



Place prioritization for biodiversity conservation using probabilistic surrogate distribution data

Sahotra Sarkar,* Christopher Pappas, Justin Garson, Anshu Aggarwal and Susan Cameron

Biodiversity and Biocultural Conservation Laboratory, Section of Integrative Biology, University of Texas at Austin, 1 University Station, C3500, Austin, TX 78712–1180, USA

ABSTRACT

We analyse optimal and heuristic place prioritization algorithms for biodiversity conservation area network design which can use probabilistic data on the distribution of surrogates for biodiversity. We show how an Expected Surrogate Set Covering Problem (ESSCP) and a Maximal Expected Surrogate Covering Problem (MESCP) can be linearized for computationally efficient solution. For the ESSCP, we study the performance of two optimization software packages (XPRESS and CPLEX) and five heuristic algorithms based on traditional measures of complementarity and rarity as well as the Shannon and Simpson indices of α -diversity which are being used in this context for the first time. On small artificial data sets the optimal place prioritization algorithms often produced more economical solutions than the heuristic algorithms, though not always ones guaranteed to be optimal. However, with large data sets, the optimal algorithms often required long computation times and produced no better results than heuristic ones. Thus there is generally little reason to prefer optimal to heuristic algorithms with probabilistic data sets.

Keywords

Area selection, conservation planning, integer programming, linear programming, probabilistic data analysis, reserve network selection, site selection.

*Correspondence: Section of Integrative Biology, University of Texas at Austin, TX 78712-1180, USA. Tel.: 1(512) 232 7133. Fax.: 1(512) 471 4806. E-mail: sarkar@mail.utexas.edu

INTRODUCTION

Place prioritization on the basis of biodiversity content for the design of conservation area networks (CANs) emerged as an explicit goal of systematic conservation planning in the 1980s (Margules & Usher, 1981). Conservation areas include, but are not limited to, traditional parks and reserves. They are defined as places at which some biodiversity conservation plan is implemented (this is why we use the term ‘place prioritization’ instead of the older ‘reserve selection’). Prioritization is necessary because resource limitations preclude the conservation of all areas of biological interest. In general, two methods for place prioritization have been used (Cabeza & Moilanen, 2001; Cowling *et al.*, 2003): (i) sets of places have been selected on the basis of expert advice (Dinerstein *et al.*, 2000); or (ii) sets of places have been selected using algorithmic procedures which incorporate biological criteria such as complementarity or rarity. These algorithms typically attempt to meet explicit conservation targets such as specified levels of representation for biota or specified proportions of land to be put under a conservation plan (Margules *et al.*, 1988; Margules & Pressey, 2000). While these methods are sometimes presented as incompatible alternatives, both have a useful role in conservation planning and can be

integrated, for instance, by using expert advice to filter algorithmic results.

This note studies the relatively unexplored class of place prioritization algorithms (PPAs) that are capable of using probabilistic data (past efforts are listed in Appendix 1). In what follows, the term ‘surrogate’ will be used to refer to features that represent biodiversity in models; these surrogates can be both biological features (species or other taxa, vegetation types, etc.) as well as nonbiological attributes of places (soil types, temperature, precipitation, etc.) (Sarkar & Margules, 2002). The important properties that adequate surrogates must have is that their distributions can be accurately and easily assessed in the field or reliably modelled.

In general, PPAs can be evaluated on the basis of six criteria:

- 1 *economy* (Margules *et al.*, 1988): PPAs should achieve the desired representation of surrogates in a minimum number of places or maximal representation of surrogates in a fixed number of places (this is sometimes called ‘efficiency’, but that term, in accordance with usage in computer science, will be reserved here for the temporal performance of algorithms);
- 2 *efficiency* (Pressey *et al.*, 1996): PPAs must be able to resolve data sets rapidly. This issue has become increasingly important as planning processes, especially those incorporating multiple

criteria (for instance, economic, political, and social criteria), may require the evaluation of hundreds of alternative scenarios;

3 flexibility (Church *et al.*, 1996): ideally, a PPA should allow the incorporation of a wide variety of criteria (e.g. a preference for size, compactness, or connectivity, in a CAN);

4 transparency: it should be clear why each individual place is selected. This is important because, should such an area be removed from a CAN, planners should be able to determine what biodiversity feature is lost;

5 universality: a PPA should be able to resolve data sets on any feature of biodiversity from anywhere in the world. It should not be specific to some surrogate set or region;

6 modularity: two different kinds of modularity are valuable. Flexibility and universality require that a PPA as a whole be a module that can be easily transported to different planning contexts. Flexibility and transparency both require internal modular organization: it should be possible to turn on or off individual criteria used in an algorithm.

Traditionally PPAs have assumed the availability of binary (1 or 0) presence-absence data, with 1 representing presence and 0 absence. However, surrogate presence or occurrence data available to conservation planners are often probabilistic for three main reasons: **1** the only data available are from inadequate (incomplete or unsystematic) surveys or from activities that do not constitute proper surveys. All such data must be regarded as uncertain. If probabilities are used to represent surrogate occurrence, it is sometimes possible to take these uncertainties into account;

2 much of conservation planning must use modelled rather than observational data. In general such models predict only probabilities of the occurrence of some surrogate, for instance, a species in a habitat;

3 ideally, a conservation plan should not be based on the present distribution of surrogates but on the likelihood of their persistence into the future. These can only be predicted as probabilities.

In the past, probabilistic data have often been converted to binary data using plausible thresholds because algorithms permitting the use of probabilities were not available. For instance, Margules & Nicholls (1987) used a 0.95 probability threshold for species data: a species with a probability of occurrence of less than 0.95 was considered absent, one with a probability greater than or equal to 0.95 was considered present. There are three problems with this approach: (i) any chosen threshold is arbitrary; (ii) treatment using a threshold results in a loss of relevant information (Arthur *et al.*, 2002). The difference between a probability of persistence of 0.90 and 0.80, is lost when a 0.95 probability threshold is used; that between a probability of 0.94 and 0.96 gets more emphasis than is justified; and (iii) at least in some cases plans using thresholds are less economical than those using the probabilities themselves (Arthur *et al.*, 2002).

THEORETICAL FRAMEWORK

Formal Problems

The theoretical framework developed here refers to individual places prioritized for conservation action as 'cells.' These cells

may have variable shapes and areas. Given a list of cells, Σ ($\sigma_j \in \Sigma$, $j = 1, 2, \dots, n$), a list of surrogates, Λ ($\lambda_i \in \Lambda$, $i = 1, 2, \dots, m$), a target of expected coverage, τ_i ($i = 1, 2, \dots, m$), for each surrogate, and a probability, p_{ij} ($i = 1, 2, \dots, m; j = 1, 2, \dots, n$), of the occurrence of λ_i (the i -th surrogate) at σ_j (the j -th cell), the place prioritization problem takes two canonical forms:

1 select the smallest set of cells, Γ , such that every surrogate meets its assigned target;

2 given the size (cardinality), κ ($\leq n$), of a set, Γ , select those cells that maximize the expected coverage of each surrogate.

After this subsection, we restrict our attention to the first form because it is more commonly encountered in conservation contexts and, unlike the second form, it has not been previously systematically studied using probabilistic data (see, however, Faith & Walker (1996)). In the context of probabilistic data, the first form will be called the Expected Surrogate Set Covering Problem (ESSCP); the second form will be called the Maximal Expected Surrogate Covering Problem (MESCP) provided that explicit targets are imposed (past work, and the origin of these names, is described in Appendix 1).

For the second canonical form, explicit targets have not been imposed in past work. However, in biological contexts, there are at least two reasons why explicit targets of representation for surrogates should be imposed for this problem: (i) a very low level of expected coverage may leave a surrogate vulnerable to extinction; and (ii) in the absence of a target an algorithm may continue to try to increase the expected coverage for a surrogate beyond what is biologically relevant at the cost of insufficient attention to other surrogates.

Moreover, in the absence of explicit targets, there is no known algorithm that guarantees finding the optimal solution of the second canonical form short of sequential exploration of every subset of Σ , iteratively increasing the cardinality (that is, the method of exhaustion). Let E_j be the event that some surrogate (say the i -th surrogate) does not occur in the j -th cell. Let $P\left(\bigcap_{j=1}^s E_j\right)$ be the probability that it does not occur in all s selected cells. In the notation being used in this argument, $1 - p_{ij} = P(E_j)$. Then:

$$1 - P\left(\bigcap_{j=1}^s E_j\right) = 1 - P(E_1)P(E_2|E_1)P(E_3|E_1 \cap E_2) \dots P\left(E_n \mid \bigcap_{j=1}^{s-1} E_j\right)$$

is the probability of its occurrence somewhere. This parameter is almost impossible to compute in practice, let alone assess in the field.

Two assumptions of independence about the data set are routinely used to simplify the problem. Suppose that:

1 the probability of the occurrence of the i -th surrogate at the j -th cell is independent of the probability of its occurrence at every other cell (that is, $\forall i, j, e, p_{ij}$ and p_{ie} are independent). This is almost always a poor assumption because the distributions of biological surrogates are usually strongly spatially correlated; and

2 the probability of occurrence of one surrogate in a given cell is independent of the probability of the occurrence of every other

surrogate in that cell (that is, $\forall i, j, e, p_{ij}$ and p_{ej} are independent). This assumption is also often false because it implies that no ecological relation holds between any pair of biological surrogates.

With these assumptions, $\prod_{j=1}^s (1 - p_{ij})$ is the probability that the surrogate does not occur anywhere in the set of places. Then the probability that it does occur is given by $1 - \prod_{j=1}^s (1 - p_{ij})$ (Cocks & Baird, 1989). Polasky *et al.* (2000) emphasized the computational complexity of maximizing this nonlinear quantity; they explored a heuristic algorithm for its solution which is one of those we use below (namely, C); Arthur *et al.* (2002) and ReVelle *et al.* (2002) also noted that it cannot be linearized. Without the imposition of targets, the second canonical form is intractable.

However, once targets are imposed, these simplifying assumptions are not necessary for finding computationally efficient optimal algorithms for solving either the ESSCP or the MESCP. Both problems can then be linearized (see ‘Mathematical Programming Problems’ below). However, the development of such algorithms requires the use of expectations instead of probabilities.

Given the p_{ij} , $s_i = \sum_{j=1}^n p_{ij}$ is the expected number of cells in which the i -th surrogate will occur in Σ . Similarly, $b_j = \sum_{i=1}^m p_{ij}$ is the expected number of surrogates occurring in the j -th cell, σ_j . These interpretations do not require any assumption about independence (of the probabilities of occurrence of different surrogates within a cell or the same surrogate in different cells). This is a significant advantage of using expectations of coverage in algorithms rather than the more traditional probabilities of occurrence (or, sometimes, persistence) since such independence assumptions are routinely violated in biological contexts. The ESSCP now becomes that of finding a set Γ of cells of lowest cardinality in which all the expected occurrences of the surrogates, λ_i meet their required targets. The MESCP is that of finding, for a given $\kappa \leq n$, a set of cells, Γ , of cardinality κ that maximizes the expected number of surrogates that meet their targets.

MATHEMATICAL PROGRAMMING PROBLEMS

Exact solutions to CAN selection problems can be generated using techniques of mathematical programming (linear programming (LP), integer programming (IP), mixed integer programming (MIP), and nonlinear integer programming (NLIP)) by modelling the problem as one of constrained optimization, that is, the maximization or minimization of an objective function subject to specified constraints on the variables of that function. Many known efficient algorithms exist for solving LPs, but IPs and NLIPs belong to a class of problems labelled NP-hard for which there is no known polynomial-time algorithm (Karp, 1972). Problems associated with CAN design cannot be properly modelled as LPs because some variables, such as the number of cells in a CAN, must have integer values. Mixed integer linear programs (MIPs) are similar to LPs with the added restriction that at least one of the variables assumes an integer value.

CAN design involves MIPs. NLIPs are even more difficult problems.

To solve the ESSCP and the MESCP using mixed integer linear programming, the variables X_j ($j = 1, 2, \dots, n$) and Y_i ($i = 1, 2, \dots, m$) are defined as follows:

$$X_j = \begin{cases} 1, & \text{if cell } \sigma_j \in \Gamma \\ 0, & \text{if cell } \sigma_j \notin \Gamma \end{cases}; \quad \text{and} \quad Y_i = \begin{cases} 1, & \text{if } \sum_{\sigma_j \in \Gamma} p_{ij} > \tau_i \\ 0 & \text{otherwise} \end{cases}.$$

The ESSCP consists of the problem: Minimize $\sum_{j=1}^n X_j$ such that $\sum_{j=1}^n X_j p_{ij} \geq \tau_i$ for $\forall \lambda_i \in \Lambda$. The MESCP consists of the problem: Maximize $\sum_{i=1}^m Y_i$ such that $\sum_{j=1}^n X_j = \kappa$ where κ is the fixed number of cells. Previous work probably missed the possibility of linearizing the second canonical form of the original problem only because explicit targets were not imposed.

However, the use of targets in this way comes at a potential cost: it is possible to select a cell so that the coverage of one surrogate is minimally increased to achieve its target while another remains entirely unrepresented (this problem was encountered in the field during conservation planning in New South Wales (Dan Faith, personal communication)).

HEURISTIC RULES

Most heuristic PPAs go back to the pioneering algorithm introduced by Margules *et al.* (1988) which was based on complementarity. Many studies since have shown that the most economical solutions are found with complementarity and rarity (Csuti *et al.*, 1997; Pressey *et al.*, 1997). We use both below along with two indices of α -diversity, which have not previously been used in PPAs.

The heuristic PPAs we develop can be used to solve either the ESSCP or the MESCP though we only report results on the former in this note. The measures used to generate our heuristic rules are:

1 rarity: in our framework the rarest surrogate, λ_i , is the one with the lowest s_i ; the cell with the highest rarity rank is the one for which the rarest surrogate has the highest probability of occurrence (note that several different cells can have the same rarity rank). Rarity behaves very differently with probabilistic data compared to binary data. For the latter, there are far fewer different rarity classes because the s_i must have integral values. This results in far more ties after the use of rarity;

2 complementarity: the complementarity value for a new cell, σ_p , to be potentially added to Γ' (the potential Γ that is being updated during algorithm execution) will be the sum $b_j = \sum_{i=1}^{m'} p_{ij}$ where the m' indicates that the summation is carried out only for those surrogates for which the target has not yet been met;

3 the Shannon index of α -diversity: let q_{ij} be the relative proportion of λ_i in σ_j ; this requires that $\sum_{i=1}^m q_{ij} = 1$. Then, for each cell, σ_p the Shannon index, $Sh(j) = -\sum_{i=1}^m q_{ij} \ln q_{ij}$. This index has been used

in ecological contexts since its introduction by Margalef (1958) and measures the evenness of surrogate occurrence in a cell (Magurran, 1988), reducing the chance that any surrogate occurs with minimal probability in a cell (which usually implies a high future vulnerability);

4 the *Simpson index* of α -diversity: the Simpson index of α -diversity is given by $S_i(j) = 1 - D_j$ where $D_j = \sum_{i=1}^m q_{ij}^2$ (another form is $1/D_j$; in all our computations, it did not give results different from $1 - D_j$). This index also measures the evenness of surrogate occurrence in a cell;

5 *redundancy*: a cell, σ_i , in Γ' is redundant if it can be removed from Γ' with no surrogate losing its ability to meet its target. In heuristic PPAs redundant cells may be present in Γ' because a cell selected early may have been made redundant by later selected cells.

METHODS

Data Sets and Problems

Our analysis used 10 artificially created data sets and one empirical data set. The computational profiles of these data sets are summarized in Table 1. The artificial data sets were created with associated targets of expected coverage for surrogates. A data set and an associated target specifies a 'problem.' Problems with artificial data sets are characterized by three parameters: the number of cells, n ; the number of surrogates, m ; and the target of repre-

sentation for the place prioritization procedure, t . Once these parameters are specified, the artificial data sets were constructed as follows: a matrix is created with the number of rows equal to the number of cells, n , and the number of columns equal to the number of surrogates, m . A maximum target, t , is selected; and a target is assigned for each surrogate by randomly selecting a number between 1 and t .

The empirical data set used is the modelled distribution of 46 major vegetation types of Ecuador (Sierra, 1999; Sierra *et al.*, 2002). This data set has been used to identify potential additions to Ecuador's biodiversity reserve network (Cameron, 2003). It is being used here only to test the efficiency of the algorithms when presented with a typical data set of actual relevance to conservation planning. The original grid of cells consisted of 6×10^6 200×200 m² data units. To reach a scale more relevant to conservation planning this data set was modified to a 2×2 km² scale by merging 100 of the original cells into a new cell. After modification, there were 61,554 cells. The probability of occurrence of each vegetation type in a cell was set equal to the proportion of the original data cells that contain that vegetation type (each original data cell had exactly one vegetation type). Thus, within each cell, these probabilities sum to 1. 23,827 of the cells were excluded from consideration because they were anthropogenically modified, leaving 37,727 cells. For this data set, two problems were generated, one with a target of 1 for the expected coverage of each surrogate, and the other with a target of 10% of the total surrogate coverage for each surrogate. These will be referred to as Ecuador (1) and Ecuador (10%).

Table 1 This table describes the computational profiles of the 12 problems corresponding to 11 total data sets (for a description of the data sets, see Methods section on Data Sets and Problems). Columns 1–10 correspond to the 10 artificial data sets; columns 11 and 12 are for the Ecuador data set with two different targets. The first and second rows depict the number of cells (n) and surrogates (m), and the third row, 'Data Points' (DP) shows the total number of elements in the matrix ($n \times m$). The fourth row, 'Surrogate Target Range', shows the range within which surrogate targets were randomly chosen, from 1 to t . The row 'Data Points with Surrogate Presence' (DPSP) depicts the number of nonzero elements in the ($n \times m$) matrix, and the 'Density' calculation for each dataset is DPSPDP. The 'Average Surrogate Probability per Cell for All Cells' (ASPC) is the average of the probabilities of all the surrogates in a cell. The 'Average Number of Surrogates per Cell' (ANSC) is just what it says (Data Set Profiles)

	1	2	3	4	5	6	7	8	9	10	Ecuador (1)	Ecuador (10%)
Cells	100	1000	1000	1000	1000	1000	1000	1000	1000	10,000	37,727	37,727
Surrogates	20	20	20	20	20	20	20	20	20	20	46	46
Data Points	2000	20,000	20,000	20,000	20,000	20,000	20,000	20,000	20,000	200,000	1,735,442	1,735,442
Surrogate Target Range (1 – t)	1–5	1–5	1–5	1–10	1–10	1–5	1–5	1–5	1–10	1–5	n/a	n/a
Data Points with Surrogate Presence	596	4516	4718	4806	8225	9313	9841	10,832	10,868	45,279	52,667	52,667
Density	0.30	0.23	0.24	0.24	0.41	0.47	0.49	0.54	0.54	0.23	0.03	0.03
Average Surrogate Probability per Cell For All Cells	0.10	0.08	0.08	0.08	0.15	0.17	0.18	0.21	0.21	0.08	0.02	0.02
Average Number of Surrogates per Cell	5.96	4.52	4.72	4.81	8.23	9.31	9.84	10.83	10.87	4.53	1.40	1.40

COMPUTATIONS

Exact (optimal) methods

We used two publicly available mathematical programming software packages in an attempt to solve the ESSCP problems to optimality:

1 we submitted the datasets in GAMS (General Algebraic Modelling System) format through the *NEOS Server 4.0* (<http://www-neos.mcs.anl.gov/neos/>). NEOS is an optimization portal that sent the IPs to be solved over various UNIX servers running the *Dash XPRESS* optimization software; and

2 we formatted the IPs in the CplexLP modelling language to run them interactively on a server with the *ILOG CPLEX 7.5* software package. The server was running the Windows™ Terminal Server operating system with dual Pentium III 733 MHz processors, 1024 MB of memory, and 2.5 gigabytes of free hard drive space.

Both optimization methods had space and run time constraints. The XPRESS constraints were more stringent. Neither space nor time restrictions are published by NEOS, but run times were not allowed more than 20 min. For CPLEX, there was no direct time constraint. We had unrestricted access during our sessions. However, in some cases a session was terminated due to secondary memory limitations and a guaranteed optimal solution was not obtained.

When no guaranteed optimal solution was found, two 'sub-optimal' methods were used to obtain approximate solutions:

1 in this method, which can only be used with CPLEX, for the first step, we 'relaxed' the IPs to linear programs by lifting the integer requirements on the variables. Each solution from this step was analysed to see which, if any, of the variables (which correspond to cells) had assumed the integral upper bound of 1. A variable that had a value of 1 in the LP solution represented a cell chosen for inclusion in that solution. The variables which assumed integer values in the relaxed solution were then set to 1 in the IP problem which was then solved to satisfy the constraints for the remaining variables. The resulting solution is not guaranteed as optimal;

2 we allowed XPRESS and CPLEX to run on the original IP until resource limitations were violated, at which point the best feasible integer solution, along with its run time and approximation value, were recorded (appendix 2 discusses the details of the approximation calculations).

Heuristic algorithms

We encoded five heuristic algorithms as modifications of the ResNet software package (Garson *et al.*, 2002), the first three of which are commonly used, though almost always only with binary data, while the last two are being explored for the first time. Csuti *et al.* (1997) and Pressey *et al.* (1997) report results (using binary data) on variants of our first three algorithms. We selected the three that were found to be the most efficient in their and other work (see Sarkar *et al.* (2002)):

1 C: both initialization and the iterative step used complementarity;

2 RC-Ra: the first cell was selected by rarity. Then, rarity was used first, followed by complementarity, with remaining ties broken by lexical order. For binary data this algorithm is generally the most economical (Csuti *et al.*, 1997; Pressey *et al.*, 1997);

3 RC-C: the first cell was selected by complementarity. Then, rarity was used first, followed by complementarity, with remaining ties broken by lexical order.

4 Sh: only the Shannon index of α -diversity was initially used. At the end of the selection regime, when several cells have only 1 surrogate that has not met its target, ties were broken using complementarity;

5 Si: the Simpson index of α -diversity in the form $1 - D$ was used. Ties at the end of the selection regime were broken by complementarity.

In each algorithm, redundant cells were always removed at the end. If there was more than one redundant cell, ties were broken by iteratively using rarity, with the cells with rarer surrogates being preferentially retained.

RESULTS

Table 2 reports results obtained using the exact algorithms. Problems 1, 4, 5 and 9 were fully resolved to optimality by both our exact methods. Problem 7 was resolved to optimality using XPRESS though not CPLEX. For Problems 2, 3, 6, and 10, approximate solutions were obtained by our first suboptimal method (using CPLEX). For Problem 8 the first suboptimal method did not produce a solution. When the first suboptimal method was used, in all cases the tractability of the resultant IPs was partially dependent on the number of variables that were assigned a value of 1 in the LP problem. For example, in the LP solution for Problem 2, there were 14 variables which were assigned 1, whereas the LP solution for Problem 8 had only 2. This difference was reflected in their run times: Problem 2 was solved in the IP step in less than 267 s, but Problem 8 was only able to be solved to within a 9.25% approximation of the optimal solution in a run that took over 9 h.

Table 3 reports the corresponding results obtained for the 5 heuristic PPAs. For the 5 problems solved to optimality in Tables 2 and 4 were also solved to optimality by the heuristic methods (Problems 1, 4, 5 and 7); Sh solved all 4; while C solved only Problems 1 and 4. C, RC-Ra, and RC-C all did better than the optimal PPAs for Ecuador (1). Sh and Si did as well as CPLEX for Ecuador (10%) though the last took over five hours for the computation. C only did marginally worse (by 0.3%). In general, among the heuristic PPAs, C performed best with respect to economy, with RC-R and Sh ranking next. Sh outperformed C for Problems 3, 5, and Ecuador (10%). RC-Ra and RC-C fared worse than how their counterparts generally perform with binary data (Csuti *et al.*, 1997; Pressey *et al.*, 1997).

The computation times for the optimal PPAs, as indicated in Table 2, span a wide range. The only general feature is that, if these PPAs found an optimal solution, they did so very rapidly (≤ 0.25 s). (The only exception was Problem 9: XPRESS took 2.08 s). For the first 10 problems, the heuristic PPAs resolved the data in at most 3 s. For Ecuador (1), the heuristic PPAs again

Table 2 This table depicts the results of XPRESS and CPLEX on the 12 problems. The columns are the same as in Table 1. For both procedures, the best solution sizes found from the various methods are listed first followed by the total computation time taken to obtain that solution. A 0 means that the solution time was less than 1 second. Solutions obtained using the first suboptimal method, are marked with an ‘*’, and solutions obtained using the second suboptimal method are marked with a ‘†’. In both cases, the integer solution is not guaranteed to be an optimal solution but is still guaranteed to be within some approximation of the size of the optimal solution. The approximation is listed in parentheses and explained below. In the case of the first suboptimal method for CPLEX, the second row corresponds to the number of cells that were set to 1 in the modified IP (see Methods section on Computations). Solutions marked with neither an ‘*’ nor a ‘†’ are optimal solutions (Solutions Using Optimal Algorithms)

Exact Methods	1	2	3	4	5	6	7	8	9	10	Ecuador (1)	Ecuador (10%)
CPLEX	10	25*	21*	20	20	14*	13*	13† (9.25%)	20	20*	65† (10.71%)	3783† (0.27%)
Number of Cells Set to 1	n/a	14	11	n/a	n/a	6	4	2	n/a	10	n/a	n/a
Total Solution Time (seconds)	0	266	1419	0	0	850	5061	34,819	0	6187	16,018	19,080
XPRESS	10	26† (8.33%)	22† (10%)	20	20	14† (7.69%)	13	13† (8.33%)	20	20† (5.26%)	56† (19.15%)	n/a
Total Solution Time (seconds)	0	97	61	0	0	1164	161	2	2	271	532	n/a

Table 3 This table depicts the results of five heuristic methods used on the 12 different problems. The columns are the same as in Table 1. The solution sizes for the five heuristic procedures (C, RC-Ra, RC-C, Sh and Si) are listed as the top entry of each cell. The time to solution (in seconds) is listed as the bottom entry as in Table 2; and 0 means that the solution was obtained in less than 1 second (Solutions Using Heuristic Algorithms)

Heuristic Algorithms	1	2	3	4	5	6	7	8	9	10	Ecuador (1)	Ecuador (10%)
C	10 0	27 0	23 0	20 0	22 1	16 0	15 0	14 0	23 1	22 2	53 18	3784 837
RC-Ra	16 0	29 0	26 0	41 0	32 1	17 0	16 0	15 0	28 1	25 2	50 17	3797 838
RC-C	16 0	28 0	26 0	41 0	32 1	17 0	15 0	15 0	27 1	24 2	51 23	3797 835
Sh	10 0	40 0	22 0	20 1	20 0	20 0	20 1	21 0	20 1	25 3	62 37	3783 1075
Si	10 0	46 0	22 0	20 1	20 0	25 1	23 0	20 1	21 0	26 4	62 36	3783 1064

resolved the data in less than 1 minute. With Ecuador (10%) they took up to 18 min.

DISCUSSION

The only new theoretical result reported in this note is the linearization of the second canonical form of the original problem (‘Formal Problems’) to produce the MESCP. That linearization depends on the imposition of targets of representation. This move is justified because of the biological reasons already mentioned (in ‘Formal Problems’). For both the ESSCP and the MESCP, success in devising efficient algorithms depended on the use of expected coverages rather than probabilities of occurrence as an explicit goal. We plan to explore the value of using expectations rather than probabilities in conservation planning in future work.

Turning to the computations, our results show a general trend that, for small data sets, optimal PPAs, even when run only to produce approximate solutions, often give more economical

results than the best heuristic PPAs. For the rest of this discussion, we do not distinguish between optimal and approximate solutions produced by the optimal PPAs because, in the practical context of conservation planning, the distinction is not important. In the past, analyses of binary data showed only that optimal PPAs were rarely more economical than heuristic PPAs when they were able to resolve data sets optimally (Csuti *et al.*, 1997; Pressey *et al.*, 1997; however, Rodrigues *et al.* (2000) and Rodrigues & Gaston (2002) claim exceptions which will be discussed below). With probabilistic data and small sets, the advantage of optimal PPAs with respect to economy is more pronounced. However, the optimal PPAs found less economical solutions for the Ecuador (1) problem than any heuristic PPA except Sh and Si. CPLEX performed no better than the heuristic solutions for the best solution for Ecuador (10%). Rodrigues *et al.* (2000) claim that heuristic algorithms are significantly inferior with respect to economy to optimal ones (*viz.*, CPLEX). Our results contradict that conclusion. Rodrigues *et al.* (2000) report results

from the literature on heuristic algorithms but make no attempt to find the best heuristics. The heuristic algorithms they report do not seem to have always been optimally implemented.

The chief disadvantage of optimal PPAs remains their efficiency. Our first suboptimal method could not solve Problem 8 even to suboptimality even after CPLEX ran for almost 10 h. The three traditional heuristic PPAs (C, RC-C and RC-Ra) found marginally worse solutions than CPLEX in less than 1 second and XPRESS did equally well in 2 s. For Problem 7 and 10, CPLEX took more than an hour. For Ecuador (1) XPRESS took 14 times as much time as the worst heuristic PPA while finding less economical solutions; CPLEX took 430 times as much time and did even worse. Our results contradict those of Rodrigues & Gaston (2002) who claim that CPLEX is tractable enough for use with data from practical contexts of CAN implementation (for instance, Ecuador).

Because the additional economy achieved by optimal PPAs is not significant in all cases studied so far, this lack of efficiency is an impediment for their practical use in conservation planning. This is particularly true when hundreds of alternative plans incorporating different constraints have to be devised from complex data sets so that these plans can be judged against socio-political criteria that cannot be formally modelled. For instance, Cameron (2003) analysed the Ecuador (10%) problem using a slightly different implementation of RC-Ra to find 100 distinct solutions to determine the selection frequency of individual cells. This was possible because each solution was found in less than 10 min. It is hard to imagine attempting such an analysis with the optimization software packages we used. At the very least, optimization techniques must be devised which are more specifically geared to the structure of biological data sets.

The flexibility of optimal PPAs remains an open question. Many biological criteria — for instance, a preference for adjacent cells — can be represented as linear constraints (for an extended discussion, see Rodrigues *et al.* (2000)). However, typically, in a heuristic PPA, new criteria are used in a temporal order, that is, to break ties in cell selection that remain unbroken by previously used criteria. For instance, adjacency may be used only if the use of complementarity leaves ties. This type of hierarchical order, which has a straightforward biological interpretation, cannot be easily incorporated into optimal PPAs.

Partly because of the problem just discussed, transparency remains a serious problem for optimal PPAs. Optimal PPAs simultaneously utilize a large number of criteria which also leads to a loss of internal modularity. Additionally, since steps in optimization algorithms are motivated by mathematical considerations rather than biological ones, they typically do not have any biological interpretation. However, because of the generally increased economy of the solutions found by the optimal PPAs, for small data sets it is probably desirable to use optimal PPAs along with heuristic ones to get some quantitative measure of the loss of economy entailed by the use of the latter. It must then be judged whether the benefits offset the losses in that planning situation.

Among the heuristic algorithms, C outperformed the others systematically. This is in striking contrast to the generally better

performance of RC algorithms with binary data (Csuti *et al.*, 1997; Pressey *et al.*, 1997). This probably happened because we introduced no threshold of probability of occurrence below which a cell cannot be selected because of rarity. Sh and SI performed very well, especially with smaller data sets. Both have the natural biological interpretation of being a measure of α -diversity and are thus obvious candidates for use in methods designed to maximize biodiversity. They have previously not been studied in this context probably because they cannot be used with binary data. We plan to study various ways of combining Sh, Si, and C to generate economical and efficient heuristic PPAs.

Finally, while we informally compared algorithm performance for the complexity of the data sets (Problems 1–10) we could not identify any set of features that predicted the performance of an algorithm. Even size was not an unequivocal predictor. However, the various PPAs, both optimal and heuristic, tracked the problems in qualitatively similar ways. All found the same ones difficult or easy which suggests that there are definite characteristics of the data sets that explain the performance of various PPAs. This is a problem that will merit further study.

Software availability

The artificial data sets that were created for this analysis as well as the programs are freely available on request from the Biodiversity and Biocultural Conservation Laboratory, University of Texas at Austin (consbio@uts.cc.utexas.edu; <http://uts.cc.utexas.edu/~consbio/Cons/Labframeset.html>). ResNet can also be downloaded from that source.

ACKNOWLEDGEMENTS

We thank J. Wesley Barnes, Dan Faith, Steve Polasky, Kerrie Wilson, and an anonymous referee for helpful conversations and comments on earlier drafts.

REFERENCES

- Arthur, J.L., Hachey, M., Sahr, K., Huso, M. & Kiester, A.R. (1997) Finding all optimal solutions to the reserve site selection problem: formulation and computational analysis. *Environmental and Ecological Statistics*, **4**, 153–165.
- Arthur, J.L., Haight, R.G., Montgomery, C.A. & Polasky, S. (2002) Counterpart analysis of the threshold and expected coverage approaches to the probabilistic reserve site selection problem. *Environmental Modelling and Assessment*, **7**, 81–89.
- Cabeza, M. & Moilanen, A. (2001) Design of reserve networks and the persistence of biodiversity. *Trends in Ecology and Evolution*, **5**, 242–248.
- Cameron, S. (2003) Place prioritization and multiple criterion synchronization for biodiversity conservation in Ecuador using modelled vegetation classes as surrogates. Honours Thesis. Department of Biology, University of Texas at Austin.
- Camm, J.D., Polasky, S., Solow, A. & Csuti, B. (1996) A note on optimal algorithms for reserve site selection. *Biological Conservation*, **78**, 353–355.

- Church, R. & ReVelle, C.S. (1974) The maximal covering location problem. *Papers of the Regional Science Association*, **32**, 101–118.
- Church, R., Stoms, D.M. & Davis, F.W. (1996) Reserve selection as a maximal coverage location problem. *Biological Conservation*, **76**, 105–112.
- Cocks, K.D. & Baird, I.A. (1989) Using mathematical programming to address the multiple reserve selection problem: an example from the Eyre Peninsula, South Australia. *Biological Conservation*, **49**, 113–130.
- Cowling, R.M., Pressey, R.L., Sims-Castley, R., le Roux, A., Baard, E., Burgers, C.J. & Palmer, G. (2003) The expert or the algorithm? — Comparison of priority conservation areas in the Cape Floristic region identified by park managers and reserve selection software. *Biological Conservation*, **112**, 147–167.
- Csuti, B., Polasky, S., Williams, P.H., Pressey, R.L., Camm, J.D., Kershaw, M., Kiester, A.R., Downs, B., Hamilton, R., Huso, M. & Sahr, K. (1997) A comparison of reserve selection algorithms using data on terrestrial vertebrates in Oregon. *Biological Conservation*, **80**, 83–97.
- Daskin, M. (1983) A maximum expected covering location model: formulation, properties, and heuristic solution. *Transportation Science*, **17**, 48–70.
- Dinerstein, E., Powell, G., Olson, D., Wikramanayake, E., Abell, R., Loucks, C., Underwood, E., Allnutt, T., Wettengel, W., Ricketts, T., Strand, H., O'Connor, S. & Burgess, N. (2000) *A workbook for conducting biological assessments and developing biodiversity visions for ecoregion-based conservation. Part 1: Terrestrial ecoregions*. Conservation Science Program, WWF-USA, Washington, DC.
- Faith, D.P. & Walker, P.A. (1996) Integrating conservation and development: incorporating vulnerability into biodiversity-assessment of areas. *Biodiversity and Conservation*, **5**, 417–429.
- Garson, J., Aggarwal, A. & Sarkar, S. (2002) *Resnet ver 1.2 manual. Report Biodiversity and Biocultural Conservation Laboratory*. University of Texas at Austin, Austin.
- Haight, R., ReVelle, C. & Snyder, S. (2000) An integer optimization approach to a probabilistic reserve site selection problem. *Operations Research*, **48**, 697–708.
- Karmakar, N. (1984) A new polynomial time algorithm for linear programming. *Combinatorica*, **4**, 375–395.
- Karp, R.M. (1972) Reducibility among combinatorial problems. *Complexity of Computer Computations* (eds R.E. Miller and J.W. Thatcher), pp. 85–103. Plenum Press, New York.
- Khachian, L. (1979) A polynomial algorithm in linear programming. *Soviet Mathematics — Doklady*, **20**, 191–194.
- Magurran, A.E. (1988) *Ecological diversity and its measurement*. Princeton University Press, Princeton.
- Margalef, R. (1958) Information theory in ecology. *General Systems Yearbook*, **3**, 36–71.
- Margules, C.R. & Nicholls, A.O. (1987) Assessing the conservation value of remnant 'islands': mallee patches on the western Eyre Peninsula, South Australia. *Nature conservation: the role of remnants of native vegetation* (eds D.A. Saunders, G.W. Arnold, A.A. Burbidge and A.J.M. Hopkins), pp. 89–102. Surrey Beatty & Sons, Sydney.
- Margules, C.R., Nicholls, A.O. & Pressey, R.L. (1988) Selecting networks of reserves to maximize biological diversity. *Biological Conservation*, **43**, 63–76.
- Margules, C.R. & Pressey, R.L. (2000) Systematic conservation planning. *Nature*, **405**, 242–253.
- Margules, C.R. & Stein, J.L. (1989) Patterns in the distribution of species and the selection of nature reserves: an example from eucalyptus forests in south-eastern New South Wales. *Biological Conservation*, **50**, 219–238.
- Margules, C.R. & Usher, M.B. (1981) Criteria used in assessing wildlife conservation potential: a review. *Biological Conservation*, **21**, 79–109.
- Polasky, S., Camm, J., Solow, A., Csuti, B., White, D. & Ding, R. (2000) Choosing reserve networks with incomplete species information. *Biological Conservation*, **94**, 1–10.
- Possingham, H., Day, J., Goldfinch, M. & Salzborn, F. (1993) The mathematics of designing a network of protected areas for conservation. *Decision sciences: tools for today, proceedings of the 12th Australian operations research conference* (eds D. Sutton, E. Cousins and C. Pierce), pp. 536–545. ASOR, Adelaide.
- Pressey, R.L., Possingham, H.P. & Day, J.R. (1997) Effectiveness of alternate heuristic algorithms for identifying indicative minimum requirements for conservation reserves. *Biological Conservation*, **80**, 207–219.
- Pressey, R.L., Possingham, H.P. & Margules, C.R. (1996) Optimality in reserve selection algorithms: when does it matter and how much? *Biological Conservation*, **76**, 259–267.
- ReVelle, C.S., Williams, J.C. & Boland, J.J. (2002) Counterpart models in facility location science and reserve selection science. *Environmental Modelling and Assessment*, **7**, 71–80.
- Rodrigues, A.S., Cerdeira, J.O. & Gaston, K.J. (2000) Flexibility, efficiency, and accountability: adapting reserve selection algorithms to more complex conservation problems. *Ecography*, **23**, 565–574.
- Rodrigues, A.S. & Gaston, K.J. (2002) Optimisation in reserve selection procedures — why not? *Biological Conservation*, **107**, 123–129.
- Salkin, H.M. & Mathur, K. (1989) *Foundations of integer programming*. North-Holland, New York.
- Sarkar, S., Aggarwal, A., Garson, J., Margules, C.R. & Zeidler, J. (2002) Place prioritization for biodiversity content. *Journal of Biosciences*, **27** (S2), 339–346.
- Sarkar, S. & Margules, C.R. (2002) Operationalizing biodiversity for conservation planning. *Journal of Biosciences*, **27** (S2), 299–308.
- Sierra, R., ed. (1999) *Propuesta preliminar de un sistema de clasificación de vegetación para el Ecuador continental*. Proyecto INEFAN/GEF-BIRF y EcoCienca, Quito.
- Sierra, R., Campos, F. & Chamberlin, J. (2002) Assessing biodiversity conservation priorities: ecosystem risk and representativeness in continental Ecuador. *Landscape and Urban Planning*, **59**, 95–110.
- Toregas, C., Swain, R., ReVelle, C. & Bergman, L. (1971) The location of emergency service facilities. *Operations Research*, **19**, 1363–1373.
- Underhill, L.G. (1994) Optimal and suboptimal reserve selection algorithms. *Biological Conservation*, **70**, 85–87.

Williams, P.H. & Araújo, M.B. (2000) Using probability of persistence to identify important areas for biodiversity conservation. *Proceedings of the Royal Society (London)*, **267**, 1959–1966.

Williams, P.H. & Araújo, M.B. (2002) Apples, oranges and probabilities: integrating multiple factors into biodiversity conservation with consistency. *Environmental Modelling and Assessment*, **7**, 139–151.

APPENDIX 1

Background: past work

Past Analyses of Probabilistic Data

Attempts to use probabilistic data in the context of place prioritization go back to Margules & Nicholls (1987) and Margules & Stein (1989) who, however, convert probabilistic data into binary data using thresholds. Recent work that focuses on the general problems raised by probabilistic data (and not only on the design of algorithms) include Haight *et al.* (2000), Polasky *et al.* (2000), Williams & Araújo (2000, 2002) and Arthur *et al.* (2002).

Past Work on the Formal Prioritization Problems

In the context of binary data, the first canonical form was originally explicitly identified by Margules *et al.* (1988). In operations research it was known as the Location Set Covering Problem (LSCP), originally identified by Toregas *et al.* (1971). The biological version is the Species Set Covering Problem (SSCP) and its connection to the LSCP was first noted by Possingham *et al.* (1993). The second canonical form is the Maximal Covering Location Problem (MCLP) of operations research (Church & ReVelle, 1974); the corresponding biological problem is the Maximal Covering Species Problem (MCSP) (ReVelle *et al.*, 2002). With all surrogate targets implicitly set to 1, it was formulated by Cocks & Baird (1989) and analysed by Underhill (1994), Camm *et al.* (1996), Church *et al.* (1996), Rodrigues *et al.* (2000) and Rodrigues & Gaston (2002). Arthur *et al.* (1997) show how all optimal solutions can be obtained. The probabilistic version of the first canonical form is being explicitly studied here for the first time and will be called the Expected Surrogate Set Covering Problem (ESSCP). However, a variant of the problem, with targets interpreted as probabilities of persistence, was analysed by Faith & Walker (1996). The probabilistic version of the second canonical form is called the Maximal Expected Covering Location Problem (MEXCLP) in operations research; it was first formulated by Daskin (1983). A biological version was first analysed by Cocks & Baird (1989) and studied by Polasky *et al.* (2000). Arthur *et al.* (2002), Polasky *et al.* (2000) and ReVelle *et al.* (2002) emphasized the computational complexity of this problem. However, none of this previous work imposed targets for the surrogates. The version studied here will be called the Maximal Expected Surrogate Covering Problem (MESCP) only after explicit targets are imposed.

APPENDIX 2

The use of optimal algorithms to produce approximate solutions

CPLEX and XPRESS both use the branch-and-bound algorithm (Salkin & Mathur, 1989). Let r be the size of the best known integer solution and let q be the best objective function value, that is, $\sum_{j=1}^n X_j$ of all unexplored nodes. The best objective function value is calculated by taking the value of the objective function at every unexplored node in a search tree, where only some variables are constrained to be integers and the others are relaxed. Then the approximation value was calculated as follows: $(r - q)/r$. The approximation value denotes the quality of the solution and is possible without knowledge of the optimal solution itself, since q is always a lower bound on the size of the optimal solution. Thus, once an integer solution is found the quality of that solution can be roughly assessed by approximating how far it is from the LP solution at that step. At each step, the variables already 'branched on' will be constrained to be integers, while the others are allowed to take continuous values. Since polynomial time algorithms exist for linear programming problems (Khachian, 1979; Karmakar, 1984), LP solutions can usually be obtained in a reasonable amount of time and therefore are usually calculated during substeps of the problem. Both the XPRESS and CPLEX solvers dynamically invoke various techniques in order systematically to reduce the search space (or feasible region) associated with a problem. However, exactly what methods are used at every step and why they are invoked is not always transparent to the user. While attempting to solve Problems 2, 3, 6, 8 and 10, XPRESS was instructed to terminate once it obtained a feasible integer solution within a 10% approximation of the optimal solution. For Ecuador (1) the termination value was set at 20% and Ecuador (10%) was too large to be solved using XPRESS.