AN INTEGER OPTIMIZATION APPROACH TO A PROBABILISTIC RESERVE SITE SELECTION PROBLEM

ROBERT G. HAIGHT

U.S.D.A. Forest Service, North Central Research Station, 1992 Folwell Avenue, St. Paul, Minnesota 55108, rhaight.fs.fed.us

CHARLES S. REVELLE

Johns Hopkins University, Department of Geography and Environmental Engineering, 313 Ames Hall, Baltimore, Maryland 21218, revelle@jhu.edu

STEPHANIE A. SNYDER

Minnesota Department of Transportation, 108 Cecil Street SE, Minneapolis, MN 55919, snyder01@yahoo.com

(Received June 1998; revisions received February 1999, June 1999; accepted June 1999)

Interest in protecting natural areas is increasing as development pressures and conflicting land uses threaten and fragment ecosystems. A variety of quantitative approaches have been developed to help managers select sites for biodiversity protection. The problem is often formulated to select the set of reserve sites that maximizes the number of species or ecological communities that are represented, subject to an upper bound on the number or area of selected sites. Most formulations assume that information about the presence or absence of species in the candidate sites is known with certainty. Because complete information typically is lacking, we developed a reserve selection formulation that incorporates probabilistic presence-absence data. The formulation was a discrete 0/1 optimization model that maximized the number of represented vegetation communities subject to a budget constraint, where a community was considered represented if its probability of occurrence in the set of selected sites exceeded a specified minimum reliability threshold. Although the formulation was nonlinear, a log transformation allowed us to represent the problem in a linear format that could be solved using exact optimization methods. The formulation was tested using a moderately sized reserve selection problem based on data from the Superior National Forest in Minnesota.

Human economic and agricultural activities contribute to the endangerment of over 900 species that are currently listed or proposed for listing under the federal Endangered Species Act in the United States (Dobson et al. 1997). One approach to conserving the elements of biological diversity—including plants, animals, and ecological communities—is to establish and enhance biological reserves in which economic development is curtailed (Ando et al. 1998). For example, between 1927 and 1998, the U.S. Forest Service established 427 research natural areas (RNAs) covering over 500,000 acres of land in national forests of the United States. These biological reserves are permanently protected and maintained in their natural condition for the purposes of conserving species and ecosystems, conducting nonmanipulative research and monitoring, and fostering education. Examples of land-use planning for the protection of biodiversity abound, including cases in Australia, South Africa, and Norway (Pressey et al. 1997). The establishment and enhancement of biological reserves is viewed as the cornerstone of biodiversity conservation throughout the world (Noss and Cooperrider 1994, Pimm and Lawton 1998).

Given the reality that protected reserve status may prohibit other land and resource activities on a site, it may not always be financially possible to protect all the species or ecosystems in a region. This issue is of particular importance when dealing with public lands in which multiple and conflicting resource demands are the norm. Difficult decisions must often be made that recognize and evaluate the tradeoffs between biodiversity protection goals and alternative land and resource uses. A haphazard selection of protected reserves such as RNAs may do little to contribute to biodiversity protection goals. Automated, quantitative methods that can efficiently and effectively identify sets of sites for reserve status could enhance the decision maker's or land manager's ability to make sound decisions regarding levels of reserve protection.

A number of quantitative methods have been developed over the past 15 years to address the reserve selection problem, as it is commonly referred to in the biological conservation literature. A common approach is to select the set of sites that maximizes the number of species that are represented by the reserve network, where a species is considered represented if at least one site with a known, viable population of the species is protected. Most of those models assume that the presence and absence of all the species in the candidate reserve sites are known with certainty. In practice, however, predictions of the presence and absence of species can be

Subject classifications: Environment: protected natural reserves on national forests. Programming, integer: discrete 0/1 by exact solution methods. Probability, applications: presence-absence data of species in reserves.
Area of review: ENVIRONMENT, ENERGY AND NATURAL RESOURCES.

erroneous. For example, species assumed to be absent because they were not encountered in partial surveys of sites might actually be present. On the other hand, species assumed to be present because of their association with known communities might actually be absent. Ecologists are beginning to quantify errors in predictions of species' occurrence (Flather et al. 1997) and to use those errors to estimate probabilities of occurrence (Dean et al. 1997). Information on the likelihood of species presence and absence should not be ignored in the development of reserve selection models.

We formulated a probabilistic reserve selection model that allowed the presence and absence of vegetation communities within potential reserve sites to be represented by probabilistic, rather than deterministic values, the COMPRES model (Covering Model for Probabilistic Reserve Selection). We focused on vegetation communities rather than species because protecting examples of a wide array of vegetation communities should conserve most species, biotic interactions, and ecological processes (Faber-Langendoen 1996). The model can also be applied to representation of species or other elements of biodiversity. The model was formulated as a 0/1 integer optimization problem that maximized the number of represented communities subject to a budget constraint, where a community was considered represented if its probability of occurrence in the set of selected reserve sites was greater than a specified minimum reliability threshold. This threshold represented the degree of risk aversion on the part of the decision maker. Although our formulation was nonlinear, a logarithmic transformation allowed us to represent the problem in an analogous, linear format that could be solved using exact optimization methods. The approach was illustrated with a research natural area selection problem on the Superior National Forest in Minnesota.

1. LITERATURE REVIEW

The reserve selection problem has been examined by researchers in a number of disciplines. Although a variety of ecological or biological protection goals can be specified (Pressey et al. 1993), two reserve selection problems are commonly addressed in the conservation biology literature: maximize the number of species that can be represented within a specified number of sites, or identify the smallest number of sites sufficient to represent all the species of concern.

The first quantitative methods developed to solve the reserve selection problem were straightforward scoring and ranking procedures based upon metrics such as reserve size or species richness (e.g., Kirkpatrick 1983, Margules et al. 1988, Cocks and Baird 1989). Sites are selected for protection in a sequential fashion, based upon score, until a resource constraint, such as cost or area, is reached (e.g., Margules and Usher 1981, Terborgh and Winter 1983, Pressey and Nicholls 1989). A significant drawback of this approach is that sites are scored and selected independently of the composition of previously ranked sites. As a result, strictly choosing the highest ranked sites may lead to solutions that are ineffectual or inefficient. An advantage of this approach to reserve selection, however, is the ability to rapidly generate feasible solutions to what could be very large and complex problems.

A second approach to solving reserve selection problems involves the use of greedy-adding heuristics (Margules et al. 1988, Rebelo and Siegfried 1990, Vane-Wright et al. 1991, Bedward et al. 1992, Nicholls and Margules 1993, Pressey et al. 1993). Like the scoring and ranking methods, greedy heuristics identify a prioritized sequence of sites for reserve status. The first step is to select the best site in terms of the principal selection criterion (e.g., species richness). Next, the value of each remaining site is calculated, accounting for the species already represented. The site that best supplements the species represented in previously selected sites is added to the solution. This re-calculation and selection continues until an appropriate resource constraint (e.g., total cost) or a stopping rule (e.g., all species of concern are represented) is met. In contrast to scoring and ranking methods, greedy heuristics avoid redundancy or omissions of representation by accounting for species represented in previously selected sites and the species still in need of representation. The principal drawback of greedy heuristics is that they do not guarantee optimal solutions (e.g., finding the maximum number of species that can be represented by a specified number of sites or finding the smallest number of sites sufficient to represent all of the species or ecosystems of concern), and there is no way to determine the degree of suboptimality (Underhill 1994). Work is continuing in the development of more sophisticated heuristics, including simulated annealing and genetic algorithms (Pressey et al. 1996), that may provide a better approximation of the optimal solution. As with the scoring methods, greedy-adding heuristics have the advantage of being able to rapidly generate feasible solutions.

A third approach involves integer optimization models that can be solved to optimality using conventional linear programming and branch and bound algorithms (Cocks and Baird 1989; Saetersdal et al. 1993; Camm et al. 1996; Church et al. 1996; Davis and Stoms 1996; Willis et al. 1996; Williams and ReVelle 1996, 1997, 1998; Snyder et al. 1999). Church et al. (1996) pointed out that the two reserve selection problems commonly addressed in the conservation biology literature are applications or modifications of two classic formulations from the location science literature: the maximal covering location problem (Church and ReVelle 1974), which maximizes the number of entities or amount of demand that could be covered or represented by a specified number of facilities and the location set covering problem (Toregas and ReVelle 1973), which minimizes the number of facilities necessary to cover or represent all demand nodes. Both of these formulations are amenable to integer optimization, an approach which guarantees optimal mathematical solutions. Optimization differs from scoring and heuristics approaches by identifying and evaluating entire sets of sites according to the selection criteria, rather than sequentially selecting sites based on the characteristics

of the sites chosen in previous iterations. Furthermore, in contrast to scoring and heuristic approaches, the solutions derived from optimization models are in no way dependent on or sensitive to starting conditions or order of site selection. However, this approach to reserve selection is not without its drawbacks. Integer optimization formulations can be difficult to solve to optimality, proving intractable in some cases for moderately sized reserve selection formulations (Pressey et al. 1996).

Through the progression of the development of the solution techniques outlined above, the reserve selection problem has been approached and solved with greater degrees of solution accuracy and efficiency. A next logical step in the development of more realistic reserve selection formulations is to address the issue of incomplete and uncertain data. Polasky et al. (2000) were the first to address the issue of a probabilistic reserve site selection formulation. They developed a model to maximize the expected number of species represented in a reserve network when the presence of species at potential reserve sites was represented as a probability, rather than a known value of 1 (present) or 0 (absent). A greedy-adding heuristic and a variant of the greedy-adding heuristic, e.g., an "expected greedy algorithm," were developed and utilized to solve the problem. Solutions derived from this probabilistic formulation were compared to solutions from a deterministic formulation when the probabilistic data were transformed into presence-absence data. That is, all probabilities greater than or equal to a specified value (e.g., 0.6) were set to a value of 1.0, and all probabilities less than this were set to a value of 0.0. As one would assume, the authors found that transforming the probabilistic data into presence-absence data and solving the respective reserve selection formulations changed the set of sites that would be chosen and the expected number of species represented. The authors also found that the expected number of represented species was lower initially than the number represented in a reserve network using transformed presence-absence data. However, the expected value of an additional site was generally higher than it was with presence-absence data. The model suggested that there is value in selecting multiple sites in which a species has some possibility of being present to ensure some chance that the species is actually represented in the network. Building upon the work of Polasky et al. (2000), as well as the many deterministic reserve selection formulations, we developed a specification of the probabilistic reserve selection problem that takes into account the risk tolerance of the decision maker. Furthermore, our 0/1 optimization formulation can be solved using exact optimization methods guaranteeing optimal solutions.

2. MODEL DESCRIPTION

A 0-1 integer optimization model was formulated to select the network of sites that maximized the number of vegetation communities represented, subject to an upper bound on the total area of the reserve network. Again, a community was considered represented if its probability of occurrence in the selected set of sites was at least as large as the specified minimum reliability level (e.g., 95%). The following notation was utilized in the model:

- i, I index and set of communities
- j, J index and set of potential reserve sites
- T upper bound on reserve network area
- A_j area of site j
- P_{ij} probability that community i is present in site j
- α_i threshold reliability level for community i
- N_i set of sites that may contain (have a nonzero probability) community i
- X_j {1/0 variable; 1 if site j is selected for inclusion in the reserve network, and 0 otherwise}
- Y_i {1/0 variable; 1 if the probability that community i is represented by the selected set of sites is at least α_i, and 0 otherwise}

The model was formulated as follows:

$$\text{Maximize} \quad Z = \sum_{i \in I} Y_i, \tag{1}$$

subject to:

$$\sum_{j \in J} A_j X_j \leq T, \tag{2}$$

$$\prod_{j \in N_i} (1-P_{ij})^{X_j} \leq (1-\alpha_i)^{Y_i} \qquad \forall i \in I, \tag{3}$$

$$X_j, Y_i \in \{0, 1\} \qquad \forall i \in I, \ \forall j \in J. \tag{4}$$

The objective (1) maximized the number of communities whose probability of occurrence, based on the selected set of sites, exceeded the specified reliability level. The first constraint (2) ensured that the total area of the selected set of sites did not exceed T, the upper bound on network area. The second set of constraints (3) defined the conditions under which communities were considered represented. This constraint stipulated that for any community i to be considered represented, the probability of its *absence* $\prod_{j \in N_i}(1-P_{ij})^{X_j}$ from the selected set of sites had to be less than the specified risk threshold level of absence $(1-\alpha_i)$. Thus, if the specified threshold reliability level for a community's presence was 95%, the probability that the community was *not* present in the selected set of sites had to be no greater than 5%. If $\prod_{j \in N_i}(1-P_{ij})^{X_j} > (1-\alpha_i)$, then the corresponding Y_i in Equation (3) had to equal zero, indicating that the selected sites did not represent community i with the required probability. If $\prod_{j \in N_i}(1-P_{ij})^{X_j} \leq (1-\alpha_i)$, then $Y_i = 1$, indicating that community i was represented with the required probability. The last set of constraints (4) defined the integer requirements for the decision variables.

Restating the nonlinear constraint in linear form facilitated a solution with exact optimization methods. This was accomplished by transforming Equation (3) to an equivalent linear set of equations with a log transform:

$$\sum_{j \in N_i} X_j \ln(1-P_{ij}) \leq Y_i \ln(1-\alpha_i) \qquad \forall i \in I. \tag{5}$$

Because we were dealing with probabilities, the logarithms in both sides of Equation (5) were always negative values. This allowed us to multiply through by (-1) and switch the direction of the inequality:

$$Y_i \leqslant \frac{\sum_{j \in N_i} X_j \ln(1-P_{ij})}{\ln(1-\alpha_i)} \qquad \forall i \in I. \tag{6}$$

If the probability of absence of a community i was greater than the specified absence threshold $(1 - \alpha_i)$, then the quotient in constraint (6) would be less than one, forcing the corresponding Y_i variable to be equal to zero, due to the integrality restrictions on the variable. Solving the reserve selection problem with Equation (6), the linear equivalent of Equation (3), enforced the same condition as that specified in (3). It is important to note that the manner in which constraint (6) was defined allowed a linear representation of what was inherently a nonlinear relationship, and as such, allowed this formulation to be solved as a linear, integer formulation with conventional solution methods, such as the simplex method in conjunction with the branch-and-bound algorithm. Without this transformation, this probabilistic formulation would have required heuristic solution methods, which could not guarantee optimal solutions.

The form of our probabilistic representation constraint (3) evolved from an idea in ReVelle and Hogan (1988, 1989) for the Maximum Availability Location problem. Those authors suggested that the probability of a vehicle being available to serve a demand region within a given time standard could be computed as one minus the product of the busy fractions of vehicles positioned in the region, where each busy fraction had a 0-1 exponent representing the vehicle placement variable. The left-hand side of our representation constraints (3), which computes the probability that a community is absent from the selected sites, is analogous to the computation of the probability of vehicle availability. Our contribution was the formulation of the right-hand side of constraint (3), in which the representation variable Y could equal one only if the probability of community absence was less than the required threshold. This constraint structure has not been utilized in the literature of reserve site selection, nor to our knowledge has it been utilized in the more general location literature.

To demonstrate the model, we generated trade-off curves comparing the maximum number of communities represented given a limited total network acreage and a specified threshold reliability level. By varying the threshold reliability level and re-solving the problem for different levels of T, the "costs," in terms of the number of communities considered represented, were evaluated for different levels of risk aversion.

Ensuring Representation of Priority Communities

In the base model above, the objective was to maximize the number of communities considered represented without requiring representation of any given communities. Situations may arise, however, when a manager may want to ensure that certain, perhaps priority or threatened, communities are represented in the set of selected sites at some specified reliability level. To enforce this condition, the following constraint set was added to the base formulation:

$$1 - \left(\prod_{j \in N_i} (1 - P_{ij})^{X_j} \right) \geqslant \beta_i \qquad \forall i \in M, \tag{7}$$

where M is the set of identified priority communities, and β_i is the minimum required reliability for each priority community i. With this constraint set, the probability of presence of each priority community must exceed the specified reliability level β_i. These constraints supplemented the corresponding constraints in Equation (6), rather than replaced them. Again, just as with Equation (3), log transformation and algebraic manipulation were needed to state this set of nonlinear constraints as an analogous set of linear ones. The following, analogous constraint set was created by subtracting 1 from both sides, multiplying through by (-1) and taking the log of both sides:

$$\sum_{j \in N_i} X_j \ln(1 - P_{ij}) \leqslant \ln(1 - \beta_i) \qquad \forall i \in M. \tag{8}$$

In this format, each constraint required the probability of absence of a priority community i to be less than or equal to specified threshold level of absence $(1 - \beta_i)$. Equation (8) was added to the base formulation and the problem re-solved to generate an additional trade-off curve. This modified formulation still maximized the number of communities represented by the minimum reliability level α_i, while ensuring that a set of priority community was represented by a higher reliability level, β_i.

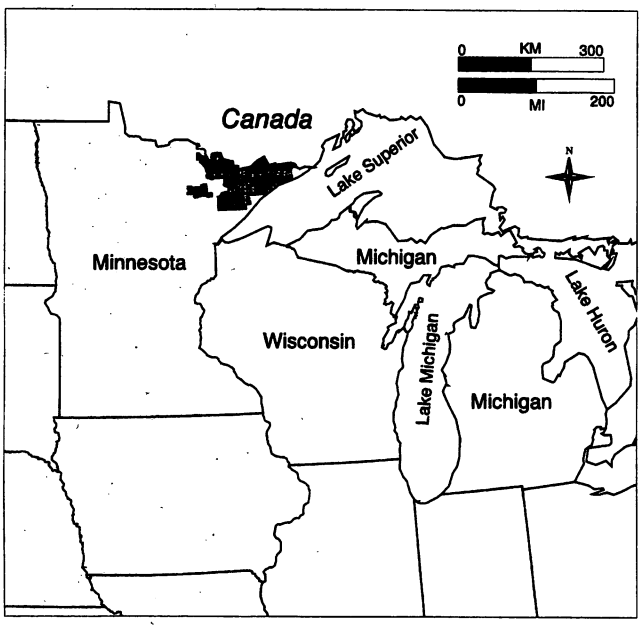
3. APPLICATION

3.1. Problem Setting

We used the model to address a research natural area (RNA) selection problem on the Superior National Forest, which is the largest national forest in the eastern United States and covers over 2.1 million acres in northeastern Minnesota (Figure 1) (USDA 1986). The goal of the RNA selection problem was to select a network of sites for protection that maximized the number of plant community types considered represented by the specified level of reliability.

The analysis was conducted using 33 potential RNAs. The 33 sites were part of a larger set of sites that had been identified using high-altitude aerial photography, as having potential to be high-quality examples of some of the ecosystems found on the Superior National Forest (Vora 1997). A rapid assessment using aerial and ground surveys of these 33 sites was conducted in 1997 to map boundaries and inventory plant communities of the sites (Anderson 1997). The 33 sites were selected from the larger set because site boundaries had been mapped and partial inventories of plant communities had been conducted by the time of our study (Figure 2). The sites ranged in size from 600 to 19,000 acres and covered a total of 126,000 acres (Table 1). The use of

Figure 1. Location of the Superior National Forest (shaded area).



these 33 potential RNAs does not imply that other sites do not merit further study as RNA candidates on the Superior National Forest. Furthermore, the analyses conducted with this data set were done to illustrate a methodological approach and do not imply or endorse any policy implications.

Community types were defined using a combination of land and vegetation classes. Land classes were the subsection level of the National Hierarchy for Ecological Units, a system of mapped land units used in national forest planning (McNab and Avers 1994, Keys et al. 1995). A subsection is a unit of land distinguished by similar climatic regimes, geologic structure, and other physio-graphic characteristics covering thousands of square miles. Five subsections are present in the Superior National Forest, shown by the shaded areas in Figure 2. Vegetation classes were defined using the alliance level of The Nature Conservancy's (TNC) National Vegetation Classification hierarchy (Faber-Langendoen 1996). An alliance is a unit of vegetation distinguished by the plants in the uppermost canopy or layer of vegetation. Twenty-five different alliances were known to be present in one or more of the 33 potential RNAs. Alliance names are listed in the left-hand column of Table 2.

Figure 2. Ecological subsections and potential research natural areas of the Superior National Forest. (Lightly shaded areas with alphanumeric labels represent subsections.)

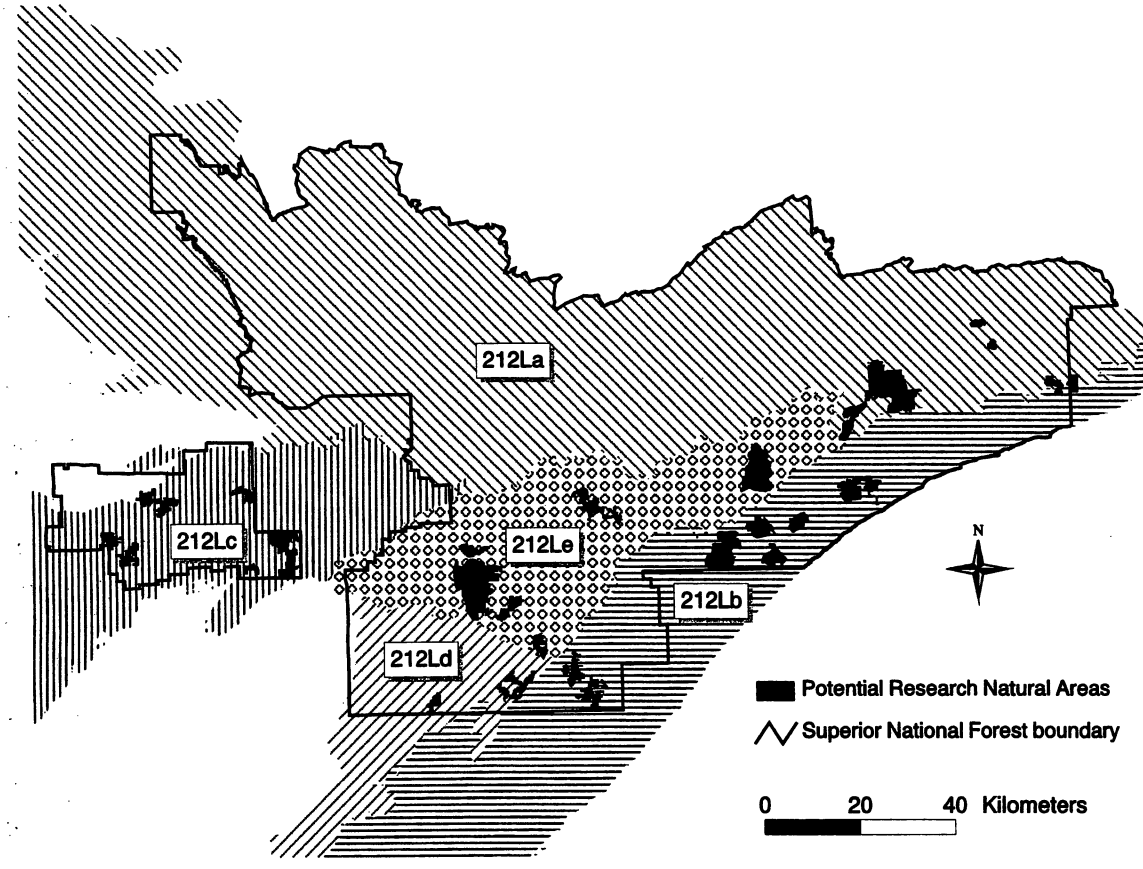


Table 1. Potential research natural areas on the Superior National Forest.

Potential RNA	Subsection	Area (acres)
Cascade Lake	212La	16,956
Locket Lake		791
Lullaby Creek		664
Rice Chain		3,886
Trout Lake		1,657
Barker Lake	212Lb	4,311
Beaver River		1,259
Cabin Creek		9,719
Fall River Fen		1,039
Heartbreak Creek		2,545
Lillian Creek South		3,434
Lutsen Addition		2,061
Pearl Lake		4,040
Schroeder Addition		3,495
South Brule River		1,573
Watertank Lake		2,709
Candle Lake	212Lc	1,872
Deepwater Lake		2,209
Loka Lake		4,079
Pike Mountain		896
Rice Lakes		1,786
Slow Creek		2,804
Watercress Lake		2,594
Wynne Creek		3,196
Sullivan Creek	212Ld	4,032
Wolf Lake		1,470
Big Lake—7 Beavers	212Le	18,578
Dragon Lake		3,778
Dunka		1,947
Little Isabella River		1,098
South Greenwood Creek		2,596
Timber-Frear		10,883
White Pine Picnic		2,124

Combining the 25 alliances with the 5 subsections, we obtained 125 community types in need of representation, assuming each of the alliances could be present within each of the five subsections.

Because of limited resources, only a portion of each of the 33 potential RNA sites was surveyed for plant community types. This partial survey verified the presence of a total of 63 community types across the set of 33 sites. In addition to the plant communities in the portion of each site that was field surveyed, each site could include additional communities that were not encountered. This uncertainty about the presence of community types was the motivation for our model.

To demonstrate the model, we constructed a matrix of probabilities for the presence of community types in sites to supplement the verified community data. With 125 community types and 33 eligible sites, the presence-absence matrix consisted of 4,125 elements. Because each site was eligible only to be located in one subsection, each site could cover a maximum of 25 different community types. Therefore, a maximum of 20% (825) of the elements of the probability matrix could be nonzero. The results of the field survey documenting the presence of community types in each site allowed us to set 17% of these 825 elements equal to one. Because we did not have means to assign the probabilities of the presence of other community types in each site, we arbitrarily set 33% of the 825 elements equal to probabilities drawn from a uniform distribution on the range of 0 to 1. The remaining elements were assumed to be zero. Each site had 8 to 15 nonzero probabilities, of which 6 to 11 were randomly generated. For illustration, partial vectors of probabilities for two sites within the same subsection are shown in Table 2.

Table 2. List of eligible community types and partial vectors of probabilities for two sites within the same subsection.

	RNAs	
Matrix Plant Communities	1	2
Jack Pine Forest Alliance	0.181	0.000
Red Pine Forest Alliance	0.459	0.000
White Pine-(Red Pine)-Quaking Aspen Forest Alliance	0.917	1.000
White Pine Forest Alliance	0.037	0.000
White Spruce-Balsam Fir-Quaking Aspen Forest Alliance	1.000	0.000
Sugar Maple-Yellow Birch-(American Beech) Forest Alliance	1.000	0.626
Quaking Aspen-Paper Birch Forest Alliance	0.897	0.000
Large Patch Plant Communities		
Black Spruce Forest Alliance	0.000	0.706
Eastern White Cedar-Yellow Birch Forest Alliance	0.521	0.454
Eastern White Cedar Forest Alliance	0.836	0.000
Red Oak-Sugar Maple-(White Oak) Forest Alliance	0.000	0.000
Paper Birch Forest Alliance	0.748	0.000
Black Spruce Saturated Forest Alliance	1.000	1.000
Eastern White Cedar Saturated Forest Alliance	1.000	1.000
Black Ash-Red Maple Saturated Forest Alliance	0.000	0.000
Black Spruce Saturated Woodland Alliance	0.000	0.366
Speckled Alder Seasonally Flooded Shrubland Alliance	0.000	1.000
Leatherleaf Saturated Dwarf-Shrubland Alliance	0.000	0.000
Cattail-(Bulrush) Semipermanently Flooded Herbaceous	0.948	0.000
Few-Seeded, Wiregrass Sedge Saturated Herbaceous	0.000	0.824
Rock Outcrop/Butte Sparse Vegetation	0.000	0.000
Small Patch Plant Communities		
Bog Birch—(Willow) Saturated Shrubland Alliance	0.814	0.000
Sedge (C. rostrata, uticulata) Seasonally Flooded Herbaceous	0.000	0.304
Open Bluff/Cliff Sparse Vegetation	0.633	0.829
Cobble/Gravel Shore Sparse Vegetation	0.000	0.000

0.457 is the probability that the community type i is present within site j (p_{ij}).

3.2. Trade-Off Analysis

To analyze the effects of imposing an upper bound on network area, we used the optimization model (Equations (1) through (4)) to compute a trade-off curve with $\alpha_i = 1$ for all i so that representation required certainty of occurrence. The trade-offs between the number of communities considered to be represented with 100% reliability and network area were obtained by decreasing the upper bound on network area, in increments of 1,000 acres from the maximum of 126,000 acres, and re-solving the optimization problem.

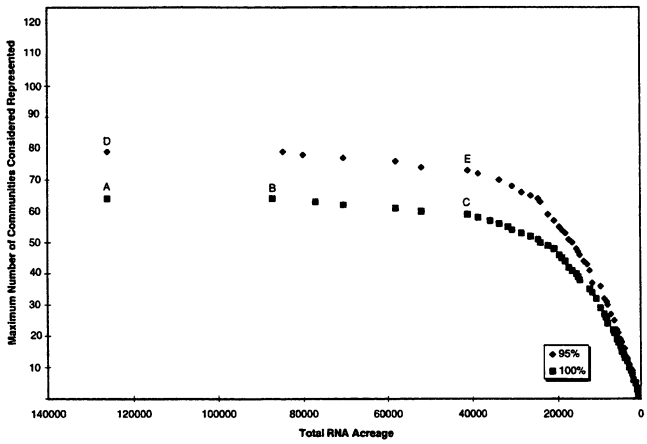
If the decision maker was willing to gamble that more communities could be represented by a different network of a given size, then the optimization model could be solved using a lower threshold reliability level for representation. We examined the effect of reduced risk aversion by lowering the threshold reliability level (applied uniformly across the community types) to 95% and recomputing the trade-off curve.

A separate set of runs was conducted with the extended formulation to evaluate the trade-offs when representation, at a high reliability level, was *required* for certain priority communities. To illustrate this, five alliances in the 212Lb subsection were identified as priority communities and were required to have representation with at least the 95% reliability level ($\beta_i = 0.95$), while the remaining community types required an 80% level of reliability to be considered represented. Note, this approach is different from setting multiple levels of α_i in the base formulation in which a higher α_i simply specifies that you want to be more certain that a community is present before you consider it represented, not that you *must* represent it at a higher reliability level. Again, the value of T, maximum network area, was incrementally varied and the problem re-solved to produce another trade-off curve between network area and the number of communities considered represented with 80% reliability, given that priority communities must be present with a reliability of 95% for representation.

3.3. Software

The model was solved on an IBM300PL personal computer using the integrated solution package GAMS/OSL 2.25 (GAMS Development Corporation 1990), which was designed for large and complex linear and mixed integer programming problems. Input files were created using GAMS (General Algebraic Modeling System), a program designed to generate data files in a format that standard optimization packages can read and process. The model was solved using the Primal-Dual Predictor-Corrector Barrier Interior point algorithm as the LP solver, in conjunction with the branch-and-bound algorithm for integer-variable problems. Both of these solution algorithms were part of IBM's OSL (Optimization Subroutine Library), a FORTRAN-based subroutine library designed to solve optimization problems. The interior point solution algorithm was chosen because it proved considerably more efficient, in terms of solution time, than the primal revised simplex in conjunction with the branch-and-bound algorithm in preliminary trials.

Figure 3. Trade-off curves for differing levels of minimum reliability.



4. RESULTS

4.1. Trade-Offs Between Community Representation and Network Area

The bottom curve in Figure 3 shows the trade-offs between maximum number of communities considered represented with certainty for decreasing upper bounds on network area. Each point on the curve represents a unique network of RNAs. The flat portion of the curve between points A and B shows that a maximum of 64 of the 125 communities (51%) could be represented with certainty. Further, these 64 communities could be represented by different networks covering a wide range of areas. If all 33 potential RNAs were selected, the protected network would cover 126,000 acres (Point A, Figure 3). Because many communities were found in more than one site, the model was able to select a smaller set of sites without reducing representation. For example, Point B (Figure 3) represented all 64 communities with 21 sites, covering 87,000 acres—a 31% decrease in network area. Additional networks can be found between points A and B on the trade-off curve, all providing representation of 64 communities.

We found that the upper bound on network area could be decreased without great reductions in community representation. For example, with an upper bound of 41,000 acres—a 67% reduction in network area from the maximum of 126,000 acres—59 communities were considered represented with certainty (Point C, Figure 3), an 8% reduction in representation from the maximum of 64 communities. With an upper bound of 33,000 acres—a 74% reduction in network area from Point A (Figure 3)—87.5% of the communities were still represented with certainty. The decline in the number of communities represented was more pronounced as the upper bound on network area dropped below 30,000 acres.

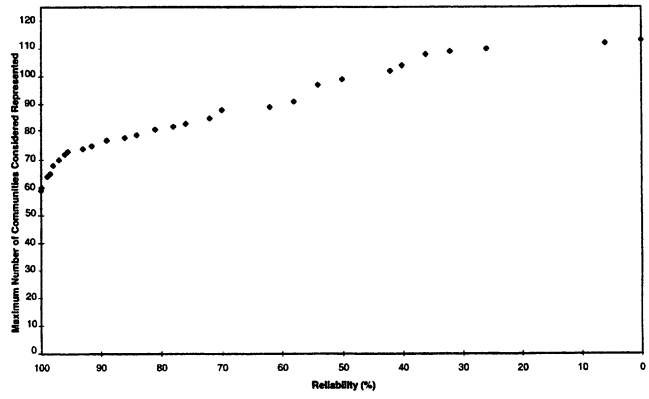
Lowering the threshold reliability level to 95% increased the number of communities that were considered represented under each area upper bound (top curve, Figure 3). A maximum of 79 of the 125 communities (63%) could be represented with at least 95% reliability (Point D, Figure 3). Community representation remained high for substantial reductions in the upper bound on network area. As an example, when this upper bound was 41,000 acres, a 67% reduction network area from the maximum of 126,000 acres (Point E, Figure 3), 73 communities were represented with 95% reliability—an 8% reduction in representation from the maximum of 79 communities.

Comparing the optimal solutions for a given upper bound on network area highlights the trade-offs associated with the different reliability thresholds. With an area upper bound of 41,000 acres, the optimal solution associated with the 100% reliability threshold (Point C, Figure 3) was a network of 18 sites that included 59 communities represented with certainty, and 9 communities represented with reliability between 95% and 99%. Although these latter nine communities were present with relatively high reliability, they were not factored into the objective function, nor did they influence the selection of the corresponding reserve sites because they did not meet the specified reliability threshold. The optimal solution associated with the 95% reliability threshold (Point E, Figure 3) was a different network of 18 sites that included 55 communities represented with certainty, and 18 communities represented with reliability between 95% and 99%. Thus, an additional trade-off to consider is whether a network that represents more communities with lower reliability is worth more to the decision maker than a network that represents fewer communities with higher reliability.

We emphasize that the objective function counts only those communities that meet or exceed the threshold reliability level, and it does not weigh communities by how much they exceed or fall short of the threshold. Thus, communities that are counted in a given solution may be represented with reliabilities that are much greater than the reliability threshold. Further, the quality of a given solution is unaffected by how close the probabilities of representation are to the threshold.

To investigate the sensitivity of optimal solutions to the threshold level of reliability, we solved the optimization problem with incrementally smaller reliability thresholds while holding the upper bound on network area constant at 41,000 acres. The number of plant communities considered to be represented increased with decreasing reliability thresholds (Figure 4). Furthermore, the rate of increase in the number of represented communities was highest for reliabilities in the range of 100% to 95%. The implication was that for any given level of reliability in this range, several communities were represented at a reliability slightly below the threshold. For a very small decrease in reliability in this range (e.g., 0.5%), the marginal gains that could be realized in terms of the number of communities expected to be represented were large. In this situation, the decision maker would be wise to evaluate the solution networks derived from the optimization model at several reliability thresholds in this range. Backing off slightly from the highest reliability scenario might allow the decision maker to identify and select networks that may contain a greater number of communities considered to be represented or to better meet additional, unmodeled criteria (e.g., political factors).

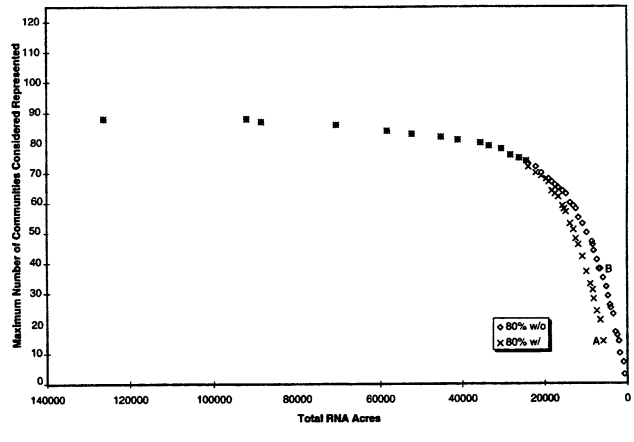
Figure 4. Maximum number of communities represented under different levels of reliability when total network area was 41,000 acres.



4.2. Ensuring Representation for Priority Community Types

Figure 5 contains two trade-off curves that illustrate the impacts of requiring representation of priority communities. The top curve shows community representation as a function of the upper bound on network area, using an 80% reliability threshold and assuming that no priority communities were specified. The bottom curve was derived from the modified formulation in which five priority communities were required to be represented with at least 95% reliability.

Figure 5. Comparison of trade-off curves with and without required minimum reliability representation of priority communities.



A comparison of these two trade-off curves allowed the impacts associated with priority representation at one level of reliability to be assessed in terms of the total number of communities that could be represented with a lower level of reliability. Priority community constraints influence the solution only when the upper bound on total network area is less than 24,000 acres. Above this upper bound, the solutions were identical, indicating that the five priority communities would be represented with at least 95% reliability without the explicit requirement. Below this acreage threshold, trade-offs were significant, and careful thought should be given to the importance of representing these priority communities. For example, Point A (Figure 5) represented a network of 5800 acres, the smallest network sufficient to represent at least the five priority communities at the 95% reliability level. This network represented a total of 14 communities: the 5 with at least 95% reliability and the remaining 9 with at least 80% reliability. An alternate network with this same acreage (Point B, Figure 5) represented 35 total communities with at least 80% reliability, but it did not enforce representation of the five priority communities. Therefore, to be certain that the five communities were represented with a high reliability, the decision maker must reduce the total number of communities that can be represented with at least 80% reliability.

5. RESERVE SELECTION MODELS IN PRACTICE

To date, reserve selection models utilized in practice assume that information about species or community presence in potential reserves is known with certainty. For example, in consultation with U.S. Forest Service planners, we used a deterministic version of our model to address three basic questions (Snyder et al. 1999):

1. What is the minimum area of RNAs required to represent all the plant communities known to exist in the Superior National Forest?
2. How sensitive is the siting of RNAs to the way in which plant communities are defined?
3. How do requirements to protect priority plant communities affect RNA siting?

In addition, we provided cost curves showing the trade-offs between the maximum number of plant communities represented and total area of the selected set of RNAs, and we estimated the costs of various constraints on the number, size, and location of RNAs. The planners used this information to propose alternative sets of RNAs to include in comprehensive land use plans. Similar published accounts of the application of reserve selection models in land use planning include the construction of potential reserve systems in the Sierra Nevada region of California (Davis et al. 1996) and the Agulhas Plain in South Africa (Lombard et al. 1997).

Reserve selection models utilize various kinds of data, depending on model specifications and objectives. When the objective involves plant community representation, information about community presence is obtained from site surveys and maps of vegetation and land-use patterns (e.g., Lombard et al. 1997, Snyder et al. 1999). For objectives involving species representation, information about species presence is obtained from species distribution maps; when basic distribution data are lacking, predictions are made based on habitat associations (e.g., Flather et al. 1997).

Most reserve selection models ignore errors in the input data. For example, our deterministic model (Snyder et al. 1999) failed to recognize that surveys of plant communities in many of the potential RNAs were incomplete. As a result, communities that we assumed to be absent because they were not encountered might actually be present (errors of omission). In other cases, errors occur when species or communities are assumed to be present based on distribution maps but are actually absent (errors of commission). Accuracy assessments of species distribution maps show that error rates vary widely among species and reach 50% of the numbers of species predicted or observed (Flather et al. 1997).

Our COMPRES model is among the first to account for errors in reserve selection input data, and our application demonstrated that RNA siting can be sensitive to errors of omission. For example, if the decision maker was willing to accept a lower level of reliability of community representation, then a set of reserves could be selected that potentially include a larger number of communities for a given upper bound on total RNA area. This information is important because it quantifies the impacts of different levels of risk assumed by the decision maker. Furthermore, the results could help identify where to focus additional site surveys.

While our COMPRES model accounts for errors in the input data, its application requires estimation of the probabilities, or subjective probabilities, of species and community presence. (Refer to Von Winterfeldt and W. Edwards 1986 for a general description of elicitation of subjective probabilities.) One approach to eliciting subjective probabilities is to obtain expert opinions about the likelihoods of community presence based on aerial photography, physical characteristics of the RNAs such as soil type and moisture regime, and information about the co-occurrence of community types and rareness. For example, community presence in a given RNA could be categorized as either likely, unlikely, or not at all likely. These responses could then be translated into subjective probabilities such as 0.6 for likely, 0.2 for unlikely, and 0.0 for or not at all likely. Another approach is to obtain a plant community distribution map and calculate errors of omission and commission based on ground surveys. The map and error information could then be translated into probabilities of community occurrence. We are pursuing both of these approaches in cooperation with U.S. Forest Service planners.

Our application of the COMPRES model suggests that problems with up to 33 sites can be readily solved using exact solution methods. This scale of analysis is typical of RNA site selection problems in national forests. For our data set, average solution times were 4.65 and 12.8 minutes

for the 80% and 95% reliability curves, respectively. Average solution time for the modified formulation, requiring priority representation, was 2.2 minutes—less than half the average solution time for the corresponding base formulation without required representation. Run times of these lengths would allow the COMPRES model to be used as a real-time decision tool. Although our application did solve well computationally, it is possible that larger data sets could require prohibitively long solution times. As problem size increases, either in terms of the number of communities in need of representation or of the number of communities for which a probabilistic presence value is needed, solution times may become larger than practical. Integer programs are, in general, a very difficult class of problems to solve to optimality, and the threshold constraint utilized in this model would not be considered "integer friendly," or likely to produce integer-valued variables without a certain amount of branch-and-bound (ReVelle 1993). Thus, our formulation may not be tractable for exact optimization methods when applied to larger problems. In that case, a heuristic would be required, and our model could be used to assess the accuracy of a heuristic.

Decision makers involved in nature reserve selection can be influenced by ecological considerations that we did not consider. For example, in our application, we assumed that a community was adequately represented if its probability of occurrence met a specified reliability level, regardless of the area, quality, or health of the represented community. A community could satisfy the reliability constraint while occupying only a small amount of land that may not, in fact, be adequate to maintain a viable community. In reality, constraints may need to require a minimum quality level or amount of acreage—and further, contiguous acreage—before a community was considered represented. Although progress has been made in the formulation and solution of deterministic reserve selection problems that provide buffer zones around interior or core areas and promote contiguity and compactness (Williams and ReVelle 1996, 1998), more work is needed to account for errors in input data in problems with spatial constraints. Other considerations in RNA selection include the quality and successional status of plant communities, disturbances such as fires or storms that alter the structure and composition of plant communities, and climatic changes that might affect the demography of protected plants and animals. Whether or not these concerns can be adequately incorporated in reserve selection models is an open question. Nevertheless, reserve selection models as currently designed address practical questions, and results are being used in conjunction with other ecological and political considerations to advise decision makers.

6. CONCLUSION

We addressed the problem of incorporating probabilistic data into a reserve selection optimization model. Most reserve selection models formulated to date require that the presence-absence of species or communities within candidate reserve sites be known with certainty. However, in real-world planning situations, complete information is rare and is often prohibitively expensive to obtain. The ability to consider and include probabilistic data adds a much needed element of realism to reserve selection planning.

We developed a 0/1 integer optimization model that incorporated probabilistic data and could be solved using exact optimization methods. Although the formulation was nonlinear, we were able to convert it to an analogous linear statement using a log transformation. This transformation allowed us to solve the model using conventional optimization software and obtain optimal solutions. We demonstrated the model by generating trade-off curves showing the impacts of changing both the maximum network area and minimum threshold reliability level on the maximum number of communities that could be considered represented. These trade-offs provide the decision maker with valuable information on the impacts associated with different bounds on total RNA area and different levels of risk aversion.

While the obvious application of this model is to natural reserve site selection problems when species information is uncertain, the model may have applicability or extensions to other probabilistic problem settings. One such area could be in the siting of detection devices, under budget limitations, to maximize the number of "violations" or occurrences detected. One example might be siting groundwater pollution detection devices when one has probabilistic information on where plumes might be located or the direction in which they are moving. Additionally, one could use this modeling approach to locate or dispatch police squad cars to maximize the number of incidents that are detected or encountered, utilizing historical data on crime rates to determine probabilities of crimes occurring in certain neighborhoods.

Providing land managers with quantitative decision tools to gain insight into complex resource allocation problems can enhance their ability to make informed and effective decisions. The selection of natural areas for protected status is rarely a simple matter, but rather is full of conflict, compromise, and trade-offs. In this sense, optimization models offer a useful and powerful approach to such problems through their ability to generate and assess trade-offs and to determine effects of policy parameters. The ability to account for incomplete information in reserve selection is a step toward addressing a more realistic planning situation.

ACKNOWLEDGMENT

This research is supported by the North Central Research Station. The authors thank Chel Anderson and Carmen Converse of the Minnesota Department of Natural Resources for providing field survey information for potential candidate Research Natural Areas in the Superior National Forest, Mark Nelson of the North Central Research Station for assisting with site mapping, and Kristin Snow of The Nature Conservancy for providing crosswalks between natural community classifications.

They also gratefully acknowledge the insight and assistance that Lucy Tyrrell, Regional RNA Coordinator, North Central Research Station, provided for this research.

REFERENCES

Anderson, C. E. 1997. Evaluation of Selected Potential Candidate Research Natural Areas as Representative of Ecological Landtype Associations on the Superior National Forest. Minnesota Department of Natural Resources, Division of Fish and Wildlife, Biological Report No. 58 (plus maps and appendixes).

Ando, A., J. D. Camm, S. Polasky, A. R. Solow. 1998. Species distributions, land values, and efficient conservation. *Science* **279** 2126–2128.

Bedward, M., R. L. Pressey, D. A. Keith. 1992. A new approach for selecting fully representative reserve networks: Addressing efficiency, reserve design, and land suitability with an iterative analysis. *Biol. Conservation* **62** 115–125.

Camm, J. D., S. Polasky, A. R. Solow, B. Csuti. 1996. A note on optimal algorithms for reserve site selection. *Biol. Conservation* **78** 353–355.

Church, R. L., C. S. ReVelle. 1974. The maximal covering location problem. *Papers in Regional Sci.* **32** 101–118.

——, D. M. Stoms, F. W. Davis. 1996. Reserve selection as a maximal covering location problem. *Biol. Conservation* **76** 105–112.

Cocks, K. D., I. A. Baird. 1989. Using mathematical programming to address the multiple reserve selection problem: An example from the Eyre Peninsula, South Australia. *Biol. Conservation* **49** 113–130.

Davis, F. W., D. Stoms. 1996. A spatial analytical hierarchy for Gap Analysis. J. M. Scott, T. H. Tear, F. Davis, eds. *Gap Analysis: A Landscape Approach to Biodiversity Planning.* American Society for Photogrammetry and Remote Sensing, Bethesda, MD, 15–24.

——, ——, R. L. Church, W. J. Okin, K. N. Johnson. 1996. Selecting biodiversity management areas. In *Sierra Nevada Ecosystem Project: Final Report to Congress.* **1**(58) University of California at Davis, Centers for Water and Wildland Resources.

Dean, D. J., K. R. Wilson, C. H. Flather. 1997. Spatial error analysis of species richness for a gap analysis map. *Photogrammatic Engrg. Remote Sensing* **63** 1211–1217.

Dobson, A. P., J. P. Rodriguez, W. M. Roberts, D. S. Wilcove. 1997. Geographic distribution of endangered species in the United States. *Science* **275** 550–553.

Faber-Langendoen, D., ed. 1996. State Natural Heritage Program Ecologists. Terrestrial Vegetation of the Midwestern United States. In *International Classification of Ecological Communities: Terrestrial Vegetation of the United States.* The Nature Conservancy, Arlington, VA.

Flather, C. H., K. R. Wilson, D. J. Dean, W. C. McComb. 1997. Identifying gaps in conservation networks: of indicators and uncertainty in geographic-based analyses. *Ecological Appl.* **7** 531–542.

GAMS Development Corporation. 1990. *General Algebraic Modeling System.* Version 2.25.090. Washington, DC.

Keys, J., Jr., C. Carpenter, S. Hooks, F. Koenig, W. H. McNab, W. E. Russell, M. L. Smith. 1995. Ecological units of the eastern United States—first approximation (map and booklet of map unit tables). Atlanta, GA. (USDA, Forest Service, presentation scale 1:3,500,000; colored.)

Kirkpatrick, J. B. 1983. An iterative method for establishing priorities for the selection of nature reserves: An example from Tasmania. *Biol. Conservation* **25** 127–134.

Lombard, A. T., R. M. Cowling, R. L. Pressey, P. J. Mustart. 1997. Reserve selection in a species-rich and fragmented landscape on the Agulhas Plain, South Africa. *Conservation Biol.* **11** 1101–1116.

Margules, C. R., A. O. Nicholls, R. L. Pressey. 1988. Selecting networks of reserves to maximize biological diversity. *Biol. Conservation* **43** 63–76.

——, M. B. Usher. 1981. Criteria used in assessing wildlife conservation potential: A review. *Biol. Conservation* **21** 79–109.

McNab, W. H., P. E. Avers, comps. 1994. Ecological subregions of the United States: Section descriptions. Administrative Publication WO-WSA-5. U.S. Department of Agriculture, Forest Service, Washington, DC.

Nicholls, A. O., C. R. Margules. 1993. An updated reserve selection algorithm. *Biological Conservation* **64** 165–169.

Noss, R. F., A. Y. Cooperrider. 1994. *Saving Nature's Legacy.* Island Press, Washington, DC.

Pimm, S. L., J. H. Lawton. 1998. Planning for biodiversity. *Science* **279** 2068–2069.

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

Pressey, R. L., C. J. Humphries, C. R. Margules, R. I. Vane-Wright, P. H. Williams. 1993. Beyond opportunism: Key principles for systematic reserve selection. *Trends Ecol. and Evolution* **8** 124–128.

——, A. O. Nicholls. 1989. Efficiency in conservation planning—Scoring versus iterative approaches. *Biol. Conservation* **50** 199–218.

——, H. P. Possingham, J. R. Day. 1997. Effectiveness of alternative heuristic algorithms for identifying indicative minimum requirements for conservation reserves. *Biol. Conservation* **80** 207–219.

——, ——, C. R. Margules. 1996. Optimality in reserve selection algorithms: When does it matter and how much? *Biol. Conservation* **76** 259–267.

Rebelo, A. G., W. R. Siegfried. 1990. Protection of fynbos vegetation: Ideal and real world options. *Biol. Conservation* **54** 15–31.

ReVelle, C. S. 1993. Facility siting and integer friendly programming. *Eur. J. Oper. Res.* **65**(2) 147–148.

——, K. Hogan. 1988. A reliability constrained siting model with local estimates of busy fractions. *Environ. Planning* **B15** 143–152.

——, ——. 1989. The maximum availability location problem. *Trans. Sci.* **23** 192–200.

Saetersdal, M., J. M. Line, H. J. B. Birks. 1993. How to maximize biological diversity in nature reserve selection: Vascular plants and breeding birds in deciduous woodlands, western Norway. *Biol. Conservation* **66** 131–138.

Snyder, S. A., L. E. Tyrrell, R. G. Haight. 1999. An optimization approach to selecting research natural areas in national forests. *Forest Sci.* **45**(3) 458–469.

Terborgh, J. W., B. Winter. 1983. A method for siting parks and reserves with special reference to Columbia and Ecuador. *Biol. Conservation* **27** 45–58.

Toregas, C., C. S. ReVelle. 1973. Binary logic solutions to a class of location problems. *Geographical Anal.* **5** 145–155.

Underhill, L. G. 1994. Optimal and suboptimal reserve selection algorithms. *Biol. Conservation* **70** 85–87.

USDA Forest Service. 1986. Superior National Forest Land and Resource Management Plan. Eastern Region, USDA Forest Service.

Vane-Wright, R. I., C. J. Humphries, P. H. Williams. 1991. What to protect?—Systematics and the agony of choice. *Biol. Conservation* **55** 235–254.

Von Winterfeldt, D., W. Edwards. 1986. *Decision Analysis and Behavioral Research.* Cambridge University Press, Cambridge, UK.

Vora, R. S. 1997. Identification of potential natural areas, including representative ecosystems, on the Superior National Forest. Unpublished report on file with the USDA Forest Service, North Central Forest Experiment Station, St. Paul, MN.

Williams, J. C., C. S. ReVelle. 1996. A 0-1 programming approach to delineating protected reserves. *Environ. Planning* **B23** 607–624.

——, ——. 1997. Applying mathematical programming to reserve selection. *Environ. Model. Assessment* **2** 167–175.

——, ——. 1998. Reserve assemblages of critical areas: A zero-one programming approach. *Eur. J. Oper. Res.* **104**(3) 497–509.

Willis, C. K., A. T. Lombard, R. M. Cowling, B. J. Heydenrych, C. J. Burgers. 1996. Reserve systems for limestone endemic flora of the Cape Lowland fynbos: Iterative versus linear programming techniques. *Biol. Conservation* **77** 53–62.