due to high traffic volume, retaining only secondary roads, other roads and track footpaths. We calculated the distance from each record to the nearest minor road and introduced bias in two ways. Record-based bias simulated a situation in which sites were visited less often as distance from the nearest road increased, and some species were undetected as they were unidentifiable or difficult to see in one or a few visits. For the site-based bias, distant sites were less likely to be visited at all, but if they were visited, every species was observed.

For both types of roads bias, we developed algorithms to subsample with exponentially decreasing probability as distance from the nearest minor road increased, hence eliminating the need to select a single arbitrary cut-off distance. Algorithm parameter values produced a simulated roads bias in which *c.* 30% of all records were retained, *c.* 90% of subsampled records were within 2 km of the nearest minor road, and sampling declined rapidly beyond 2 km. In comparison, 52% of the records in the complete data set were within 2 km of the nearest minor road and the number of records declined gradually as distance increased.

Abundance-biased Subsamples

We used absolute species abundance at a site as an index of detectability. Abundance in the database was recorded in 13 classes: historical records only, 1, 2, 3, 4, 5, 6, 7, 8, 9 plants, 10–100 plants, 100–10 000 plants and > 10 000 plants. We replaced missing data values for 1836 records with the mean of all remaining abundance records for that species.

We developed an algorithm to probabilistically subsample the data set based on species abundance. Rather than imposing an arbitrary cut-off abundance, sampling declined continuously as abundance decreased beyond an abundance threshold value of 100. We chose this threshold because we believed it was realistic to miss a population of ≤ 100 individuals in a 20 ha area, and the majority of the records had population values in the 10–100 and 100–10 000 classes. Therefore, choosing a threshold ≤ 10 or > 10 000 either retained or excluded *c.* 90% of the records in the complete data set. A threshold value of 100, in which all records with abundances > 100 were always included and records with abundances of ≤ 100 were sampled probabilistically, retained *c.* 60% of all records.

Random subsamples

We generated a random subsample of equal size to each replicate of the biased subsamples to separate the confounded effects of bias and sampling effort within each bias type. This yielded a total of 3000 random subsamples to match the 1000 replicates of each biased subsampling type. For the record-based subsampling schemes (AB and RB1) sample size was defined as number of records in the data set, thus random subsamples were matched by number of records. For the site-based subsampling scheme (RB2) sample size was defined as the number of sites in the data set, thus random samples were matched by number of sites. Despite the difference in subsampling units between RB1 and RB2, the number of records in each was similar. Hereafter, we refer to these random subsamples as record-based roads random (RR1), site-based roads random (RR2) and abundance random (AR).

Reserve selection

Software parameters

We used the decision support tool MARXAN (Ball & Possingham 2000; Possingham *et al.* 2000) to generate the near minimum set of planning units required to meet a given conservation goal. Because we used a heuristic algorithm (described below) we were not guaranteed to find the absolute minimum set of sites. However, for convenience we use the term 'minimum set' throughout this manuscript to refer to the set of reserves selected by MARXAN. Our goal was not to design a reserve network suitable for on-the-ground conservation, but to examine the effects of data quality on hypothetical reserve networks. Although not necessarily appropriate for real-world conservation planning, we chose the simplest scenario to obtain baseline reserve systems that represented all species at minimal cost (Possingham *et al.* 1993). We did not consider spatial parameters such as boundary length, aggregation or separation of reserves. We also excluded the cost threshold, which constrains the reserve system to a user-specified maximum cost, and set the representation target for each species (number of required occurrences of a species in the reserve system) and the cost of each planning unit to one. We set the species penalty factor, the penalty assigned to a reserve system for failing to represent a species, to 10 to guarantee that every detected species would be represented (Ball & Possingham 2000). MARXAN allows for the inclusion of abundance data; however, for simplicity we used presence–absence only. Although many alternative minimum sets may be generated, for most analyses we recorded only the solution selected by MARXAN.

Heuristic algorithm

For each of the 6001 data sets (1000 replicates of each biased and random subsample type, plus the complete data set), we implemented 1000 MARXAN runs with the summed rarity algorithm and normal iterative improvement (Ball & Possingham 2000). We chose summed rarity because initial comparisons showed it consistently generated more efficient solutions than either adaptive simulated annealing or any of the other heuristic algorithms. Although studies have shown that simulated annealing may be better at decreasing the cost or improving the clustering in a reserve design (Pressey *et al.* 1997; McDonnell *et al.* 2002), summed rarity produced more efficient reserve networks with our data and we were not concerned here with optimizing the spatial configuration of reserves or simultaneously optimizing an array of competing objectives. Furthermore, there is evidence that the simpler algorithms (greedy and rarity based) may be more robust to variations in survey effort and spatial bias in sampling when applied to binary data (Possingham *et al.* 2000). We therefore expected our results to provide a conservative estimate of the sensitivity of outcomes from complementarity-based algorithms to changes in data quality and quantity.

Computation

To complete the large number of analyses required for this study (1.2002×10^7 including the efficiency cost runs described below) we used Grid computing (Cummings & Huskamp 2005) through The Lattice Project (Bazinet & Cummings 2007). The MARXAN executable was converted to a Grid service (Bazinet *et al.* 2007) to distribute files among hundreds of computers where analyses were conducted asynchronously in parallel. Grid computing allowed us to complete > 1000 CPU days of computation in a few weeks.

Measured variables

Species detection in subsamples

Sampling (either random or biased) may result in some species being undetected. Species detection is critical to the outcome of reserve selection algorithms because only species known to be present in a region can be explicitly targeted, and therefore guaranteed representation in reserve networks. To quantify the efficacy of limited and biased sampling, we calculated the number of species detected and the percentage of species undetected. To evaluate the impact of species distribution characteristics on detection levels, we also calculated the percentage of species in each rarity class.

Chance representation in reserve networks

Reserve networks can capture undetected species by chance if such species are included in reserve networks without being explicitly targeted. To determine whether chance representation might mitigate the impact of low sampling effort or bias on species representation, we used the complete distribution data to calculate the total number of species actually represented in the minimum sets generated from subsampled data. Species that were not detected in the subsamples but occurred within at least one planning unit contained in the minimum set were considered represented by chance.

Minimum-set size

To assess the impact of incomplete and biased data on the efficiency of reserve networks, we calculated the number of planning units contained in the minimum sets generated with subsampled and complete data. We also calculated the percentage of the complete minimum-set size included in the minimum sets derived from subsampled data.

Efficiency cost

We created an index of 'efficiency cost' to quantify the additional cost of complete representation in a situation where a reserve network was initially selected with incomplete or biased data and subsequently expanded when more data became available. We defined efficiency cost as the difference in minimum-set size between this expanded reserve network, and a reserve network initially designed with complete data. Thus, efficiency cost depends on two components: (1) the difference between the initial minimum-set size generated from subsampled data and the initial minimum-set size generated from complete data; and (2) the number of additional sites required for complete representation. To calculate efficiency cost we fixed the minimum set generated from each replicate subsample in a reserve network, and then generated 1000 expanded reserve networks using complete data. We then subtracted the number of planning units in the minimum set initially selected with complete data from each expanded minimum set.

Minimum-set similarity

We quantified overlap in the minimum sets generated from the subsampled and complete data sets using the Jaccard (1912) coefficient following Warman *et al.* (2004). For comparison, we also quantified similarity among all alternative minimum-set solutions generated from the complete data set.

Irreplaceability comparisons

Irreplaceability was defined as the number of times a planning unit was selected for inclusion in all reserve network solutions. To compare the importance of each planning unit in the subsampled and complete data set solutions, we calculated the number of irreplaceable planning units (i.e. selected in all 1000 MARXAN runs) in the complete data set, and the mean number of irreplaceable planning units for all replicates of each subsample type. We also calculated the correlation between all planning unit irreplaceability scores from the complete and subsampled data and examined the average of the 1000 correlations. We used rank correlation (Spearman 1904) to examine the relationship among irreplaceability scores following Warman *et al.* (2004).

Hypothesis testing

We used permutation tests (Manly 1991; Good 1994; Maritz 1995) with 10 000 permutations to assess the significance of the difference between matching biased and random subsample distributions. We conducted each permutation by randomly assigning each of the 1000 matched pairs to the biased or random group, summing the differences across all pairs, and locating the position of the actual value with respect to the distribution of the 10 000 permutations. To determine the significance of the difference between subsampled distributions and the complete data set, we located the position of the complete data set value with respect to the distribution of the 1000 replicate subsamples.

RESULTS

Species detection in subsamples

Differences in detection levels between all subsampled data sets and the complete data set were statistically significant ($P = 0.001$–0.01) with the exception of abundance random ($P = 0.291$) (Fig. 1). The median number of species missing from the biased data sets was 4–21 (1–5%) and from the random subsets was 1–3 (0.3–0.8%). Although all biased subsamples detected fewer species than their random counterparts ($P = 0.0001$), differences were much greater in both types of roads-biased subsamples than in abundance-biased subsamples. In comparison with the roads-biased subsamples, surprisingly few species (median = 0.3–1%) were missing from all other subsampling schemes despite elimination of 40–70% of occurrence records.

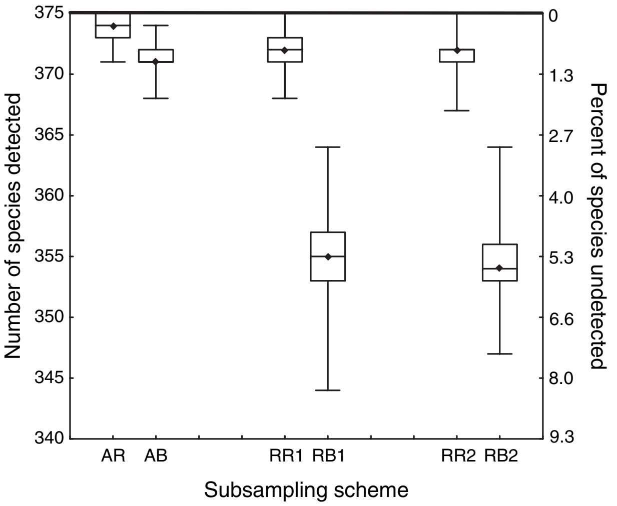

Figure 1 Number of species detected (and per cent of species undetected) by each subsampling scheme of Proteaceae in the Cape Floristic Region. Data are from 1000 replicates of each subsampling type (AR, abundance random; AB, abundance biased; RR1, record-based roads random; RB1, record-based roads biased; RR2, site-based roads random; RB2, site-based roads biased). The dark horizontal line indicates the total number of species in the complete data set. All pairs of biased and matching random subsample distributions are significantly different ($P = 0.0001$). Boxplots illustrate median, interquartile range, minimum and maximum.

Species detection as a function of rarity class

Detection levels varied with species rarity class. The abundance-random subsamples detected a median of 100% of species in all rarity classes and abundance-biased subsamples detected 100% of species in all rarity classes except WRD and LRS (Fig. 2a–b). Both forms of roads-biased and roads-random subsamples detected a median of < 100% of the species in rarity classes WRD, LRD and LRS, although detection in these classes was much lower in biased than random subsamples (Fig. 2c–f). Locally distributed (L), habitat restricted (R) species with sparse populations (S) were the most severely impacted in all subsampling schemes.

Chance species representation in reserve networks

On average, no species were represented by chance in abundance-random and abundance-biased subsamples, one additional species was represented by chance in both roads-random subsamples, and two additional species were

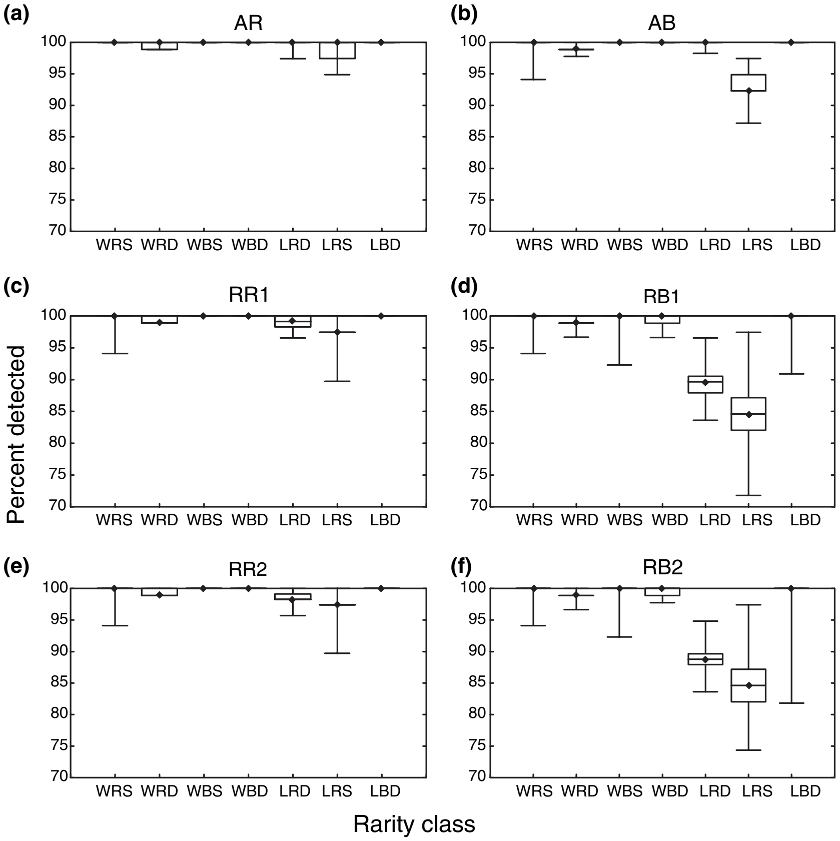

Figure 2 Percentage of species in each Rabinowitz rarity class detected by each subsampling scheme of Proteaceae in the Cape Floristic Region. Data are from 1000 replicates of each subsampling type (AR, abundance random; AB, abundance biased; RR1, record-based roads random; RB1, record-based roads biased; RR2, site-based roads random; RB2, site-based roads biased; W, widespread distribution; L, localized distribution; B, broad habitat specificity; R, restricted habitat specificity; D, dense populations; S, sparse populations). Boxplots illustrate median, interquartile range, minimum and maximum.

represented by chance in both roads-biased subsamples. No subsampling scheme on average achieved complete representation even after accounting for chance-represented species.

Minimum-set size

All biased subsamples resulted in significantly larger minimum sets (despite targeting fewer species for representation) than their random counterparts ($P = 0.0001$) (Fig. 3). This difference was slightly greater in abundance-biased subsamples than in both types of roads-biased subsamples. All subsampling schemes also resulted in significantly larger minimum sets than the complete data set (complete data minimum-set size = 96, $P = 0.001$–0.008). The biased subsampling schemes generated minimum sets that were a median of 9–17% larger than the minimum set generated with complete data. Record-based roads bias resulted in larger minimum sets than site-based roads bias.

Efficiency cost

All biased subsampling schemes had significantly higher efficiency costs than their random counterparts ($P = 0.0001$), although the difference was much greater in both types of roads-biased subsamples than in abundance-biased subsamples (Fig. 4) mainly as a consequence of relatively low species detection levels (component 2). Record-based roads bias resulted in slightly higher efficiency cost than site-based roads bias due to the larger initial minimum-set size (component 1).

Minimum-set similarity

Overall, spatial congruence as measured by the median Jaccard similarity of planning-unit locations in the minimum sets generated with subsampled and complete data was low, ranging from 0.11 to 0.25 (Fig. 5). All biased subsampling schemes generated minimum sets that were significantly less similar to the minimum set generated from complete data than their random counterparts ($P = 0.0001$). For comparison, although the similarity among all alternative minimum sets generated with complete data was low ($n = 32$, median = 0.43), it was considerably higher than between minimum sets generated from subsampled and complete data.

Irreplaceability comparisons

Reserve networks selected with complete data contained 46 irreplaceable planning units (Table 1). All subsampling

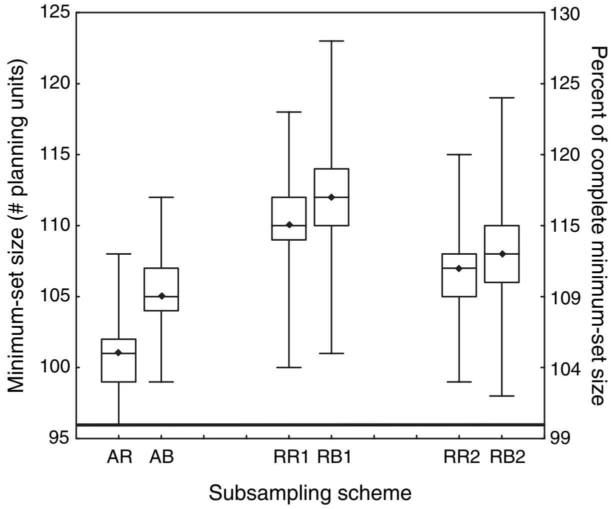

Figure 3 Minimum-set sizes (and per cent of complete minimum-set size) generated from each subsampling scheme of Proteaceae in the Cape Floristic Region. Data are from 1000 replicates of each subsampling type (AR, abundance random; AB, abundance biased; RR1, record-based roads random; RB1, record-based roads biased; RR2, site-based roads random; RB2, site-based roads biased). The dark horizontal line at 96 2 km^2 planning units indicates the minimum-set size generated from the complete data set. All pairs of biased and matching random subsample distributions are significantly different ($P = 0.0001$). Boxplots illustrate median, interquartile range, minimum and maximum.

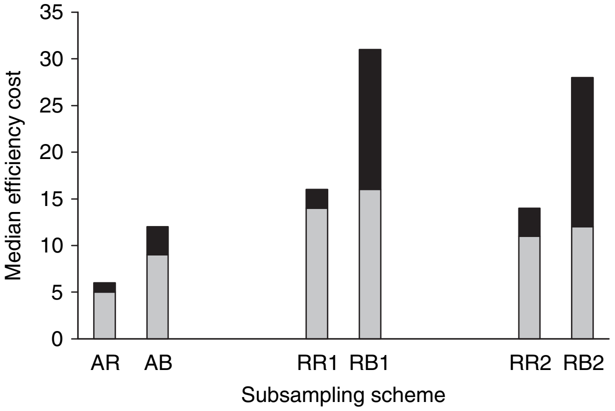

Figure 4 Median efficiency cost of each subsampling scheme of Proteaceae in the Cape Floristic Region. Efficiency cost represents the difference in minimum-set size between reserve networks selected initially with complete data, and reserve networks selected initially with biased or incomplete data and then subsequently expanded (using the complete data set) until all species were represented. Data are from 1000 replicates of each subsampling type (AR, abundance random; AB, abundance biased; RR1, record-based roads random; RB1, record-based roads biased; RR2, site-based roads random; RB2, site-based roads biased). The grey portion (component 1) represents the median difference between the initial minimum-set size generated from subsampled data and the minimum-set size generated with complete data. The black portion (component 2) represents the median number of additional planning units required for complete representation of complete data (i.e. all 375 taxa). All pairs of biased and matching random subsample distributions are significantly different ($P = 0.0001$).

schemes generated more irreplaceable planning units than the complete data set ($\bar{x} = 52–63$) and all biased subsampling schemes generated more irreplaceable planning units than their random counterparts. The average Spearman's rank correlations between the irreplaceability scores generated from all replicates of each subsample type and those generated from complete data were low ($\bar{x} = 0.30–0.57$). All biased subsampling schemes had lower correlations with the complete data set than did their random counterparts.

DISCUSSION

Conservationists are often required to make decisions with incomplete and biased data. Our results demonstrate that incomplete data, with or without sampling bias, have major implications for reserve design. However, biased sampling clearly impacts the results of reserve selection algorithms more severely than decreased sampling effort alone. This pattern applies to both spatially biased and detectability biased data. Our results suggest that reserve networks based on biased data require more area to protect fewer species and identify different locations than those selected with unbiased or complete data.

Biased sampling can lead to inadequate representation in reserve networks due to omission of undetected species from target species lists. The lack of detection of localized (L), habitat-restricted (R) species with sparse populations (S) was not unexpected for abundance-biased subsampling because sparse populations were intentionally poorly sampled; however, it is notable that the impact of both types of roads-biased subsampling on LRS species was even more severe. This increased severity was clearly not a result of lower sampling effort alone because random subsampling had a much weaker impact on LRS species (Fig. 2c–f). It is the spatial character of roads bias that severely limits the type of species whose distributions can be adequately characterized because species that are narrowly distributed, restricted to habitat types that do not exist in close proximity to roads, or are unable to thrive in disturbed environments, are rarely sampled. Because localized and habitat-restricted species tend to be spatially clustered, spatially-biased sampling could easily fail to detect all occurrences of a species. It is precisely the localized and habitat-restricted species that should be the focus of sampling and conservation efforts.

Furthermore, our results indicate that chance representation of species that were omitted from target species lists

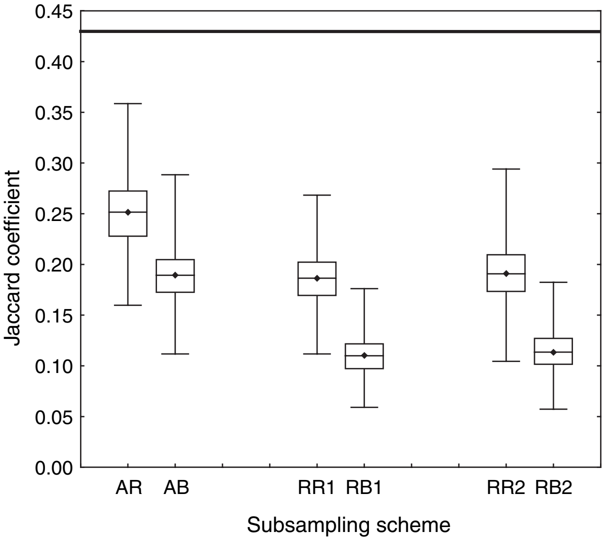

Figure 5 Jaccard coefficients for each subsampling scheme of Proteaceae in the Cape Floristic Region. The Jaccard coefficient measures the spatial congruence (overlap) of planning units between the minimum sets generated with subsampled and complete data. Data are from 1000 replicates of each subsampling type (AR, abundance random; AB, abundance biased; RR1, record-based roads random; RB1, record-based roads biased; RR2, site-based roads random; RB2, site-based roads biased). The dark horizontal line at a Jaccard coefficient of 0.43 indicates the median overlap of planning units among all alternative minimum sets generated with complete data. All pairs of biased and matching random subsample distributions are significantly different ($P = 0.0001$). Boxplots illustrate median, interquartile range, minimum and maximum.

was surprisingly low. At least for the Proteaceae, chance representation may do little to mitigate the failure to explicitly represent the biodiversity of the region in reserves. Because lack of species detection was substantially more severe with spatially-biased sampling than with random sampling at 30% sampling effort, our results highlight the importance of spatially unbiased surveys regardless of level of sampling effort. Although detectability-biased sampling appears to be less of a problem with regard to species detection, this phenomenon must be explored further to eliminate the possibility that it is due to a less extreme reduction in sampling effort.

The effect of biased data on minimum-set size may be the most consequential of all results presented here. Paradoxically, our results suggest that more planning units are required to represent fewer species when biased data are used for decision making. The apparent reason for this result is the decrease in compositional overlap among our 2 km^2 planning units due to the lack of species detection at poorly sampled or unsampled sites within them. An inadequately sampled planning unit that is included in a minimum set will therefore represent fewer species, requiring more planning units to be included before full representation is achieved.

Table 1 Number of irreplaceable planning units (i.e. selected in all 1000 MARXAN runs) for the complete data set and mean number of irreplaceable planning units of the 1000 replicates of each subsample type, and mean Spearman's rank correlations between the irreplaceability scores of all planning units generated from the 1000 replicates of each subsample type and those generated from the complete data set of the Proteaceae in the Cape Floristic Region.

Sampling scheme	Number (or mean number) of irreplaceable planning units	Mean correlation with complete data irreplaceability scores
Complete	46	N/A
Abundance random (AR)	52	0.57
Abundance biased (AB)	57	0.47
Record-based roads random (RR1)	61	0.36
Record-based roads biased (RB1)	63	0.30
Site-based roads random (RR2)	57	0.40
Site-based roads biased (RB2)	60	0.31

The relatively high efficiency cost of designing reserves with both forms of roads-biased data provides further evidence that spatially biased sampling protocols are ineffective at generating distribution data that can be used to design effective and efficient reserve networks. This cost appears to be exacerbated when sites far from roads are incompletely sampled (RB1), as opposed to being sampled either completely or not at all (RB2). However, the proportion of habitats into which the road network extends is likely to have a large impact on the quality of roads-biased data sets. Because sampling intensities differed between the roads-biased and abundance-biased sampling, quantifying the relative efficiency costs of these two types of bias requires further investigation.

The low similarity or spatial congruence among planning units in the minimum sets generated from subsampled and complete data suggests that planning-unit location is highly sensitive to quality and quantity of species occurrence information. Use of incomplete data of any kind, and biased data in particular, as input for reserve selection algorithms has the potential to generate a very different set of priority sites than if complete data are used. Although the observed level of similarity among alternative minimum sets generated from the complete data set indicates somewhat high flexibility in minimum-set solutions, it also suggests that the lack of similarity between minimum sets generated with subsampled and complete data was not simply an artifact of generally high flexibility in potential reserve designs. Thus, designing reserve networks with limited data leads to the selection of different (and more) locations to achieve the same objectives. It is important to recognize, however, that we compared only one possible minimum set generated from each subsample to one possible minimum set generated from the complete data set. The analysis with complete data, for example, generated 32 alternative minimum sets that might have produced slightly different spatial patterns. Nevertheless, our conclusions are supported by the fact that spatial congruence was considerably greater among all alternative minimum sets generated from complete data than between minimum sets generated from subsampled and complete data (Fig. 5).

Finally, the low correlations between irreplaceability scores for reserve network solutions generated from subsampled and complete data suggest that the importance of planning units is also highly sensitive to quality and quantity of species occurrence information. All forms of subsampling, but especially biased forms, increased the number of irreplaceable planning units. This may be due to the failure to detect all but a single or small number of occurrences of a species, thereby increasing species rarity as defined by the summed rarity algorithm. Irreplaceability, defined here as selection frequency of a planning unit, provides a fundamental measure of the conservation value of a planning unit (Pressey *et al.* 1993) and is therefore a critical component of efficiency (Stewart *et al.* 2003; Stewart & Possingham 2005). The failure of reserve selection algorithms to accurately identify irreplaceable planning units when incomplete data are used, contributes to the efficiency cost of the resulting reserve networks.

Several limitations in our analyses may temper our findings. For example, our results may be sensitive to the choice of reserve selection algorithm. Because each algorithm uses different criteria to assign a score to each planning unit, it is possible that other heuristic or optimization algorithms might produce different results. However, we believe that our findings are relatively conservative as there is evidence that the rarity algorithms are robust to variation in survey effort and spatial bias (Possingham *et al.* 2000). Nevertheless, further research on the impact of data quality and quantity on the results of different complementarity-based reserve selection algorithms is necessary. Other unexamined variables which we intend to explore (e.g. variations in sampling effort, the number of required occurrences of each species, planning-unit resolution, and spatial configuration of reserve networks) would also contribute to more fully understanding the consequences of using incomplete and biased data to inform critical conservation planning decisions.

In summary, our findings have important implications for design of conservation reserve networks. Biodiversity data will never be complete, because exhaustive surveys over broad regions are rarely feasible and take longer than conservation planners can afford to wait. Unfortunately, incomplete data impacts conservation planning in critical ways. Given that impacts are significantly more severe with biased than with randomly sampled data, survey methods should be unbiased. Encouragingly though, our results demonstrate that exhaustive surveys are not essential to generate reserve networks that closely approximate the levels of representation and efficiency achieved with complete data. Random sampling with only 60% of the effort of an exhaustive survey (subsampling scheme AR) can detect 99.7% of species (Fig. 1), generate minimum sets that are only 5% less efficient than ideal reserve networks (Fig. 3), and have a relatively low efficiency cost (*c.* 6% increase in cost) (Fig. 4). Given this performance additional data collection may not be warranted. Limited sampling effort of this level however, even if unbiased, does lead to low spatial and irreplaceability value congruence with ideal reserve networks, indicating that different planning units would be chosen for conservation. Conservation practitioners must be wary of biases in their data sets and weigh the costs and benefits of allocating limited time and resources toward additional data collection.

ACKNOWLEDGEMENTS

We greatly appreciate the technical assistance provided by Adam Bazinet, lead developer of The Lattice Project at the University of Maryland, the hard work of all atlassers who contributed to the Protea Atlas Project, and the helpful advice of the WWF Conservation Science Program staff. We thank Hugh Possingham and one anonymous referee for helpful comments on an earlier version of this manuscript. JG was supported by a National Science Foundation (NSF) Postdoctoral Fellowship in Biological Informatics and MCP was supported in part by grants from the Department of Energy and NSF.

REFERENCES

Ball, I.R. & Possingham, H.P. (2000). *MARXAN (V1.8.2): Marine Reserve Design Using Spatially Explicit Annealing, A Manual.* Available at: URL http://www.ecology.uq.edu.au.

Bazinet, A.L. & Cummings, M.P. (2007). The Lattice Project: a Grid research and production environment combining multiple Grid computing models. In: *Distributed & Grid Computing – Science Made Transparent for Everyone. Principles, Applications and Supporting Communities* (ed. Weber, M.H.W.). Tectum Publishing House, Marburg, in press.

Bazinet, A.L., Myers, D.S., Fuetsch, J. & Cummings, M.P. (2007). Grid Services Base Library: a high-level, procedural application programming interface for writing Globus-based Grid services. *Future Gener. Comp. Sy.*, 23, 517–522.

Cowling, R.M., Pressey, R.L., Rouget, M. & Lombard, A.T. (2003a). A conservation plan for a global biodiversity hotspot – the Cape Floristic Region, South Africa. *Biol. Conserv.*, 112, 191–216.

Cowling, R.M., Pressey, R.L., Sims-Castley, R., le Roux, A., Baard, E., Burgers, C.J. *et al.* (2003b). The expert or the algorithm? – comparison of priority conservation areas in the Cape Floristic Region identified by park managers and reserve selection software. *Biol. Conserv.*, 112, 147–167.

Cummings, M.P. & Huskamp, J.C. (2005). Grid computing. *EDUCAUSE Rev.*, 40, 116–117.

Environmental Systems Research Institute (1999–2004). *ArcGIS 9.0*. ESRI, Redlands, CA.

Forshaw, N. (1998). *Protea Atlas Project*. Available at: URL http://protea.worldonline.co.za

Freitag, S. & Van Jaarsveld, A.S. (1998). Sensitivity of selection procedures for priority conservation areas to survey extent, survey intensity and taxonomic knowledge. *P. Roy. Soc. Lond. B. Bio.*, 265, 1475–1482.

Freitag, S., Nicholls, A.O. & van Jaarsveld, A.S. (1996). Nature reserve selection in the Transvaal, South Africa: what data should we be using? *Biodivers. Conserv.*, 5, 685–698.

Freitag, S., Nicholls, A.O. & van Jaarsveld, A.S. (1998). Dealing with established reserve networks and incomplete distribution data sets in conservation planning. *S. Afr. J. Sci.*, 94, 79–86.

Gaston, K.J. & Rodrigues, A.S.L. (2003). Reserve selection in regions with poor biological data. *Conserv. Biol.*, 17, 188–195.

Gladstone, W. & Davis, J. (2003). Reduced survey intensity and its consequences for marine reserve selection. *Biodivers. Conserv.*, 12, 1525–1536.

Goldblatt, P. & Manning, J. (2000). Cape plants: a conspectus of the Cape Flora of South Africa. *Strelitzia*, 9, 1–743.

Good, P. (1994). *Permutation Tests, A Practical Guide to Resampling Methods for Testing Hypotheses*. Springer-Verlag, New York.

Harrison, J., Miller, K. & McNeely, J. (1984). The world coverage of protected areas: development goals and environmental needs. In: *National Parks, Conservation and Development: The Role of Protected Areas in Sustaining Society. Proceedings of the World Congress on National Parks, Bali, Indonesia, 11–22 October 1982* (eds. McNeely, J.A. & Miller, K.R.). Smithsonian Institution Press, Washington, DC, pp. 24–33.

Jaccard, P. (1912). The distribution of the flora of the alpine zone. *New Phytol.*, 11, 37–50.

Justus, J. & Sarkar, S. (2002). The principle of complementarity in the design of reserve networks to conserve biodiversity: a preliminary history. *J. Biosci.*, 27, 421–435.

Lombard, A.T., Cowling, R.M., Pressey, R.L. & Rebelo, A.G. (2003). Effectiveness of land classes as surrogates for species in conservation planning for the Cape Floristic Region. *Biol. Conserv.*, 112, 45–62.

Manly, B.F.J. (1991). *Randomization and Monte Carlo Methods in Biology*. Chapman & Hall, London.

Maritz, J.S. (1995). *Distribution-free Statistical Methods, Second Edition. Monographs on Statistics and Applied Probability 17*. Chapmap & Hall, London.

McDonnell, M.D., Possingham, H.P., Ball, I.R. & Cousins, E.A. (2002). Mathematical methods for spatially cohesive reserve design. *Environ. Model. Assess.*, 7, 107–114.

Meir, E., Andelman, S. & Possingham, H.P. (2004). Does conservation planning matter in a dynamic and uncertain world? *Ecol. Lett.*, 7, 615–622.

Polasky, S., Camm, J.D., Solow, A.R., Csuti, B., White, D. & Ding, R. (2000). Choosing reserve networks with incomplete species information. *Biol. Conserv.*, 94, 1–10.

Possingham, H.P., Day, J., Goldfinch, M. & Salzborn, F. (1993). The mathematics of designing a network of protected areas for conservation. In: *Decision Sciences: Tools for Today, 12th National ASOR Conference* (eds. Suttion, D.I., Pearce, C.E.M. & Cousins, E.A.). ASOR, Adelaide, pp. 536–545.

Possingham, H.P., Ball, I.R. & Andelman, S. (2000). Mathematical methods for identifying representative reserve networks. In: *Quantitative Methods for Conservation Biology* (eds Ferson, S. & Burgman, M.). Springer-Verlag, New York, pp. 291–305.

Pressey, R.L. & Tully, S.L. (1994). The cost of ad hoc reservation: a case study in western New South Wales. *Aust. J. Ecol.*, 19, 375–384.

Pressey, R.L., Humphries, C.J., Margules, C.R., Vanewright, R.I. & Williams, P.H. (1993). Beyond opportunism – key principles for systematic reserve selection. *Trends Ecol. Evol.*, 8, 124–128.

Pressey, R.L., Possingham, H.P. & Day, J.R. (1997). Effectiveness of alternative heuristic algorithms for identifying indicative minimum requirements for conservation reserves. *Biol. Conserv.*, 80, 207–219.

Pressey, R.L., Possingham, H.P., Logan, V.S., Day, J.R. & Williams, P.H. (1999). Effects of data characteristics on the results of reserve selection algorithms. *J. Biogeogr.*, 26, 179–191.

Rabinowitz, D., Cairns, S. & Dillon, T. (1986). Seven forms of rarity and their frequency in the flora of the British Isles. In: *Conservation Biology: The Science of Scarcity and Diversity* (ed. Soulé, M.E.). Sinauer Associates, Inc., Sunderland, MA, pp. 182–204.

Rebelo, A.G. & Siegfried, W.R. (1992). Where should nature reserves be located in the Cape Floristic Region, South Africa? Models for the spatial configuration of a reserve network aimed at maximizing the protection of floral diversity. *Conserv. Biol.*, 6, 243–252.

Ricketts, T.H., Dinerstein, E., Boucher, T., Brooks, T.M., Butchart, S.H.M., Hoffmann, M. *et al.* (2005). Pinpointing and preventing imminent extinctions. *Proc. Natl Acad. Sci. U.S.A.*, 102, 18 497–18 501.

Rodrigues, A.S.L. & Gaston, K.J. (2001). How large do reserve networks need to be? *Ecol. Lett.*, 4, 602–609.

Rodrigues, A.S.L. & Gaston, K.J. (2002). Rarity and conservation planning across geopolitical units. *Conserv. Biol.*, 16, 674–682.

Rouget, M., Reyers, B., Jonas, Z., Desmet, P., Driver, A., Maze, K. *et al.* (2005). *South African National Biodiversity Assessment 2004: Technical Report. Volume 1: Terrestrial Component*. South African National Biodiversity Institute, Pretoria.

Spearman, C. (1904). The proof and measurement of association between two things. *Am. J. Psychol.*, 15, 72–101.

Stewart, R.R. & Possingham, H.P. (2005). Efficiency, costs and trade-offs in marine reserve system design. *Environ. Model. Assess.*, 10, 203–213.

Stewart, R.R., Noyce, T. & Possingham, H.P. (2003). Opportunity cost of ad hoc marine reserve design decisions: an example from South Australia. *Mar. Ecol. Prog. Ser.*, 253, 25–38.

Virolainen, K.M., Virola, T., Suhonen, J., Kuitunen, M., Lammi, A. & Siikamaki, P. (1999). Selecting networks of nature reserves: methods do affect the long-term outcome. *Proc. R. Soc. Lond. B Bio.*, 266, 1141–1146.

Warman, L.D., Sinclair, A.R.E., Scudder, G.G.E., Klinkenberg, B. & Pressey, R.L. (2004). Sensitivity of systematic reserve selection to decisions about scale, biological data, and targets: case study from Southern British Columbia. *Conserv. Biol.*, 18, 655–666.

Wilcove, D.S., Rothstein, D., Dubow, J., Phillips, A. & Losos, E. (1998). Quantifying threats to imperiled species in the United States. *BioScience*, 48, 607–615.

Editor, Michael Hochberg
Manuscript received 30 November 2006
First decision made 5 January 2007
Second decision made 19 January 2007
Manuscript accepted 1 February 2007