RULE-BASED DECISION MAKING: A WAY TO DETERMINE WHICH ALIEN SPECIES TO CONTROL

Kent W. Bridges

ABSTRACT

Most decisions that managers make are based on information. Presumably the more relevant the data, the better the decision. Managers rarely feel that they have enough information, but it is not always possible to easily obtain more. What is required are decision-making techniques that make the best use of incomplete information. In addition, a good decision-support system should point to what data might best be collected in future studies. We are entering a period of rapid development of sophisticated decision-making tools. It is appropriate to look at how these tools can help managers decide whether alien species are likely to become problems. While it is too early to expect the new technologies to provide direct assistance, an understanding of their operations should help guide future data collection efforts.

HOW DECISIONS ARE MADE

We generally work under the premise that more relevant information enables us to make better decisions. If this is true, we are then faced with obtaining enough relevant data. For most management problems, some information will always be at hand; but the amount is rarely considered sufficient. Sometimes we know that data exist that will assist us with a particular decision-making problem, but we are unable to actually get the data. Reasons for unavailability range from location inaccessibility to inability to take the time to find data. Collection of new data is frequently unacceptable because of time constraints, or resources are not available. Sometimes it is not entirely clear what new information is needed. Biological problems cannot always be solved with the collection of new data anyway. For example, when historical patterns are being examined, "old" data have great value. More often than not, however, such data are incomplete and ambiguous.

How do we make the most effective use of data in the decision-making process? At the least, it would be reassuring to know that data are

actually being used to make decisions in an unbiased, objective manner. We can be even more demanding and pose the question of whether it is possible to have a system in which decision making is so well organized that it is effectively "reproducible," that is, other people who use the same data would be led to the same conclusion. We refer to such formal decision making as an inference process and will explain how we can use specific inference techniques to help us answer questions about weed control below.

WEED CONTROL: WHICH ALIEN SPECIES ARE LIKELY TO BECOME PESTS?

Two aspects of alien species control are examined here. The first concerns the problem of determining the existing distributions of species. This involves evaluating information on the geographic range of the weed in its native habitat and the extent of its spread in its new habitat. The second aspect involves making predictions about the likely changes in distributions. How rapidly will the weed spread, and what will be the extent of the invasion? Species distribution questions are too broad to be answered directly. Instead, they need to be divided into a series of related "operational questions."

Many operational questions are relevant to the evaluation of species distributions. For example:

1. What is the highest elevation at which each species is found?
2. What is the usual range of each species?
3. What species have broad ranges?
4. What other species are associated with a particular species?
5. What is the life-form spectrum at a particular location?
6. What species might be found at a particular location?

These questions can be analyzed in several different ways, ranging from the direct use of information to the need to do complex, indirect inferences. A naturalist who is intimately familiar with species distributions in a region will perform these analyses mentally. But the concern here is how such questions can be answered by someone relatively unfamiliar with the region, who has access to distributional data, such as herbarium records and field notes. The point is not to remove experienced people from the decision-making process, but to examine whether there are ways that this process can be assisted so that all managers will be able to make the best possible decisions.

A useful place to start examining the existing decision-making process is by evaluating how each of the operational species-distribution questions listed above can be answered. The first question shows the direct use of information: the existing species distribution records, or new collections, are searched for the highest elevation value. The value provides the answer to the question. The second and third questions involve the use of statistical concepts. The term "usual range" implies that an answer includes an interpretation of the modal (most common) occurrence of the species as well as something like its interquartile

range. Questions of this type are answered by producing some sort of statistical distribution for all of the recorded distribution information.

The fourth and fifth questions are examples of questions that require the use of complex definitions. The answer to question four involves the definition of the term "associated with," while the fifth question requires the definition of "life-form spectrum." These terms make use of relatively complex categorization processes that must be made based on analyses of large amounts of data. As a result, considerable work must be done on information relating to all the species in the geographic region of concern before specific questions can be answered.

The most complex type of question is one that requires a series of individual decisions. This is illustrated by the sixth question. The sort of individual decisions that make up the question-answering strategy will be examined after a few comments on the sources of data and how these can be organized and searched.

SOURCES OF INFORMATION AND THEIR ORGANIZATION

Information on species distributions comes in many forms. Although historical records are of considerable importance, the way in which information is recorded often reflects the period when it was produced. Some types of records, such as herbarium sheets, show relatively little change in the last hundred years in how annotations have been made. In contrast, field notes are particularly variable and strongly reflect the skills and purposes of their author.

Some information can be found in the published literature. Notes on species distributions are part of almost every ecological paper, for example, even though they are not the primary focus of the study being reported. Published papers are a diffuse source of information. They are also a relatively incomplete resource since some critical information can be lacking. The increasing trend to edit journal articles so that they do not stray from their specific topic makes incompleteness likely.

One of the most valuable sources of information on species distributions is the "common" or "general" knowledge held by people familiar with the species in a region. This sort of information is known as "heuristics." It is rarely recorded. Instead, it is either acquired independently through experience or passed along in informal discussion.

When a large amount of information has a consistent structure, it is possible to record it in a formal database. This is the case with herbarium information relating to species distributions, where species names, collection dates, and collection locations can be recorded in a simple table. It is also possible to record data on the sites on which species have been found (*e.g.*, elevation, soil type, rainfall, and temperature). Tables of such information can be used to answer specific questions.

Many tools can be applied to formal databases. Personal computers are increasingly being used both for the storage and use of databases. Complex searches, even when they involve information from several tables, are made quite easily and are usually done to produce a short list of species (or sites) that meet some set of criteria (Jacobi, this volume). For example, a search may be made of all records of species that are now listed as threatened or endangered, collected at a specific elevation on a particular island during a particular time period. Questions of this complexity can be handled routinely, provided an appropriately structured database exists. Designing such a database is a relatively simple matter, given the sophistication of the relational database tools that are now available.

Not all information can be stored in a formal database. Information such as heuristics (*i.e.*, "common knowledge") can only be stored in what can be called an "*ad hoc* organization" database. Such a database has two kinds of information, referred to as "facts" and "rules." The facts record simple observations and appear in the database like a simple English sentence. For example, "Pu'u Maka'ala has an elevation of 5,610 ft (1,700 m)." Rules are more complicated and correspond to a definition. The form of the definition is that something will be true if a set of conditions is met. For example: "A particular site will be considered a bog if it has high rainfall and poor soil drainage and topography that restricts runoff." More complete and stylistically accurate examples of facts and rules are given below.

Relatively few database tools in widespread use handle *ad hoc* database structures. The work reported here examines database organization using PROLOG, a programing language based on predicate calculus (Clocksin and Mellish 1984). PROLOG stands for "programing in logic" and is one of the logic programing languages. It provides a very natural way to do the sort of species distribution investigations that are described here, in large part because of the way information is stored, and because the statements used to pose search questions in PROLOG closely mimic English descriptions. However, the match is not exact. Since PROLOG has its own formalities, like any other computer language, and the purpose here is not to teach this language, all the following examples will be paraphrased compromises between English and PROLOG.

EXAMPLES OF AN AD HOC DATABASE SYSTEM

Examples of facts and rules that could be used in an *ad hoc* database are given below. Facts and rules use a notation in which some compound words end with the letters x, y or z. These words indicate that something appropriate has to be substituted. For example,

species-x found-on site-y on-date-z

requires three substitutions: x, y, and z. Some of the substitutions are made when the database is constructed. This would be the case in the above example. There would be many listings of species distributions in the form shown in the example, such as

passiflora-mollissima found-on puu-makaala on-date-11-24-1974.

This sort of substitution is most characteristic of the facts that are stored in the database.

Rules, in contrast, generally do not have all the substitutions made when they are stored. A rule could be stored that states:

site-y is-a-bog if
    site-y has-an-annual-rainfall-very-wet and
    site-y has-drainage-poor and
    site-y has-topography-flat or
    site-y has-topography-basin.

Note that site-y has not been identified in the rule. When PROLOG uses rules, it will substitute sites in the database, one at a time, to determine which ones satisfy the criteria listed in the body of the rule.

When PROLOG structure is applied to species distribution problems, the following sorts of facts, with the appropriate values substituted, would probably be included in the database:

species-x found-on site-y on-date-z
site-y has-soil-type-z
site-y has-elevation-z
site-y has-annual-rainfall-z
site-y has-slope-z
site-y is-on-island-z
botanist-w collected-at site-y on-date-z.

These sorts of facts are likely to come from herbarium records and published site descriptions or field notes.

PROLOG-like rules that relate to a database with such facts would be like:

species-x may-be-found-on site-y if
    species-x found-on site-y on-date-z and
    z = this-year and
    confidence = 1.0

species-x may-be-found-on site-y if
    species-x found-on site-y and
    site-y has-elevation elevation-similar-to site-z and
    site-y has-annual-rainfall rainfall-similar-to site-z and
    site-y is-on-island-z and
    site-y is-on-island-z and
    confidence = 0.75

site-y elevation-similar-to site-z if
    site-y has-elevation-x and
    site-z has-elevation-y and
    abs (x-y) less-than-200

A few observations should be made about these rules because they show so much of the power of this type of database. Several rules can define any particular situation. In the example above, there were two definitions of the "may-be-found-on" rule. The first alternative implies that one would expect to find a particular species on a site if it had been recorded on that site this year. The second alternative means that one would expect to find a particular species on a site if it had been found sometime on a site on the same island and the sites had similar elevations and amounts of annual rainfall. The alternatives are used one at a time. If the first "may-be-found-on" rule is discovered to be true, then the remaining ones are not used. If the first one is not found to be true, then the second one is tried, and so forth, until all the alternatives are exhausted.

Confidence statements were given at the end of the "may-be-found-on" rules. These are not conditions that are being tested, but values that are being saved when an alternative rule is found to be true. This provides a way of storing the likelihood of the results. In this example, it means that a recent sighting of a plant on a site is stronger evidence that it is still likely to occur there, than finding that it was only found once on a site with a similar climate. Whether these values (100% versus 75% certain) are correct is a matter of debate and is not the issue here.

Sometimes it is useful to use indirect rules. That is true in the case of the "elevation-similar-to" rule, where a simple test is performed to see if two sites are within 200 m of each other. This is useful because it makes it possible to simplify the statement of rules. It also makes it possible to change a definition such as "elevation-similar-to" in a single place, rather than having to modify every place that it has been used in other rules.

The purpose of a species distribution database is to obtain answers to the two basic, specific questions that can be asked. The first question tests whether there is evidence in the database to show that a statement is true. For the example used here **passiflora-mollissima found-on puu-makaala**: PROLOG automatically searches the database, looking at facts and testing rules, until it finds a "found-on" fact that is true (in which case it gives the response YES), or until it gets through all the rules with none of them being true (and it responds NO).

The second basic question that can be used with PROLOG provides lists of solutions, not just YES or NO answers. In particular, the responses list the values for items in statements for which specific information has not been supplied. This is seen with three questions and typical PROLOG responses (shown here as indented lines below the question):

x found-on puu-makaala on-date-11-24-1974
x = passiflora-mollissima
x = rubus-ellipticus

passiflora-mollissima found-on puu-makaala on-date-x
x = 11-24-1974
x = 9-17-82

x found-on puu-makaala on-date-y

x = passiflora-mollissima	y = 11-14-1974
x = rubus-ellipticus	y = 11-24-1974
x = passiflora-mollissima	y = 9-17-1982

IMPLICATIONS FOR MANAGEMENT

The sort of *ad hoc* database shown here has a number of implications for managers who are faced with alien species control problems. This new database structure is well adapted for storing species distribution and site information. The *ad hoc* database differs from the formal database in its ability to include complex rules as part of the overall database. This means that it should be possible to use information that was once considered to be too informal to be stored or used in a database system.

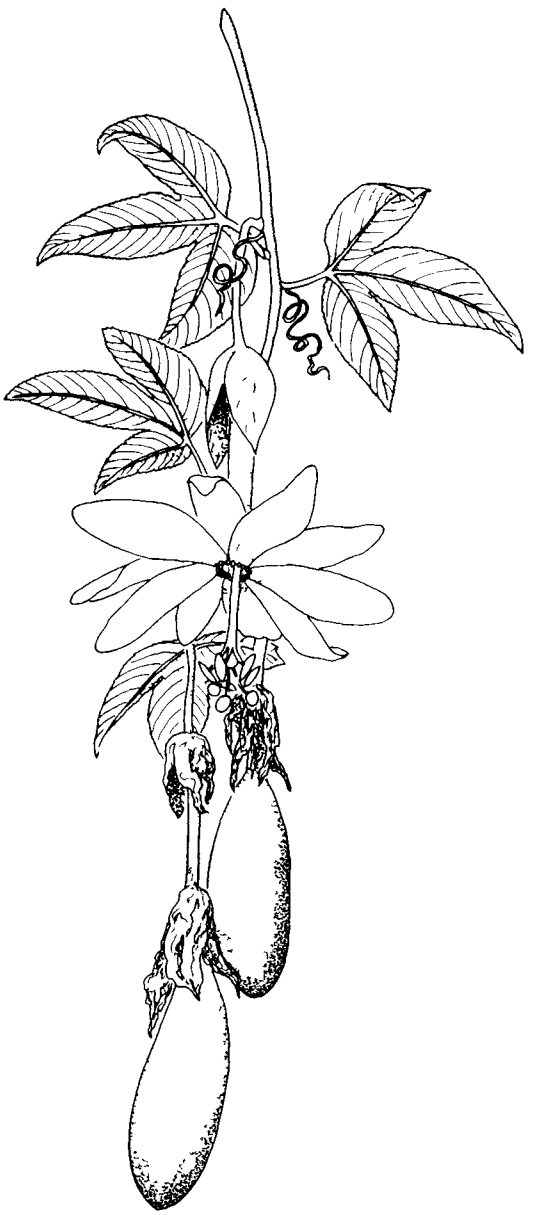
The flexibility provided managers can be seen in another example. Suppose that a database has been constructed that includes all the species distribution information from a large set of herbarium records. This information has been transcribed as faithfully as possible, but when it comes time to use it, the manager discovers that a particular plant collector was relatively inaccurate in the identification of a particular species. Rather than going back into the files and changing the species identification made by this collector, it would be possible to add a rule to the database that had the effect of saying "so-and-so wasn't good at identifying x so ignore these records." This maintains the integrity of the database and clearly identifies a constraint that will be applied when the database is used. It is obvious that it would be relatively easy to remove such a rule-based restriction, while any changes in the database would be particularly difficult to reverse.

Since rules in the database provide a way of recording "common sense," it is possible that a manager could build a database that has "institutional memory." As researchers and field workers come to understand some of the subtle relationships that they feel occur in the species distribution patterns, these can be coded as rules and entered into the database. This information can remain in the database, regardless of turnover of personnel. Eventually, the database could be an extremely rich source of very practical information that would otherwise have to be rediscovered by new personnel before it could be used. The recording of "common sense" may also have a valuable side benefit. New personnel should be able to gain considerable insight by examining the database, which could

likely shorten the apprenticeship necessary to become familiar with a new environment.

It may be possible for managers to undertake analyses of alien species problems more objectively by using a database of the sort described here. If a management question is posed as a rule, then it is possible to examine the way that the answer is obtained. At the least, this forces questions to be clearly stated.

No species distribution databases exist that have the structure described here. The potential benefits that have been shown provide motivation for the storing of information from herbarium records, field notes, and the thoughts of field workers in a readily accessible, usable form. Exciting possibilities await the manager who has an opportunity to search such a database.



Literature Cited

Clocksin, W.F., and C.S. Mellish. 1984. *Programming in Prolog*, 2nd ed. Berlin: Springer-Verlag.

Jacobi, J.D., and F.R. Warshauer. [this volume] Distribution of six alien plants in upland habitats on the island of Hawai'i.