Available online at www.sciencedirect.com

SCIENCE DIRECT®

Ecological Modelling 170 (2003) 219–226

ECOLOGICAL MODELLING

www.elsevier.com/locate/ecolmodel

Using regression trees to identify the habitat preference of the sea cucumber (*Holothuria leucospilota*) on Rarotonga, Cook Islands

Sašo Džeroski [a,*], Darrin Drumm [b]

[a] *Department of Intelligent Systems, Jožef Stefan Institute, Jamova 39, Ljubljana 1000, Slovenia*
[b] *Department of Marine Science, University of Otago, P.O. Box 56, Dunedin, New Zealand*

Abstract

In the Pacific Islands, invertebrates including sea cucumbers are among the most valuable and vulnerable inshore fisheries resources. As human activities continue to force substantial impacts on coral reef ecosystems, the management of inshore fisheries has become an increasingly important priority. Knowledge of the distribution, biology and habitat requirements of a species can significantly enhance conservation efforts. The sea cucumber (*Holothuria leucospilota*) forms an important part of the traditional subsistence fishery on Rarotonga, Cook Islands, yet little is known of this species' present spatial distribution and abundance around the island.

We apply two machine learning approaches and a classical statistical approach to predict the number of sea cucumber individuals from site characteristics. The machine learning methods used are induction of regression trees and instance-based learning. These are compared to the classical statistical approach of linear regression. The most accurate predictions are obtained using instance-based learning, while the most understandable descriptions are obtained using regression tree induction.

Keywords: Machine learning; Tropical marine ecology; Habitat preference; Sea cucumber

1. Introduction

In the Pacific Islands, invertebrates including sea cucumbers are among the most valuable inshore fisheries resources (Dalzell et al., 1996). The multi-species tropical sea cucumber fishery throughout the Pacific Ocean has existed for thousands of years. However, unsustainable harvesting rates can contribute to local species depletions and/or extinctions. As human activities continue to force substantial impacts on coral reef ecosystems, the management of inshore fisheries in the Pacific Islands is becoming an increasingly important priority. Effective management plans must be developed for these fisheries.

This paper describes an investigation into the habitat conditions preferred by one species of sea cucumber (*Holothuria leucospilota*) on Rarotonga, Cook Islands. If favourable habitat conditions can be effectively identified, steps can be taken to preserve these areas and help the population to thrive. We examine three different analytical modelling techniques that can be used to predict favourable habitat conditions for the sea cucumber. These three modelling techniques are applied to the same dataset and the resulting models are compared in terms of their overall effectiveness for "good" habitat prediction.

The Cook Islands consist of 15 small islands covering a large area of the tropical South Pacific Ocean, and are divided into a northern and a south-

* Corresponding author. Tel.: +386-1-477-3217; fax: +386-1-425-1038.
E-mail address: Saso.Dzeroski@ijs.si (S. Džeroski).

0304-3800/$ – see front matter © 2003 Elsevier B.V. All rights reserved.
doi:10.1016/S0304-3800(03)00229-1

ern group. Rarotonga is a tropical, extinct volcanic island and it lies in the southern group at approximately latitude 21°12′ S and longitude 159°48′ W. It is the most populated and largest of the Cook Islands (11.5 km East to West and 8 km North to South) and is encircled by a well-developed, lagoonal ecosystem. The study area includes the entire shallow-water ecosystem of the island of Rarotonga, Cook Islands.

The Rarotonga lagoon supports an extensive subsistence fishery that targets many invertebrate species, including sea cucumbers, sea urchins, giant clams and trochus. This traditional fishery provides a substantial component of the protein requirements for the Rarotongan people. *Holothuria leucospilota* is the most heavily targeted species of the traditional sea cucumber fishery, yet little is known of this species' present spatial distribution and abundance around the island. Sperduto and Congalton (1996) recognise that knowledge of the biology, distribution and habitat requirements of a species are important elements necessary for its conservation.

Ecosytems characteristically exhibit highly complex non-linear relationships between their associated variables. Statistical analysis of these multivariable, non-linear systems using conventional linear-regression statistical methods often does not provide solutions that meet the goals of the investigators, and so there is always an interest in exploring new solution techniques. Machine learning methods, including regression tree induction and instance-based learning, seem to offer an advantage over traditional linear-regression analysis techniques, because they do not introduce any prior assumptions about the relationships between the variables. Machine learning methods have an inherent ability to discover patterns in the data that are not possible to detect using conventional linear-regression models and are increasingly popular techniques for analysing ecological datasets (Fielding, 1999; Lek and Guegan, 1999; Recknagel, 2001).

In this paper, we apply two machine learning methods, regression tree induction and instance-based learning, to the problem of modelling habitat suitability for the sea cucumber (*H. leucospilota*) and compare the results to those of the classical statistical approach. In particular, we address the problem of predicting the number of individuals of the species present in a strip of seabed, based on environmental characteristics of the strip.

The remainder of this paper is organised as follows. Section 2 describes the machine learning methods used. Section 3 describes the data analysed. Section 4 presents the data analysis results, while Section 5 discusses these results and concludes.

2. Machine learning methods

In the following two subsections, we describe the two machine learning methods used to predict the number of sea cucumber individuals from site characteristics. Linear regression, which is also used for comparison purposes, is a standard statistical method: as such it is treated in most textbooks on statistics and will not be described here.

2.1. The M5′ program for inducing regression trees

Regression trees are a representation for piece-wise constant or piece-wise linear functions. Like classical regression equations, they predict the value of a dependent variable (called class) from the values of a set of independent variables (called attributes). Data represented in the form of a table can be used to learn or automatically construct a regression tree. In the table, each row (example) has the form $(x_1, x_2, \ldots, x_N, y)$, where x_i are values of the N attributes (e.g. site characteristics, such as percentages of sand, rubble, etc.) and y is the value of the class (e.g. the number of sea cucumber individuals).

Unlike classical regression approaches, which find one single equation for a given set of data, regression trees partition the space of examples into axis-parallel rectangles and fit a model to each of these partitions. A regression tree has a test in each inner node that tests the value of a certain attribute, and in each leaf a model for predicting the class: the model can be a linear equation or just a constant. Trees that can have linear equations in the leaves are also called model trees.

Given a new example for which the value of the class should be predicted, the tree is interpreted from the root. In each inner node, the prescribed test is performed and according to the result of the test the corresponding left or right sub-tree is selected. When the selected node is a leaf, then the value of the class for the new example is predicted according to the model in the leaf.

Tree construction proceeds recursively starting with the entire set of training examples (entire table). At each step, the most discriminating attribute is selected as the root of the (sub)tree and the current training set is split into subsets according to the values of the selected attribute. For discrete attributes, a branch of the tree is typically created for each possible value of the attribute. For continuous attributes, a threshold is selected and two branches are created based on that threshold.

Technically speaking, the most discriminating discrete attribute or continuous attribute test is the one that reduces most the variance of the values of the class variable. For continuous attributes, the values of the attribute that appear in the training set are considered as thresholds. For the subsets of training examples in each branch, the tree construction algorithm is called recursively. Tree construction stops when the variance of the class values of all examples in a node is small enough (or if some other stopping criterion is satisfied). Such nodes are called leaves and are labelled with a model (constant or linear equation) for predicting the class value.

An important mechanism used to prevent trees from over-fitting data is tree pruning. Pruning can be employed during tree construction (pre-pruning) or after the tree has been constructed (post-pruning). Typically, a minimum number of examples in branches can be prescribed for pre-pruning and confidence level in error estimates in leaves for post-pruning.

A number of systems exist for inducing regression trees from examples, such as CART (Breiman et al., 1984) and M5 (Quinlan, 1992). M5 is one of the most well-known programs for regression tree induction. We used the system M5′ (Wang and Witten, 1997), a re-implementation of M5 within the software package WEKA (Witten and Frank, 1999). The parameters of M5′ were set to their default values, except for the pruning parameter, which was set to 1.0 (default 2.0), thus pruning trees slightly less than usual.

2.2. Instance-based learning

Instance-based learning (IBL) algorithms (Aha et al., 1991) use specific instances to perform classification tasks, rather than using generalisations, such as regression trees. IBL algorithms are also called lazy learning algorithms, as they simply save some or all of the training examples and postpone all effort towards inductive generalisation until classification time. They assume that similar instances have similar classifications: novel instances are classified according to the classifications of their most similar neighbours.

IBL algorithms are derived from the nearest neighbour pattern classifier (Cover and Hart, 1968). The nearest neighbour (NN) algorithm is one of the best known classification algorithms and an enormous body of research exists on the subject (Dasarathy, 1990). In essence, the NN algorithm treats attributes as dimensions of a Euclidean space and examples as points in this space. In the training phase, the classified examples are stored without any processing. When classifying a new example, the Euclidean distance between that example and all training examples is calculated and the class of the closest training example is assigned to the new example.

The more general *k*-NN method takes the *k* nearest training examples and determines the class of the new example by majority vote for discrete classes and by taking the average for continuous classes. In improved versions of the *k*-NN method, the votes of each of the *k* nearest neighbours are weighted by the respective proximity to the new example. We used the algorithm IBk, as implemented in the software package WEKA (Witten and Frank, 1999). The parameter settings were left at their default values, except for *k*, which was set to 7: the class values of the seven nearest neighbours of a new example were averaged to obtain a prediction of the class value for the new example.

3. Spatial data preparation

A geographic information system (GIS) was used to facilitate the sampling design and the spatial analysis of the physical habitat and biological attributes. The habitats of the shallow-water (backreef and lagoon) ecosystem were classified into the following broad morphologically homogeneous zones: reef rim, rubble/rock, sand/coral matrix, sand, mudflat and passage/harbour (Fig. 1). The boundaries of these zones were delineated by the spectral signature of aerial photographs and were digitised in the GIS to produce detailed habitat maps. The habitat classifications were later confirmed by ground truthing. Sample sites were randomly located and their allocation was weighted

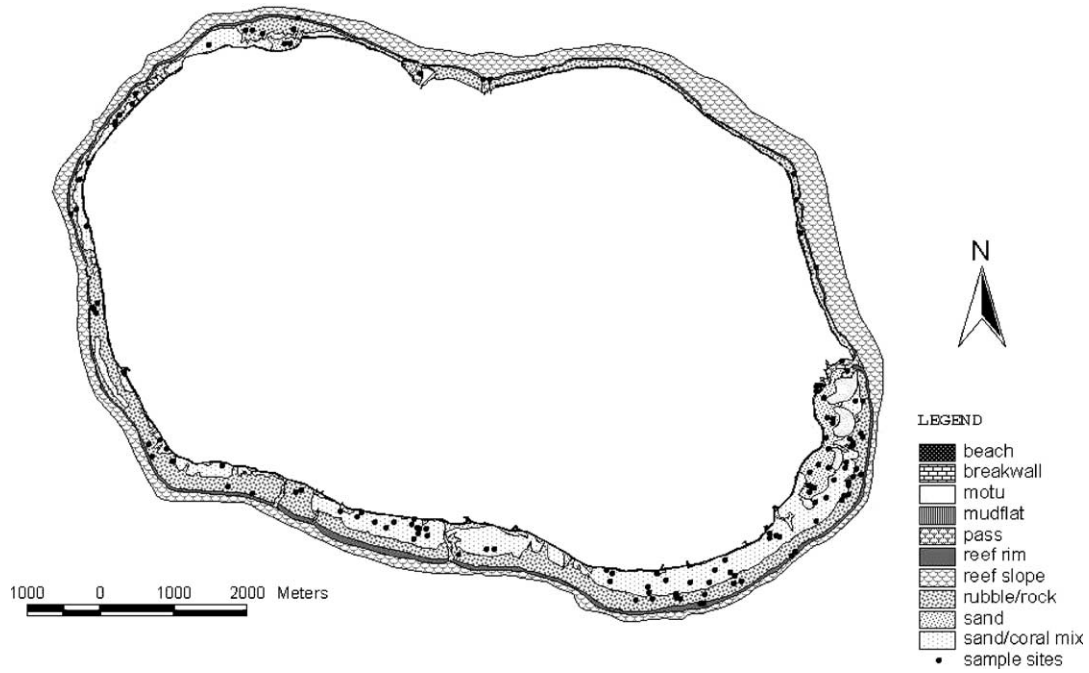

Fig. 1. Broad reef-top habitat zonation and sample site location on Rarotonga, Cook Islands.

according to the relative percentage of each habitat zone from the GIS spatial model.

A total of 128 sites were sampled for environmental and biological variables, using 2 m × 50 m (100 m^2) strip transects. This size sample unit was selected to account for the patchy distribution of the animals, and the number of *H. leucospilota* individuals encountered along each transect was recorded. In addition to the species abundance, 10 environmental variables that were expected to have an influence on the habitat preference of the sea cucumber were recorded at each of the 128 locations. These included the exposure of the site (windward or leeward side of the island), and the following microhabitat variables, estimated as a percentage (with possible values from 0 to 100%) of the total 100 m^2 area sampled: Sand, Rubble, Cons_Rubble (consolidated rubble), Boulder, reef rock/pavement (Rock_Pave), live coral (Live_Coral), dead coral (Dead_Coral), mud/silt (Mud_Silt), and Gravel.

The microhabitat variables were estimated as a percentage of the total 100 m^2 area sampled. The location of each sample site was mapped using post-processed differentially corrected global positioning system (GPS) data to provide an accurate spatial model that integrated the environmental and ecological conditions present at each site. The spatial dataset containing these variables was then prepared for analysis with the machine learning approaches that we investigated.

We predict the number of sea cucumber individuals from the site habitat characteristics. The higher the predicted number of individuals, the more suitable the habitat. The distribution of the number of individuals is given in Table 1.

Earlier work (Zhou et al., 2000) categorised each of the observed sites into two distinct classes according

Table 1
The distribution of the number of sea cucumber individuals

Number of sites (patterns)	Animal frequency range
89	0–50
15	51–100
9	101–200
11	201–500
4	501–1000
128 total	0–1000 overall

to the frequency of animals at each location within the 100 m^2 sample unit: the "average condition class" was characterised by having a density of <1 animal/m^2, while the "good condition class" was attributed to sites with >1 animal/m^2.

4. Analysis of the generated models and their predictive accuracy

For each of the three prediction techniques outlined in Section 2, we describe the predictive models obtained. Where appropriate, i.e. for regression trees and linear regression, we explain and comment on the models generated on the entire dataset of 128 sites.

We also present the results achieved in terms of predictive accuracy of these models on unseen cases, as estimated by 10-fold cross-validation. In 10-fold cross-validation, the dataset is split into 10 disjoint subsets of approximately the same size, and 10 experiments are performed. In each of these, 1 of the 10 subsets is withheld, the prediction method trained on the union of the remaining 9, then tested on the unseen examples from the withheld subset. The reported accuracies are the averages of the 10 experiments. Three measures of performance are considered: the multiple correlation coefficient (R), mean absolute error (MAE) and root mean squared error (RMSE). The results are summarised in Table 2.

Table 2
The predictive accuracy (error) of the three prediction approaches on unseen cases as estimated by onefold cross-validation, evaluated by three measures of error between the actual and predicted number of see cucumber individuals (see also text)

Method/measure	R	MAE	RMSE
Linear regression	0.55	64.51	116.93
M5′	0.50	62.06	121.74
IBk	0.58	66.81	116.96

4.1. The M5′ regression tree

The performance of the regression tree induced by M5′ from the given dataset on unseen cases (as estimated by 10-fold cross-validation) is as follows: the R is 0.50, the MAE is 62.06 and the RMSE is 121.74. The regression tree induced by the M5′ program on the entire dataset of 128 sites is given in Table 3.

The tree can be used to predict the number of sea cucumber individuals for a new site from its characteristics as follows:

- If the site exhibits less than or equal to 9.5% of rubble, the linear model LM1 is used. This model predicts the number of individuals as a linear combination of the variables Rubble, Cons_Rubble and Live_Coral: the number of individuals increases as the values of each of these variables increase (all the coefficients in the model are positive). The average number of individuals for the 69 sites to which LM1 applies is 15: these sites are examples of poor habitat for *H. leucospilota*.
- If the site exhibits more than 9.5% of rubble, less than or equal to 8.5% sand, and less than or equal to 15% reef rock/pavement, the linear model LM2 is used. This model uses the variables Sand, Rubble and Cons_Rubble: the number of individuals decreases as the percentage of sand increases, while increasing as the percentages of rubble and consolidated rubble increase. LM2 applies to 28 sites with (on average) 236 individuals per site: these sites represent optimal habitat for the sea cucumber.
- If the site exhibits more than 9.5% of rubble, less than or equal to 8.5% sand, but more than 15% reef rock/pavement, LM3 is used. This model uses only the variables Rubble and Cons_Rubble: the number of individuals increases as the percentages of rubble and consolidated rubble increase. The average number of individuals at the 14 sites to which LM3

Table 3
A regression tree for predicting the frequency (number of see cucumber individuals)

Rubble ≤ 9.5: (LM1) Freq = 6.54 + 0.423 × Rubble + 2.24 × Cons_Rubble + 0.408 × Live_Coral
Rubble > 9.5:
  Sand ≤ 8.5:
    Rock_Pave ≤ 15: (LM2) Freq = 148 − 17.2 × Sand + 1.06 × Rubble + 1.21 × Cons_Rubble
    Rock_Pave > 15: (LM3) Freq = 74.1 + 1.06 × Rubble + 1.21 × Cons_Rubble
  Sand > 8.5: (LM4) Freq = 5.33 + 1.51 × Rubble + 1.73 × Cons_Rubble

Table 4
A linear equation for predicting the number of see cucumber individuals (a) and comments on the consistency of the coefficient values with existing expert knowledge (b)

(a)

$$\text{Freq} = -32.8764 \times \text{Exposure} - 0.2826 \times \text{Sand} + 2.1727 \times \text{Rubble} + 2.4133 \times \text{Cons\_Rubble} - 0.5938 \times \text{Boulder} - 0.2135 \times \text{Rock\_Pave} + 0.2132 \times \text{Live\_Coral} - 0.1771 \times \text{Dead\_Coral} - 0.2174 \times \text{Mud\_Silt} - 0.5987 \times \text{Gravel} + 32.9036$$

(b)

Exposure:	Windward side gets more impact from waves, more rubble created by breaking off coral.
Sand:	Too much sand deprives the sea cucumber of crevices and interstitial spaces (and thus the protection from predators.)
Rubble, consolidated rubble:	These provide the necessary hiding places for protection from predators.
Boulder:	Expert thinks they provide suitable habitat (rubble on a large scale), although a large boulder would provide less habitat than the same surface of rubble. The negative coefficient does not reflect this clearly.
Rock pavement:	Does not provide suitable habitat; coefficient confirms this.
Live coral, dead coral:	Expert thinks they both provide suitable habitat by providing a canopy and some protection. Coefficients in equation confirm this for live coral, but not for dead coral.
Mud, silt, gravel:	Do not provide suitable habitat; coefficients confirms this.

applies is 81: these sites represent good habitat (but not optimal).

- Finally, if the site exhibits more than 9.5% rubble and more than 8.5% sand, the linear model LM4 applies, which uses the same variables as LM3, but has slightly larger coefficients. The average number of individuals at the 17 sites to which LM4 applies is 35: these sites represent poor (but not very poor) habitat.

4.2. Instance-based learning

The performance of instance-based learning on unseen cases (as estimated by 10-fold cross-validation) is as follows: the R is 0.58, the MAE is 66.81 and the RMSE is 116.96. The higher correlation coefficient and lower RMSE would indicate better predictive power than for regression trees, but note that the mean absolute error is higher for IBL than for regression trees. Instance-based learning does not generalise from the training examples, it just stores them for future reference. Thus, no model is generated.

4.3. The linear-regression equation

The linear equation for predicting the number of individuals, obtained using the standard linear-regression method, is given in Table 4(a).

A positive coefficient in front of a variable means that (all other variables being constant) the variable positively influences the number of sea cucumber individuals: as the value of the variable increases, the number of individuals is predicted to increase as well. Conversely, a negative coefficient means a negative influence: as the variable increases, the number of individuals is predicted to decrease (all other variables being constant).

From the coefficients in the equation above, we can see that rubble, consolidated rubble and live coral positively influence the number of sea cucumber individuals. All the other variables influence the frequency of sea cucumbers negatively. Comments on the consistency of the coefficient values with existing expert knowledge are given in Table 4(b).

The linear-regression model has the smallest estimated RMSE, while it scores between M5′ and IBk on the other two performance measures.

5. Discussion and conclusions

The sea cucumber, *H. leucospilota*, occupies a distinct ecological niche. It prefers areas with a larger grained physical structure, such as areas of rubble, consolidated rubble and boulder, rather than a fine-grained structure like sand. These preferred substrate types offer the necessary interstitial spaces and cover required for protection. Due to the strict habitat requirements of this species, one would expect rubble and consolidated rubble to have positive effects on the

abundance of animals and sand to negatively influence the frequency, and this is reflected in the generated rules, both for linear regression and regression trees.

All three approaches used achieve similar levels of predictive accuracy. Instance-based learning has the best *R*, linear regression has the best RMSE and M5′ has the lowest MAE. None of the accuracies are overwhelmingly high, but we should bear in mind that the primary goal of our analysis is to determine the essential influences of site characteristics on sea cucumber habitat, rather than exactly predict the number of sea cucumber individuals.

M5′ identifies the most important influences of the site characteristics on habitat suitability (rubble, consolidated rubble, live coral, rock pavement and sand). Of these, rubble and sand are most important, as they appear in the first and second level of the regression tree, respectively. It also identifies four types of sites and constructs different linear models to predict the number of sea cucumbers at each type of site. Each leaf of the regression tree corresponds to one site type. The sea cucumbers prefer larger percentages of rubble and consolidated rubble in all four types of sites (positive coefficients for rubble/consolidated rubble in each of the four linear models).

Two of the site types are essentially not very suitable as sea cucumber habitat: the first does not have enough rubble (less than or equal to 9.5%), while the second does have enough rubble, but too much sand (greater than 8.5%). The average number of individuals recorded at the two types of sites are 15 and 35, respectively. One site type is very suitable as sea cucumber habitat, as evidenced by the average of 236 animals found per site. This type of site is characterised by enough rubble (greater than 9.5%), little sand (less than or equal to 8.5%) and little rock pavement (less than or equal to 15%). Within this type of site, sea cucumbers prefer less sand. The last type of site represents a moderately suitable habitat for sea cucumbers: it has the same characteristics as the most suitable habitat, except for too much rock pavement (greater than 15%): 81 individuals (on average) are found at sites of this type.

The advantages of the regression tree generated by M5′ as compared to the linear-regression model are as follows. It autonomously identifies the four different types of site and constructs a model for each of these: this is also an advantage of previous work (Zhou et al., 2000), which categorised each of the observed sites by setting an arbitrary threshold of 1 animal/m^2. It focuses on the most important influences of site characteristics on sea cucumber frequency, rather than using all of the given variables. It is completely consistent with expert knowledge (both concerning the structure of the tree/importance of variables and the influence of the variables in individual models on the frequency), unlike linear regression which constructs one model for all sites and has problems with some of the variables (it is not straightforward to explain the coefficients for boulder and dead coral).

In further work, more data of better quality should be collected and analysed to obtain better models of habitat suitability for the sea cucumber (*H. leucospilota*). In particular, additional relevant variables should be measured, such as turbidity and water temperature. Information and data on predators should also be taken into account.

Acknowledgements

The authors would like to thank Tamsyn Dearlove, Grant Hopkins and Naomi Sugimura for their assistance with the field data collection. We also gratefully acknowledge the financial support provided by the Pacific Development and Conservation Trust, World Wide Fund for Nature (WWF), and the New Zealand Official Development Assistance (NZODA) programme. Generous donations of equipment were made by Pacific Kayak Ltd., Auckland; Divers World, Wellington, New Zealand; and Sokkia New Zealand. Tom Taranto at CSIRO Marine Research, Cleveland, Australia, provided invaluable guidance and instruction with the GIS component of the research. Thanks also to Cook Islands Land Information Services, Government of the Cook Islands and the Cook Islands Meteorological Service. Darrin Drumm was supported by a University of Otago, Ph.D. scholarship during the field data collection. For comments and suggestions on this work, thanks are also due to Friedrich Recknagel.

References

Aha, D., Kibler, D.W., Albert, M.K., 1991. Instance-based learning algorithms. Machine Learn. 6, 37–66.

Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J., 1984. Classification and Regression Trees. Wadsworth, Belmont.

Cover, T.M., Hart, P.E., 1968. Nearest neighbor pattern classification. IEEE Trans. Inf. Theory 13, 21–27.

Dalzell, P., Adams, T.G.H., Polunin, N.V.C., 1996. Coastal fisheries in the Pacific Islands. Oceanogr. Marine Biol. Ann. Rev. 34, 395–531.

Dasarathy, B.V. (Ed.), 1990. Nearest Neighbor (NN) Norms: NN Pattern Classification Techniques. IEEE Computer Society Press, Los Alamitos, CA.

Fielding, A. (Ed.), 1999. Machine Learning Methods for Ecological Applications. Kluwer, Dordrecht.

Lek, S., Guegan, J.F. (Guest Eds.), 1999. Application of artificial neural networks in ecological modelling. Ecol. Model. 120 (2/3) (Special Issue).

Quinlan J.R., 1992. Learning with continuous classes. In: Proceedings of the Fifth Australian Joint Conference on Artificial Intelligence, 343–348. World Scientific Singapore.

Recknagel, F. (Guest Ed.), 2001. Application of machine learning to ecological modelling. Ecol. Model. 146 (1–3) (Special Issue).

Sperduto, M.B., Congalton, R.G., 1996. Predicting rare orchid (small whirled pogonia) habitat using GIS. Photgramm. Eng. Rem. Sens. 62 (11), 1269–1272.

Wang, Y., Witten, I.H., 1997. Induction of model trees for predicting continuous classes. In: Proceedings of the Poster Papers of the European Conference on Machine Learning. Faculty of Informatics and Statistics, University of Economics, Prague.

Witten, I.H., Frank, E., 1999. Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations. Morgan Kaufmann, San Francisco.

Zhou, Q., Drumm, D., Purvis, M., 2000. Adaptive knowledge discovery techniques for identifying the habitat preference for the sea cucumber, *H. leucospilota*, on Rarotonga, Cook Islands. In: Recknagel, F. (Ed.), Book of Abstracts, 2nd International Conference on Applications of Machine Learning to Ecological Modelling. University of Adelaide, Australia.