The GARP modelling system: problems and solutions to automated spatial prediction

DAVID STOCKWELL

Research Article

The GARP modelling system: problems and solutions to automated spatial prediction

DAVID STOCKWELL

San Diego Supercomputer Center, University of California, San Diego,
9500 Gilman Drive, La Jolla, CA 92093-0505, USA
e-mail: davids@sdsc.edu

and DAVID PETERS

Tasmanian Parks and Wildlife Service, 139 Macquarie Street, Hobart, Australia

Abstract. This paper is concerned with the problems and solutions to reliable analysis of arbitrary datasets. Our approach is to describe components of a system called the GARP Modelling System (GMS) which we have developed for automating predictive spatial modelling of the distribution of species of plants and animals. The essence of the system is an underlying generic spatial modelling method which filters out potential sources of errors. The approach is generally applicable however, as the statistical problems arising in arbitrary spatial data analysis potentially apply to any domain. For ease of development, GMS is integrated with the facilities of existing database and visualization tools, and Internet browsers. The GMS is an example of a class of application which has been very successful for providing spatial data analysis in a simple to use way via the Internet.

1. Introduction

Museums, environmental groups and governments provide environmental resource information through a wide variety of media. A significant proportion of these data are records of the spatial location of species, such as their breeding sites, or observations of their presence. These 'species data' have a very incomplete geographical coverage, particularly in the vast and sparsely populated arid regions of continents. Prediction of distributions as probability surfaces can provide a complete and fine-scale spatial coverage of potential distribution, even in areas where there are no data available. These predicted distributions can then be used as building blocks for further analysis: e.g. to assess the status of nature reserves, to guide more efficient surveys, to establish the actual distribution of rare and endangered species, or to aid scientific research into biogeographical questions.

Developing maps of potential distributions on a species-by-species basis puts an enormous load on skilled personnel. Each inquiry entails accessing the database, using a statistical modelling package, and preparing and printing maps through a

GIS. Automation of this repetitive task would increase the availability of the data through decreased response time and cost, and free skilled operators for more challenging tasks. Similar views are championed by a number of authors who have envisioned future GIS which include a complete integration of GIS with spatial analysis and modelling facilities (Goodchild *et al.* 1992), intelligent processing producing end-products that meet user-specified quality requirements (Burrough 1992), and predominantly graphical and interactive user interfaces (Peters 1990).

The view motivating this work is that as well as better systems integration, GIS require the development of robust methods of analysis that transparently and explicitly address a range of data handling and modelling problems which would otherwise require substantial user intervention. We arrived at this view after comparative modelling studies showing significant difference in results between methods of species distribution prediction (Stockwell *et al.* 1990). It was also apparent that studies evaluating existing modelling methods by comparing the accuracy of methods on a limited number of species, provide little guidance on selection of methods for system users or insight into the causes of inaccuracy for system developers (Skidmore *et al.* 1996). The progress of the field requires a better understanding of the causes and solutions to inaccuracies in modelling real-world data.

The paper reports on experiences with the GARP Modelling System (GMS), an integrated spatial analysis system for predicting distributions of plants and animals. It was first implemented at the Environmental Resources Information Network (ERIN) (Boston and Stockwell 1994). The GMS inputs data from a user or a database of species location records, analyses it with a machine learning based analytical package called GARP (Genetic Algorithm for Rule-set Production) (Stockwell and Noble 1991, Stockwell 1992) and displays the output in a World Wide Web browser. The most recent implementation of GMS is as an information system for the Biodiversity Species Workshop at the web site http://biodi.sdsc.edu (figure 1).

The central themes of this paper are the problems in providing computational analysis resources for the rapid and automated development of spatial models using arbitrary spatial data points. We examine the way the GMS has addressed these problems. The problems addressed here largely arise from the use of arbitrary or *ad hoc* data collections, which are typically not the result of well designed experiments. Where data are not the result of well designed experiments, taking multivariate analysis methods beyond exploratory stages is regarded as unreliable due to variable quality and characteristics of the data (James and McCulloch 1990). Strict application of these methods potentially excludes a large number of existing collections. For example, the majority of herbaria or museum records of the species locations are the result of *ad hoc* surveys and opportunistic sampling.

The problem is not only that the assumptions of multi-variate statistics are generally not met by data extracted directly from museum databases. The types, distributions and correlations of the environmental data such as climate, soils and remote sensing data used for the independent variables can also be problematic. In a method for unsupervised, automated analysis of generic data sets based on simplified versions of Bayes' rule, forms of data used in the system were found to be restrictive (Aspinall 1992). This was due to the strong assumptions of Bayes' theorem, particularly the need for conditional independence in the predictor variables. Experimental analysis of a Bayesian-based system using environmental data for automated prediction of occurrences of species found significant inaccuracy due to a range of data characteristics, such as correlation in environmental variables
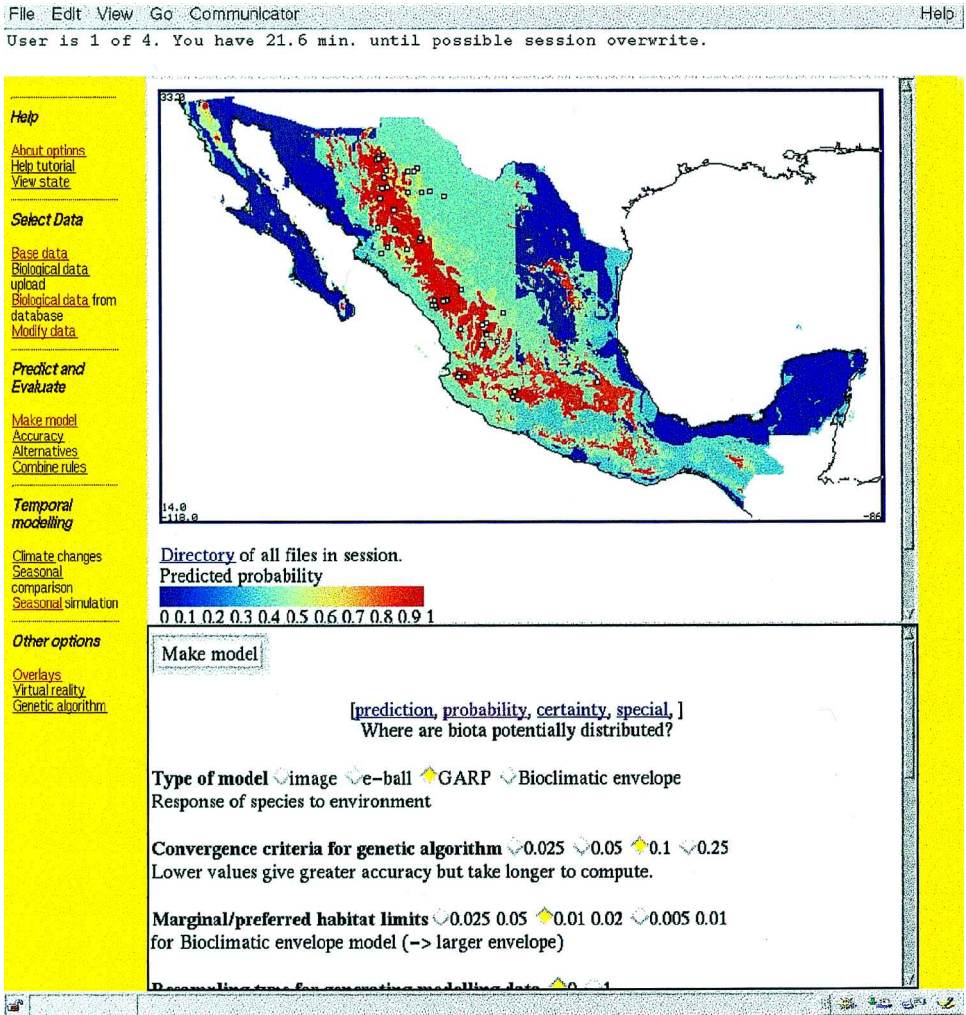
Figure 1. The user interface for the GMS at the Biodiversity Insight Systems site http://biodi.sdsc.edu. This site allows development and manipulation of sets of rules developed from modelling the users own data. Shown is the predicted distribution of a species of bird, the Eared Trogon, in Mexico.

(Stockwell 1997). Other problems were due to the structure of the analytical system, limiting applications to prediction of mutually exclusive events (Stockwell 1993).

One possible approach to robust analysis is to try to incorporate 'intelligence' in the system to determine the appropriate method to apply to given data (Burrough 1992). This intelligence can take the form of an expert system for examining the data and choosing the most appropriate method from a range of methods. There are, however, many methods available, and difficulties in determining the appropriate conditions for application. For example, one assumption of bioclimatic modelling methods such as BIOCLIM, is that the abundance of the species is determined by climatic limits (Nix 1986). As climate could potentially be a minor factor or even irrelevant to distribution in some areas, expert systems could improve reliability by

determining the accuracy of the assumption that the species are in fact sensitive to climate before applying the method.

The alternative approach is to implement robust analysis systems, i.e. systems that produce reliable results under a wide range of operating conditions, or problem domains. A fertile source of such systems are machine learning methods such as genetic algorithms (GAs), decision trees and neural nets because they are designed for analyzing poorly structured domains. The GARP modelling system uses a genetic algorithm, the basic concept of which was developed by Holland (1975). GAs have been applied to a wide range of domains, including numerical function optimization (Bethke 1981, Brindle 1981), adaptive control system design (DeJong 1980), and artificial intelligence task domains where the structure of the problem domain prohibits the use of classical statistical methods and gradient search techniques (Goldberg, 1989).

In addition to the well documented robust performance of GAs, the genetic algorithm in GARP has a feature which extends the capacity of GAs for generating and testing a wide range of possible solutions—the capacity to simultaneously generate and test a range of types of models, including categorical, range-type and logistic models. One of the main purposes of this paper is to describe this characteristic and demonstrate its utility in analysis of *ad hoc* data.

1.1. *Overview of the GMS*

Ideally a system for predicting distributions of species would take as input the species and the area of interest and produce images, maps, documentation, and easily comprehended models for explaining relationships in the data via the Internet. While it is possible that GIS could provide at least some of the functions required, the use of GIS packages in GMS implementations has been limited to preparing environmental data. Some concerns with existing packages have been: automatic license restrictions that limit the number of users that could access the system simultaneously, long delays in completion of simple data processing tasks, inefficient interaction of spatial data with spatial analysis programs, and limited flexibility for experimenting with novel types of data analysis.

Thus the GMS was developed as a number of *C* language modules, linked by PERL scripts. The intermediate data in GMS are propagated through sub-systems via common files and Unix pipes. Thus, within the range of architectures for spatial information systems, GMS is classified as a loosely coupled system (Abel *et al.* 1992). Loosely coupled, or 'open' systems, are easy to re-configure and therefore easy to integrate and customize. For example, when a user requests the prediction for a particular species, the recorded sightings are extracted from a data base in point coverage form. The first two columns contain the longitude and the latitude, and the following columns may contain an abundance value for a species, or value of a variable. These values can originate from any source; files in directory trees, databases, or GIS e.g.:

$$150.775 \quad -35.005 \quad 0$$
$$148.005 \quad -35.005 \quad 1$$
$$\ldots$$

Environmental data are stored as grey-scale raster images (called layers) with one byte value per grid cell. The raw grey-scale image format has two main advantages. It provides a compact representation of environmental information, and it is viewable by image processing tools. As each layer is a geographic grid it can potentially be very large; a grid of 258×410 cells contains 106K points, requiring

significant memory resources if stored as floating point numbers. Storing these layers with one byte per cell reduces the amount of memory needed. Image processing tools are optimized for these data, providing fast efficient data manipulation.

A typical GMS implementation uses over 30 rasterized environmental data layers of climate and topographic variables. Climate data are derived from available rasterized surfaces of continent-wide averages of temperature and precipitation. These data were previously derived by terrain-modified interpolation for each grid cell on the land surface from weather station data. GMS includes topographic (slope, aspect, altitude) and substrate variables (soil types, pH, geology) in addition to the climatic variables.

The analysis system (i.e. GARP) is composed of eight programs shown in table 1, each with a specific function. The central program, called EXPLAIN, develops the model. The model developed by GARP is composed of a set of rules, or if-then relationships. Rules have been widely used for prediction in expert systems. The difference between a rule and the more familiar regression model, say, is that a rule has preconditions that determine when it can be applied; when these conditions are not met the rule is not used. The set of rules is developed through evolutionary refinement, testing and selecting rules on random subsets of training data sets. Application of a rule set is more complicated than applying a regression equation as the prediction system must choose which rule of a number of applicable rules to apply.

The goals of the system are to maximize significance and predictive accuracy of rules without 'overfitting' or overly specializing rules. Significance is established through a χ^2 test on the difference in the probability of the predicted value (usually presence or absence of the species) before and after application of the rule. Maximizing significance and predictive accuracy (the proportion of data correctly predicted) is actually a novel goal for analytical systems; most modelling methods typically maximize significance only.

The outputs of the prediction are converted into images with image processing utilities. Scripts written in PERL format output into HTML (Hyper Text Markup Language) documents for viewing in a Web browser, providing the user interface for integrating images, textual output, and supplementary documentation. Thus the GMS is integrated into the World Wide Web (WWW) (Putz 1994).

Table 1.   The components of GMS in the ERIN-GMS implementation.

Program	Function
Data preparation	
RASTERIZ	Converts spatial data files to raster layers
PRESAMPL	Produces training and test sets by random sampling
Model development	
INITIAL	Develops an initial approximate model
EXPLAIN	Refines model using a genetic algorithm
Model application	
VERIFY	Provides predictive verification information on the output rule set
PREDICT	Takes the model and predicts probability for each value
Model communication	
IMAGE	Takes predicted probabilities and produces a number of results in required image format
TRANSLAT	Takes the model and forms natural language explanation of rules

2. Problems and solution in the GMS

The GARP programs can be logically grouped into data preparation, model development, model application and communication. Operations performed by the programs usually satisfy more than one system need. The approach has been to identify system needs by identifying the potential problems of data and modelling, and modifying the system to address the potential problem without reducing applicability.

The first problem in developing an automated modelling system is how to structure the components of the system. A 'production line' type architecture, with a linear configuration of modular components provides an efficient, simple structure. Production line architecture is a simplified form of the work-flow model for a generic environmental information management system (Davey *et al.* 1995). The use of modules with well defined, simple data types and simple functions provides flexibility, faster implementation times and lower maintenance costs (Black 1991).

2.1. *Data preparation*

Ideally an analytical technique receives a random sample of adequate size throughout the range of values to be predicted. It should also be able to use all available information, both categorical classes and continuous environmental variables. The first GMS program, RASTERIZ, prepares the data to make it more amenable to analysis by increasing uniformity and consistency of typing.

2.1.1. *Problem: non-uniform population*

Non-uniformity of the data is often due to a range of scale of the data. A recorded sighting of a species will have a location associated with it, but the precision of the measurement of locations can vary from a map sheet with a side of over 100 km to locations precise within a few metres. Some data are duplicated in the data base, due to multiple samples or return visits to a single location. Put differently, the amount of information at each point is not uniform: single outlying points have great importance while duplicates provide little new information.

The program RASTERIZ maps point data into a spatial grid at a given scale. Data redundancy due to duplicate records and records from close locations are removed by absorption into a single cell. As expected, the redundancy in an arbitrary data set increases with the coarseness of the base grid, as shown in a sample of bird data analyses for North America at 0.5°, 0.05° and 1 minute on table 2. By modifying the data in this way, the potential non-uniformity is reduced by bringing all data to the same spatial scale.

2.1.2. *Problem: range of types of the data*

Different types of data are mapped into grid cells differently. RASTERIZ recognizes three types of data. With species or presence/absence data, a cell takes a presence

Table 2. The effective number of data points as the fineness of the base data is increased for observations of a species of bird in North America.

Scale in degrees	No. of points	No. of grid cells	Fraction
0.5	62	36	0.58
0.05	62	52	0.84
0.0167	62	55	0.89

value if one or more points falls within it, otherwise it remains zero. With categorical data a cell takes the value of the mode of the values of the points that fall within it. With continuous data the cell takes the mean value. A cell is a single byte, its value determined by linearly scaling the point value between minimum and maximum values. Scaling of the environmental variables into single bytes reduces the effects of differing magnitude between variables that can effect some analytical techniques. The result is that all three types of environmental data are incorporated into the system in a single efficient format.

2.2. *Producing the data set*

The first stage in predictive modelling in GMS is to map the species locations into a grid at the same scale and extent as the environmental data. A single data point for the subsequent programs is then the values of the coordinates of a vector that passes vertically through the layers at the same geographical location. These data points, though uniform, may be biased in various ways.

Sampling bias refers to any departure of the data set from a random sample of the possible data points (or population). When a model is developed to reflect the patterns in a population, sampling bias imposes unwanted patterns on the data. A biological data set from a herbarium or museum database will have a range of omissions or disproportionate representations of information, such as spatial bias and from one to very many data points.

A process called presampling reduces the effects of sampling biases. The program PRESAMPL prepares a data set for analysis by controlled random sampling of the raster data set produced by RASTERIZ. In this way some inherent biases can be reduced, and a consistent, less biased data set presented to the modelling algorithm. By default, PRESAMPL outputs a set of 2500 points, with even proportions in each of the dependent, or predicted values, randomly selected from non-masked (e.g. non-ocean) areas. The output of the program PRESAMPL is two sets of data points, called *train* and *test*. The model is developed using the *train* data set, the *test* set is used for testing.

2.2.1. *Problem: missing values (presence-only data)*

One major form of sampling bias in ecology is in the use of presence-only data. Most museum databases, for example, record where a species was collected, but usually no information is available on where the species did not occur. This represents a type of bias where the sampled set is biased completely towards a particular value of the dependent variable. The solution provided by GARP is to generate the pseudo-absence data called 'background' by selecting points at random from the geographical space. The data set is then composed of two different types of data: presence and background data. Where true absence records are available, these can be included, with or without background data. For some species or functional groups, it may be possible to postulate a pseudo-survey region which restricts the area from which pseudo-absences can be drawn.

2.2.2. *Problem: variable prior proportions*

A data set of sightings of a species from a field survey will be composed of presences and absences in varying proportions, depending on the abundance of the species, and how much effort was spent in favourable habitat. While these proportions can in some cases represent the relative scarcity or abundance of a species, they create analytical difficulties. For instance, in rare species, if the proportion of sightings

is very low, such as less than five percent of the total, the strategy of predicting the absence of the species everywhere will have an expected accuracy of 95%. Thus a model using no predictor variables that predicts the species everywhere may be output due to its high accuracy. Most measures of the quality of the model such as significance usually become more inaccurate at values close to zero or one. Also, it is difficult to compare the predictive accuracy of models on different species when the prior proportions are not the same. Presampling the data to even proportions, i.e. 50% presences and 50% absences allows consistent comparison of accuracy between species, and produces more reliable models.

2.2.3. *Problem: large and small numbers of data*

Many modelling methods allow arbitrarily large numbers of data points, but this can often lead to long computation times for little gain in predictive accuracy. A limit to data points limits the computational time required to develop the model and typically provides sufficient information for the system. The number of points may be varied to achieve repeatable results, depending on the species.

Small numbers of records occur with rare species or those of a very restricted range. When this occurs sampling with even proportions cannot occur without replacement of the record. While PRESAMPL supports the option of sampling with or without replacement, sampling with replacement provides generality by allowing model development using the range of possible frequencies of records, including from a single datum.

Research into the implications of these decisions is ongoing. For example, although this generality may compromise assumptions of independence, experience has shown that often a very small number of records are needed to correctly define the distribution of a species, once data has been sampled in this way for subsequent analysis.

2.3. *Model development*

The two goals of modelling are repeatable and accurate results. This is usually the case with explicit numerical methods when the data sets are well suited to the methods of analysis. However, in practical application to arbitrary data, many numerical methods can fail to produce a result. This can be due to lack of convergence of algorithms, non-invertability of matrices, or memory limitations that limit the size and dimensionality of data sets. For problems without explicit solutions, exhaustive search methods take a prohibitively long time as the size of the search space increases.

2.3.1. *Problem: achieving repeatable results*

Stochastic algorithms such as GAs incorporate random elements or stochasticity into the strategy for searching the space of possible solutions. While stochastic algorithms can find solutions to problems in large search spaces, it can be at the expense of repeatability because of the random search method. In some cases the convergence on solutions can take a long time, because in the initial stages the initial population contains no useful information. Algorithmic performance in the GARP modelling system is considerably improved by developing good initial estimates of the model. Initial models are generated using standard parametric statistical methods and heuristics by the module INITIAL. The estimates are used as a starting point for the genetic algorithm in the following program.

2.3.2. *Problem: achieving consistently accurate results*

A more subtle but pervasive performance problem is the match of the class of predictive model to the actual response type present in the data. In modelling the

response of species to the environment the form of the response is not known prior to the analysis. Nor is there general agreement in the literature on this issue. While ecological theory would suggest a continuous uni-modal response modelled by quadratic logistic equations (Austin and Smith 1989), in practice the interplay of data artefacts and biases distorts this theoretical picture in unpredictable ways.

For example, in mapping vegetation distributions for forestry, the distribution of forest trees in a pristine area may correlate highly with a uni-modal response curve to soil fertility, say, representing smoothly changing quality of habitat for the species. However, forestry operations in the area will have changed the distribution from the pristine, through removal and replacement with other species. In a GIS, one would model such discrete factors using overlays. There may be many overlays, corresponding to different types of discrete phenomena. Thus the final model is a complex combination of continuous response curves and logical spatial operations, nested in a multi-level hierarchical structure (Lyndenmayer *et al.* 1991).

In an alternative approach to dealing with this problem, GMS simultaneously uses a range of forms of model to model the range of potential relationships in the data. The rules that contain each model differ in type, but are evaluated with the same criteria: statistical significance and predictive accuracy. Given a set of rules making up a GARP model, different rules are selected automatically for predictions at each cell, based on the estimated predictive accuracy of each rule.

GARP uses envelope rules, GARP rules, atomic and logit rules. The following is an example of an envelope rule. The conjunction of ranges for all of the variables is a climatic envelope or profile, indicating geographical regions where the climate is suitable for that entity, enclosing fixed percentiles of values for each parameter.

```
IF TANN=(23,29]degC AND RANN=(609,1420]mm AND GEO=(6,244]c
THEN SP=PRESENT
```

In natural language, this rule states that if the annual temperature (TANN) falls between 23 and 29° C, and the annual rainfall (RANN) falls between 609 and 1420 mm, and the value of the category of geology (GEO) falls in the range 6 to 244, then predict that the species is present. A GARP rule is similar to an envelope rule, except that variables can be irrelevant. An irrelevant variable is one where points may fall within the whole range. An example of a GARP rule modified from the above is given below:

```
IF TANN=(23,29]degC AND GEO=(6,244]c
THEN SP=ABSENT
```

An atomic rule, as the name suggests, is a conjunction of categories or single values of some variables. The natural language translation of the rule below is that if the geology is category 128 and the elevation (TMNEL) is 300 m above sea level then predict absent.

```
IF GEO=128c AND TMNEL=300masl
THEN SP=ABSENT
```

Logit rules are an adaptation of logistic regression models to rules. A logistic regression is a form of regression equation where the output is transformed into a probability. For example, logistic regression gives the output probability p that determines if a rule should be applied where p is calculated using:

$$p = 1/(1 - e^{-y})$$

and y is the sum of the linear equation in the precedent of the rule: e.g.

```
IF 0.1-GEO 0.1+TMNEL 0.3
THEN SP=ABSENT
```

The capacity of rule sets to increase the coverage and accuracy of rules is shown on table 3. The regions predicted by each type of rule in isolation are usually less than the total area. The predictive accuracy of models composed of sets of different kinds of rules typically equals or exceeds the accuracy of models composed of rules of a single type. Thus, the system makes use of high accuracy rules that apply in different areas to achieve an optimal overall coverage.

In theoretical terms, the four different types of models, and, potentially, a large number of variables, create a problem of finding a set of good models in a very large search space. Theoretical studies of DeJong (1975) and Holland (1975) and experimental studies (Bethke 1981, Brindle 1981, DeJong 1980) have shown that GAs are particularly efficient at finding solutions to problems which have many variables, are noisy and contain potentially many solutions.

The rules are developed by a process of incremental refinement by the genetic algorithm. Each iteration is referred to as a generation, in which the set of rules are tested, reproduced and mutated. A description of the GA procedure in the module EXPLAIN follows:

1. Initialize population of structures.
2. Select random subset of data.
3. Evaluate current population.
4. Save the best rules to a rule archive.
5. Terminate outputting the rule archive, or continue.
6. Select new population, using rule archive and random generators.
7. Apply heuristic operators to population.
8. Go to 2.

The GARP algorithm starts by inputting an initial set of rules generated by the INITIAL program. The first step in the GARP iterative loop is to select a data set by randomly sampling half the available data. The next step is to evaluate the rules on the sampled data. For each of n data points the following values are incremented:

1. no —the number of points the rules applies to.
2. pYs —the number of data with the same conclusion as the rule.
3. $pX Ys$ —the number of data the rule predicts correctly.

Table 3. The coverage and accuracy achieved for individual types of rules and the combined model of all rules. Errors expressed as standard errors of the means of three repeats.

Model type	Coverage	s.e.	Accuracy	s.e.
Envelope	0.66	0.1	0.87	0.01
GARP	0.48	0.00	0.87	0.00
Logit	0.97	0.01	0.78	0.00
Atomic	0.21	0.02	0.94	0.03
All	1.00	0.01	0.81	0.01

The following values are then calculated to evaluate the performance of a rule:

1. Coverage $= no/n$
2. Prior probability $= p\,Ys/n$
3. Posterior probability $= pX\,Ys/no$
4. Significance $= (pX\,Ys - no \times p\,Ys/n)/ \sqrt{no \times p\,Ys \times (1 - p\,Ys/n)/n}$

In the terminology of genetic algorithms, each rule is a member of a population. The composition of a population changes with each generation t. The members of the population $P(t+1)$ are chosen from the population $P(t)$ by a randomized selection procedure. The procedure ensures that the expected number of times a structure is chosen is proportional to that structures performance, relative to the rest of the population. That is, if x_j has twice the average performance of all the structures in $P(t)$, then x_j is expected to appear twice as frequently in population $P(t+1)$. At the end of the selection procedure, population $P(t+1)$ contains exact duplicates of the selected structures in population $P(t)$.

Variation is introduced into the new population by means of idealized genetic recombination operators. The most important recombination operator is called crossover. Under the crossover operator, two structures in the new population exchange segments. This can be implemented by choosing two points at random, and exchanging the segments between these points. In most genetic algorithms, recombination occurs on binary strings. In GARP however, recombination acts on values or ranges of values of variables, depending on the type of the rule. For example two GARP rules can exchange ranges of climatic variables in the crossover recombination.

```
Rule 1:
IF TANN=(23,29]degC AND RANN=(10,16]degC
THEN SP=PRESENT

Rule 2:
IF TANN=(35,38]degC AND TMNEL=(19,27]degC
THEN SP=PRESENT
```

Given the two rules above, suppose that the crossover point has been chosen between the variables. The resulting structures would be:

```
Rule 3:
IF TANN=(23,29]degC AND TMNEL=(19,27]degC
THEN SP=PRESENT

Rule 4:
IF TANN=(35,38]degC AND RANN=(10,16]degC
THEN SP=PRESENT
```

The mutation operator changes the value of a variable to a new value. While mutation produces small changes to the rules, crossover introduces representatives of new structures, or combinations of variables, into the population. If this structure represents a high-performance area of the search space, it will lead to further exploration in this part of the search space.

The genetic algorithm will terminate either when a fixed maximum number of generations is reached, or the modification or discovery of new rules is lower than a fixed rate. The set of rules which is statistically significant, is output once the adjustments have fallen below a fixed percentage.

2.4. *Model application*

The first problem, known as the 'overfitting problem', is ubiquitous in modelling. When models overfit, they can predict very poorly despite an apparent excellent accuracy or fit to the data they were developed on. This necessitates the evaluation of predictive accuracy of habitat models using resampling methods (Verbyla and Litvaitis 1989).

The second problem is one of model choice. Its solution is integral to GMS as sets of models represent the relationship between the species and its environment. Predicting with a set of rules is more complex than predicting using a single model because there can be more than one rule that applies in a particular situation. These difficulties include resolving conflict between rules (predictions of different outcomes), and estimating probability when probability estimates of rules differ.

2.4.1. *Problem: estimating actual accuracy*

Two strategies are used to control over-fitting. During model development in EXPLAIN, models are repeatedly evaluated for statistical significance on randomly sampled subsets of the training data. Empirical studies have shown that this approach almost eliminates overfitting (Stockwell 1992). The second strategy is to use the *test* set to provide a better estimate of the actual accuracy of the rules on independent data. To do this, the VERIFY program is run first on the *train* data set and then on the *test* data set, finally calibrating each individual rule to its true independent predictive accuracy.

The accumulation of these results is shown as a confusion matrix for prediction over the input data file. A 'confusion matrix' shows the proportions of types of successes and errors made by the model. In the listing below, the values on the diagonal falling downward to the right are correct predictions of presence of a species when present, and absent when absent. The upper right-hand corner is the proportion predicted as present when the data records an absence, while the lower left-hand corner is the proportion of predicted absences that are actually presences.

```
VERIFY—predictive accuracy of rules using training data

Results for all rules

Confusion Matrix:                    Actual
                                     Present    Absent    Background
    Predict Present                  1113       0         43
            Absent                   0          0         0
            Background               321        0         837
    Conflicts        0               0
    Unpredicted      94              0
    NoMasked         0.000
    Unpredicted      0.074
    Predicted        0.926
    Accuracy         0.843 s.d.      0.007
    Overall Acc.     0.780 s.d.      0.008
```

VERIFY modifies the performance values for each of the rules according to the performance of the rules on the given data set. Thus when VERIFY is run on the test set, the performance values reflect the expected performance of the rules on an independent test set, rather than on the data set used to derive the rules. This

independent verification gives a more reliable estimate of the true performance of the rules in independent test situations.

2.4.2. *Problem: resolving conflict between rules*

The basic strategy for selecting which rule to apply is maximizing the predictive accuracy. Adoption of a strategy where the value predicted is supplied by the rule with the highest expected accuracy, as indicated by the posterior probability of the rule, maximizes the probability of success at each prediction, and maximizes the total accuracy (Stockwell 1992). The program PREDICT uses the rules and the environmental data to generate predictions of all values at each grid cell in the environmental data layers. The output of the PREDICT utility has the value 254 when present and 1 when absent. The value zero is an unpredicted area where no rule applies and 255 is a masked area, outside the area of interest.

A probability surface representing the probability of occurrence of the species in each grid cell is also output. The probability at a grid cell can be derived from the rule set in two ways. First, using the most accurate rules that applies at a site by the rules:

1. if predicted value is present $P =$ posterior probability,
2. if predicted value is absent $P = 1 -$ posterior probability.

In practice this method was found to give sharp transitions between areas of high predicted probability of occurrence and areas of low predicted probability of occurrence. A method that yields a smoother probability surface takes the average of the posterior probability of the most accurate presence rule, and the reciprocal of the posterior probability of the most accurate absence rule. When no rule applies the area is unpredicted.

Each rule set contains a large number of rules, but frequently a few of these perform most of the final prediction. TRANSLAT outputs a list of rules ordered by their usage in the final prediction. Rules at the top of the list were applied most frequently due to general applicability and high predictive accuracy.

2.5. *Evaluation of the GMS approach*

There are two main ways of evaluating predictive modelling systems. The first is to demonstrate validity of the system on a theoretical basis using mathematics and possibly simulated data. The second is to conduct empirical trials and comparisons with alternative systems.

The approach we have taken is to empirically test the system as widely as possible. The GMS system has been in use continuously since 1995 with open access to users world-wide. In addition, tests and validation trials have been conducted to evaluate the system. The validation trials included examination of the results of the system on prediction of distribution of a range of organisms by a number of experts in the biology of those particular organisms.

2.5.1. *Problem: novelty of the system*

One trial elicited the comments of experts on predicted distributions for a range of species (Stockwell 1995). This trial included three species each of: rare species, endangered species, localized endemic, widespread and common, plants, mammals, reptiles, birds, invertebrates, fish, as well as predictive maps for selected feral animals and landscape phenomena (e.g. wetlands and snowline). The distribution data for

the validation were solicited from the experts, the models developed, and predictions sent to the experts for comment.

Based on the feedback from the experts, GMS was capable of modelling the habitat of a wide variety of biological entities at a wide range of scales. Deficiencies noted in the predictions appeared to be due to lack of data rather than intrinsic limitations in the methodology. Two examples of data deficiencies were data sets restricted to State or Territory boundaries, and lack of hydrological and water-quality related environmental data sets. Correcting these deficiencies and collating the appropriate data thus appears to be a most effective way of increasing the quality of species modelling.

2.5.2. *Problem: interpretation of outputs*

The important and unanswered question for use of the system for management of the environment is: which interpretations lead to valid uses? For example, the uses of a model can range from interesting insights, a source of evidence, a vehicle for planning surveys, or a warrant for changing the tenure of areas of land. If people are to use the results of a system, there must be guidance as to acceptable uses. What this guidance should be and how to provide it is a widespread problem in the field of applied modelling.

3.   Conclusions

The GARP modelling system provides automated analytical facilities for the GMS application for modelling the distribution of species from data on species locations extracted from a data base. The system has potential applications anywhere where spatially located data occurrences need to be modelled and the distributions predicted.

The approach of the GMS system in targeting a specific task like prediction of species distributions, and providing it in an efficient integrated system, stands in contrast to the general-purpose toolbox approach adopted in most GIS applications. The task-specific approach has been successful in providing an efficient system that is easy to use. Increased functionality may increase the range of uses of the GMS— at the expense of greater complexity and probably increased user confusion. Perhaps the future of automated spatial analysis lies in development of many applications dedicated to specific tasks. These may or may not contain similar computational components hidden from the users' view. The user would find and access these applications using search tools such as those presently used in the World Wide Web.

GMS contains a number of novel solutions to automated modelling of *ad hoc* data. Organizing the structure and functions of the system to anticipate and deal with data problems is successful in producing a robust system. It also provides a framework for assessing, and improving the system. Future research is aimed at further quantifying the impact of the problems and solutions. By examining modelling systems in this way, alternative solutions to problems other than those adopted might be suggested, and possibly incorporated into future systems.

assistance during the project. The comments from anonymous referees assisted in revising the manuscript. Townsend Peterson supplied the Mexican Bird data.

References

ABEL, D. J., YAP, S. K., ACKLAND, R., CAMERON, M. A., SMITH, D. F., and WALKER, G., 1992, Environmental decision support system project: an exploration of alternative architectures for geographical information systems. *International Journal of Geographical Information Systems*, **6,** 193–204.

ASPINALL, R., 1992, An inductive modelling procedure based on Bayes theorem for analysis of pattern in spatial data. *International Journal of Geographical Information Systems*, **6,** 105–121.

AUSTIN, M. P., and SMITH, T. M., 1989, A new model for the continuum concept. *Vegetatio*, **83,** 35–47.

BETHKE, A. D., 1981, Genetic algorithms as function optimizers, PhD Thesis, Department of Computer and Communication Sciences, University of Michigan.

BLACK, U. D., 1991, *OSI: a model for computer communications standards* (New Jersey: Prentice Hall Inc.), pp. 07632.

BOSTON, T., and STOCKWELL, D. R. B., 1994, Interactive species distribution reporting, mapping and modelling using the World Wide Web. *Computer Networks and ISDN Systems*, **28,** 231–238.

BRINDLE, A., 1981, Genetic algorithms for function optimization. PhD Thesis, Computer Science Department, University of Alberta.

BURROUGH, P. A., 1992, Development of intelligent geographical information systems. *International Journal of Geographical Information Systems*, **6,** 1–11.

DAVEY, S. M., STOCKWELL, D. R. B., and PETERS, D. G., 1995, Intelligent systems: their use in managing biological diversity. *AI Applications*, **9,** 69–89.

DEJONG, K. A., 1975, Analysis of the behavior of a class of genetic adaptive systems. PhD Thesis, Department of Computer and Communication Sciences, University of Michigan, 1975.

DEJONG, K. A., 1980, Adaptive system design: a genetic approach. *IEEE Transactions Systems, Manufacturing, and Cybernetics*, **SMC-10,** 566–574.

GOLDBERG, D. E., 1989, Genetic Algorithms in Search, Optimization and Machine Learning, Addison-Wesley.

GOODCHILD, M., HAINING, R., and WISE, S., 1992, Integrating GIS and spatial data analysis: problems and possibilities. *International Journal of Geographical Information Systems*, **6,** 407–423.

HOLLAND, J. H., 1975, *Adaptation in Natural and Artificial Systems* (Ann Arbor: University of Michigan Press).

JAMES, F. C., and MCCULLOCH, C. E., 1990, Multivariate analysis in ecology and systematics: panacea or Pandora's box? *Annual Review of Ecology and Systematics*, **21,** 129–166.

LYNDENMAYER, D., NIX, N., TANTON, M., CUNNINGHAM, R., NORTON, T., and STOCKWELL, D., 1991, A hierarchical framework for the spatial and temporal analysis of habitat: An example using a rare and endangered species. In: *Proceedings of the 9th Biennial Conference on Modelling and Simulation, Simulation Society of Australia*, Gold Coast, Queensland, Australia, MSSANZ, Modelling and Simulation Society of Australia and New Zealand Inc., pp. 434–441.

NIX, H. A., 1986, A biogeographic analysis of Australian elapid snakes. In: *Atlas of Elapid Snakes of Australia*, edited by R. Longmore, Australian Flora and Fauna Series Number 7 (Canberra: Australian Government Publishing Service), pp. 4–15.

PUTZ, S., 1994, Interactive information services using World-Wide Web hypertext. *CERN: First International Conference on the World-Wide Web, 25–27 May, 1994.*

SKIDMORE, A. K., GAULD, A., and WALKER, P., 1996, Classification of kangaroo distribution using three GIS models. *International Journal of Geographical Information Systems*, **10,** 441–454.

STOCKWELL, D. R. B., 1992, Machine learning and the problem of prediction and explanation in ecological modelling. Doctoral Thesis, Australian National University, Australia.

STOCKWELL, D. R. B., 1993, LBS: Bayesian learning system for rapid expert system develop-
      ment. *Expert Systems With Applications*, **6,** 137–147.
STOCKWELL, D. R. B., 1995, Evaluation and Training in the Use and Interpretation of GARP
      Genetic Algorithms for Prediction. Consultancy Report to Australian Nature
      Conservation Agency.
STOCKWELL, D. R. B., 1997, Generic predictive systems: an empirical evaluation using the
      Learning Base System (LBS). *Expert Systems With Applications*, **12,** 301–310.
STOCKWELL, D. R. B., DAVEY, S. M., DAVIS, J. R., and NOBLE, I. R., 1990, Using induction
      of decision trees to predict greater glider density. *AI Applications*, **4,** 33–43.
STOCKWELL, D. R. B., and NOBLE, I. R., 1991, Induction of sets of rules from animal
      distribution data: a robust and informative method of data analysis. *Mathematics and
      Computers in Simulation*, **32,** 249–254.
VERBYLA, D. L., and LITVAITIS, J. A., 1989, Resampling methods for evaluating class accuracy
      of wildlife habitat models. *Environmental Management*, **13,** 783–787.