Ecological Modelling 222 (2011) 1657–1665

Contents lists available at ScienceDirect

Ecological Modelling

journal homepage: www.elsevier.com/locate/ecolmodel

Determining factors that influence the dispersal of a pelagic species: A comparison between artificial neural networks and evolutionary algorithms

D.R. Pontin[a,*], S. Schliebs[b], S.P. Worner[c], M.J. Watts[d]

[a] Department of Ecology, Lincoln University, Canterbury, New Zealand
[b] Knowledge, Engineering and Discovery Research Institute, Auckland University of Technology, Auckland, New Zealand
[c] Bio-Protection Research Centre, Lincoln University, Canterbury, New Zealand
[d] School of Earth and Environmental Sciences, University of Adelaide, Adelaide, Australia

ARTICLE INFO

Article history:
Received 28 July 2010
Received in revised form 9 January 2011
Accepted 3 March 2011

Keywords:
Ecological informatics
Artificial neural networks
Evolutionary algorithms
Physalia
Feature selection

ABSTRACT

Because of increasing transport and trade there is a growing threat of marine invasive species being introduced into regions where they do not presently occur. So that the impacts of such species can be mitigated, it is important to predict how individuals, particularly passive dispersers are transported and dispersed in the ocean as well as in coastal regions so that new incursions of potential invasive species are rapidly detected and origins identified. Such predictions also support strategic monitoring, containment and/or eradication programs. To determine factors influencing a passive disperser, around coastal New Zealand, data from the genus *Physalia* (Cnidaria: Siphonophora) were used. Oceanographic data on wave height and wind direction and records of occurrences of *Physalia* on swimming beaches throughout the summer season were used to create models using artificial neural networks (ANNs) and Naïve Bayesian Classifier (NBC). First, however, redundant and irrelevant data were removed using feature selection of a subset of variables. Two methods for feature selection were compared, one based on the multilayer perceptron and another based on an evolutionary algorithm. The models indicated that New Zealand appears to have two independent systems driven by currents and oceanographic variables that are responsible for the redistribution of *Physalia* from north of New Zealand and from the Tasman Sea to their subsequent presence in coastal waters. One system is centred in the east coast of northern New Zealand and the other involves a dynamic system that encompasses four other regions on both coasts of the country. Interestingly, the models confirm, molecular data obtained from *Physalia* in a previous study that identified a similar distribution of systems around New Zealand coastal waters. Additionally, this study demonstrates that the modelling methods used could generate valid hypotheses from noisy and complicated data in a system about which there is little previous knowledge.

1. Introduction

Researchers in ecology face an ever growing influx of data as well as a range of methods to analyse that data and make predictions. They are also obliged to minimise the error and uncertainty of any model created, while maximising potential insights into the workings of the target system. Standard statistical methods such as Generalised Linear Models and Bayesian statistics (Bolker, 2008) have often been used to successfully analyse challenging ecological data. However, the appropriate statistical method to analyse any given dataset should be determined based on prior knowledge of the system so that the parameters included in the analysis are relevant (Hochachka et al., 2007). With novel systems, lack of prior knowledge may affect the validity of the analysis and/or any insights gained because of the use of sub-optimal parameters. Although many conventional statistical methods exist for parameter exploration and feature selection (Burnham and Anderson, 2002) data mining techniques have often been shown to be more powerful and efficient (Segurado and Araujo, 2004; Elith et al., 2006; Virkkala et al., 2010) especially when the goal is to maximise the classification accuracy of the predictions as well as generate testable hypothesis from the data.

Data mining includes a number of analysis techniques that include artificial neural networks (Rumelhart et al., 1986), evolutionary algorithms (Baluja, 1994), support vector machines (Cortes and Vapnik, 1995) and decision trees (Murthy, 1998). Despite being very different, these analytical methods or algorithms are characterised by few statistical restrictions on the form of the input data or limitations on the way such data are processed. Despite the lack of restrictions these methods have been shown to provide models that

* Corresponding author at: Bio-Protection Research Centre, Lincoln University, PO Box 84, Lincoln, Canterbury 7647, New Zealand. Tel.: +64 3 325 3838; fax: +64 3 325 3864.
E-mail address: david.r.pontin@gmail.com (D.R. Pontin).

0304-3800/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.ecolmodel.2011.03.002

can maximise predictive performance in the context of ecological problems (Hochachka et al., 2007). However, to maximise predictive performance and model ability to generalise to new data often requires feature selection or variable selection to remove redundant data. Fortunately, all these methods can be used to assess the relationship between predictor and responses variables to identify important predictor variables and thus ensure a robust model.

Historical species occurrence records are being increasingly used in ecological modelling efforts as they are transferred to electronic format (Elith et al., 2006). While many modelling techniques are extremely flexible and can overcome many issues with the data, the long held adage that the quality and quantity of the data has a considerable impact on the effectiveness of the model, remains true (Zhang et al., 2003; Fogel, 2008). As with all historical species records there is often an inherent bias in the collection of the records, as the sampling is rarely carried out in a way appropriate to meet the objectives of a particular sampling program. For example, in marine systems, inshore areas are often sampled at a higher rate than offshore areas despite that the species of interest may be more widespread (Ready et al., 2010). Also because of often inadequate sampling, records may misrepresent a species presence or absence in an area. For example, small sample sizes will fail to detect the presence of some species. Clearly, if a species is rarely encountered it can lead to an imbalance between the number of presence records compared with the number of records of species absence in the dataset. True absences are also most often never recorded and when they are, it is often not known whether the species is truly absent, or was by chance, or inadequate sampling design, not detected. Usually, modellers deal with pseudo-absences or potential absences, the number of which, are selected to represent the area to be modelled. When compared with the presence data the absence data set is most often the largest class. Class imbalance is a significant challenge as any modeller will seek to maximise predictive accuracy over the full range of instances modelled (Chen et al., 2008; Liu et al., 2008). The result is, a model that is able to classify the majority class accurately (the absences) but has a poor ability to classify the minority class (Xu and Chow, 2006; Chen et al., 2008) greatly limits inferences about the relationship between predictor and responses variables. Clearly, it is just as important to achieve balanced data as well as use only those variables considered important in the system to maximise model performance.

In this study we compare the feature selection properties of two machine learning approaches, a multilayer perceptron (MLP) and a type of evolutionary algorithm, the Versatile Quantum-inspired Evolutionary Algorithm (vQEA) as proposed in (Defoin-Platel et al., 2007) to select features from a complex and noisy ecological data set. An artificial neural network is potentially a powerful technique for modelling species populations and can be used to identify key factors that influence those populations (Joy and Death, 2004; Olden et al., 2004) especially for species that live in complex and changing environments. ANN have often been shown to outperform standard statistical techniques when applied to complex data (Lek et al., 1996; Brosse et al., 1999; Mutanga and Skidmore, 2004). Many studies have shown that MLP, in particular, do well modelling problems with noisy and complex data (Lek et al., 1996; Olden and Jackson, 2002; Joy and Death, 2004) however, their use in ecology is still not widely accepted. Evolutionary algorithms have had some acceptance and application in ecology, for example predicting species distributions using the Genetic Algorithm for Rule Set Production (GARP) (Stockwell and Peters, 1999), or to model nuisance algal blooms in coastal ecosystems (Muttil and Lee, 2005). Recently, Quantum-inspired Evolutionary Algorithms (QEAs) have been introduced (Han and Kim, 2004) and successfully applied to combinatorial benchmark problems. Following some of the principles of quantum computing, QEA employs a probabilistic model to efficiently explore a binary search space. Defoin-Platel et al. (2007) proposed an improved version of the method, namely the Versatile Quantum-inspired Evolutionary Algorithm (vQEA). The Versatile Quantum-inspired Evolutionary Algorithm was found to significantly outperform both a classical genetic algorithm and QEA when using traditional combinatorial benchmark problems (Defoin-Platel et al., 2009). Based on the wrapper approach proposed by Kohavi and Sommerfield (1995), vQEA was combined with a classification method, namely the Naïve Bayes Classifier (NBC) (Friedman et al., 1997), and successfully applied to an ecological modelling problem (Schliebs et al., 2009). In order to maximize the classification accuracy of the wrapper, vQEA evolves a suitable feature subset, while the NBC acts as the fitness function evaluating the quality of the selected features. Due to its low computational cost, NBC is very suitable to use with the wrapper context, since the evolutionary process requires the evaluation of many potential solution candidates. Although assumed to be less accurate than MLP (Kotsiantis et al., 2006), NBC is often surprisingly competitive and can even outperform some state-of-the-art algorithms on specific problems (Domingos and Pazzani, 1997; Kotsiantis, 2007).

Complete datasets of the occurrence or abundance of marine species within their associated ecological systems are rare, and usually only available for well studied species (Ready et al., 2010). Recently, there has been interest in investigating the genetic linkages between marine populations because of the historical belief that oceans provide little barrier to gene flow (Palumbi, 1992; Knowlton, 2000; Dawson and Jacobs, 2001). Of particular interest are species that rely on passive dispersal as environmental conditions combined with the larval duration determine their dispersal patterns (Cowen et al., 2000, 2006; Kinlan and Gaines, 2003). However, in general, marine ecologists have found it difficult to characterise dispersal patterns for any but those taxa dispersing over short distances because of the difficulty tracking and quantifying marine dispersal events (Kinlan and Gaines, 2003; Bradbury et al., 2008). A possible solution is the use of individual based models. A good example, is a Lagrangian particle tracking model described by White et al. (2010) used to simulate how a subtidal whelk (*Kelletia kelletii*), will potentially disperse in the Santa Barbara Channel, CA, USA. Moon et al. (2010) demonstrated, by simulating spawning and the movement of individuals of the giant jellyfish (*Nemopilema nomurai*) in the East China Sea, that the Lagrangian particle tracking model is also able to evaluate points of origin for a dispersal event. A primary requirement for the approach is that estimates of dispersal characteristics such as rate of drift of larva and their mortality rates are available (Levin, 2006). Such prior knowledge of the species biology and ecology is not often available.

A pelagic species such as the jellyfish *Physalia* sp. (Phylum Cnidaria: Siphonophora) provides an opportunity to compare methods to characterise features influencing dispersal of a passive dispersing marine species for which there is a small amount of occurrence data. *Physalia* is a classical passive disperser, relying totally on ocean winds, waves and currents and may be considered a proxy for other species of this nature. Dispersal is achieved by means of a pneumatophore (float) in which the gas cannot be regulated (Collins, 2002) so that this genus solely inhabits the surface of the ocean (Lane, 1960). Previous work indicated that ANN could be used to identify patterns occurrence in relation to oceanographic data (Pontin et al., 2008, 2009), however, only a limited area was modelled and it was difficult to generalise the results or judge how important the parameters were. The overall aim of this study was to compare the use of ANN and the vQEA to identify important predictor variables that drive *Physalia* occurrence in New Zealand waters. In this study *Physalia*, is used as an example of a marine passive disperser. Additionally, there is interest in predicting *Physalia* pres-

ence at New Zealand beaches as severe stings can produce nausea, vomiting, breathing difficulties and cardiovascular collapse, leading to possible death (Slaughter et al., 2009).

2. Materials and methods

2.1. Physalia occurrence data

Seventy-two surf lifesaving clubs in New Zealand patrol beaches during the Southern Hemisphere summer months from approximately late October until mid March. Surf lifeguards who treat members of the public for jellyfish stings are required to record these incidents. Since *Physalia* sp. is the only stinging jellyfish genus known in New Zealand waters (Slaughter et al., 2009), these records can be considered a proxy for their presence. We accessed the records of patrols carried out from the 2000/2001 season to the 2004/2005 season. Records that showed a beach head count of zero, that is there were no people on the beach, were excluded, because clearly there will be no jellyfish incidents if no one is swimming at the time. It must be noted that this data is not continuous but is dependent on when patrols were carried out, which was primarily over weekends but between mid December and the end of January, daily patrols were carried out.

2.2. Oceanographic data

Oceanographic data was sourced from National Oceanic and Atmospheric Administration/National Centers for Environmental Prediction (NOAA/NCEP) Wavewatch III model hindcast output (Tolman, 1998) represented eighty 1.25 × 1 degree global grid cells surrounding New Zealand. Each cell contained three-hourly measurements of five variables (significant wave height (m), peak wave period(s), peak wave direction (°N) and *U* and *V* wind vector components (ms^{-1})). All variables were transformed to daily data points, by averaging each of the eight data points for each day. Furthermore, from the *U* and *V* wind vector components, wind velocity (ms^{-1}) and direction were calculated. The circular mean (Fisher, 1995) was used for all directional variables. Once the transformations had been completed, each cell contained daily data for significant wave height (m), peak period (s), peak direction (°N), wind velocity (ms^{-1}) and wind direction (°N). The averaging of the oceanographic data was a necessity to ensure the same resolution of the explanatory and observed data. By reducing the resolution to daily data it is accepted that there would be a corresponding reduction in the presence of short duration phenomena. But it was expected that the underlying patterns would remain allowing the identification important predictor variables that drive *Physalia* occurrence. A variable to describe the fetch in each cell was not included because it would have been highly correlated with wind velocity and significant wave height. Moreover this relationship would be expected to be detected by the models as they are data driven rendering a fetch variable superfluous. Suitable data on speed and direction of surface currents was unavailable and was not incorporated into the models. Oceanographic data were extracted for five regions around coastal New Zealand (West Auckland, Bay of Plenty, Taranaki, Wellington and Canterbury) (Fig. 1) and combined with the *Physalia* occurrence data. For each region, data from a given cell were included if the cell was less than 250 km distant from the centre of the region. The mean wavelength for each region is shown in Table 1. Time lags were created by time-stepping the data from one to six days (Fig. 2). In other words data from each of the selected cells from one to six days prior were included in the final datasets by adding the corresponding information from the desired time step as additional explanatory variables.

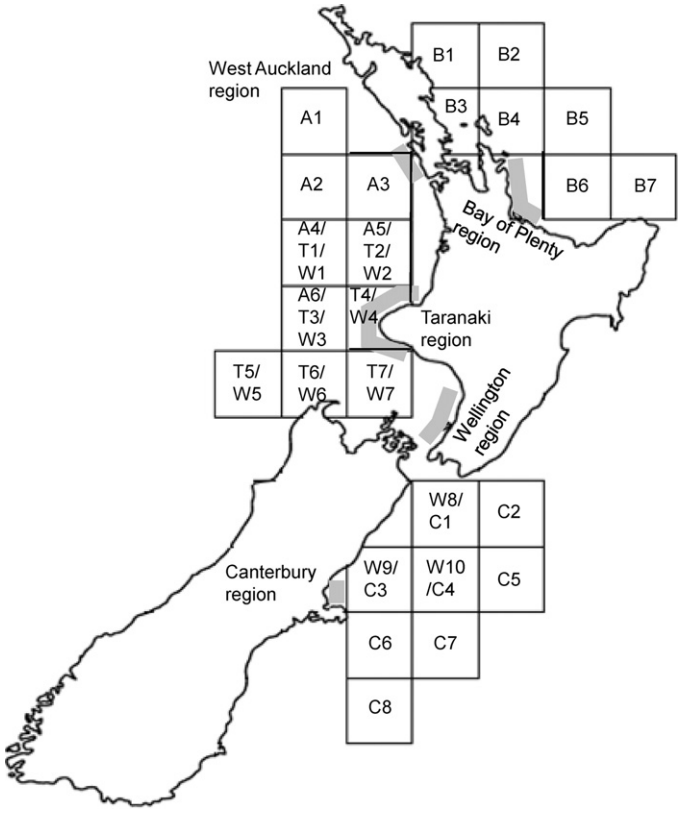

Fig. 1. Oceanic cells associated with each of the five regions examined. Cells that are associated with a particular region are shown by ID codes in which the letter indicates the associated region, except for the West Auckland region which is represented by an A, and the number identifies individual cells within a region.

2.3. Training and evaluation of MLP

Standard three neuron-layer MLP were used to model the data, and the learning algorithm used was an unmodified back-propagation algorithm with momentum (Fig. 3). The method of training and evaluating the MLP was similar to that suggested in Flexer (1996) and Prechelt (1996). To determine optimum parameters a series of runs were carried out over each region, where each run used a different combination of hidden neuron layer size, learning rate and momentum. Each run consisted of 100 trials. For each trial, the data were randomly divided into a training set, comprising two-thirds of the available data, and a test set comprising the remaining one-third. An MLP was then created with randomly initialised connection weights and trained over the training data set. Each network modelled a single region, that is, there was only one output neuron per network, where the output indicated the predicted presence or absence of *Physalia* in that region on that particular day.

Network accuracy was measured by assessing Cohen's Kappa statistic (Cohen, 1960). As the proportion of features to examples was high, especially with the larger time lags, to reduce

Table 1
Mean ocean wavelength for each region between 2000/2001 season to the 2004/2005 season in each of the modelled regions.

Region	Wavelength (m)	SEM
West Auckland	100.3	8.30
Bay of Plenty	168.0	1.34
Taranaki	146.1	10.85
Wellington	135.6	10.36
Canterbury	103.6	4.30

Date	Variable X
10	1.1125
11	1.8025
12	2.0475
13	1.63
14	1.41125
15	2.115
16	3.4075
17	3.22
18	2.715
19	1.8025

Date	Variable X	Variable X-1	Variable X-2
10	1.1125	1.8025	2.0475
11	1.8025	2.0475	1.63
12	2.0475	1.63	1.41125
13	1.63	1.41125	2.115
14	1.41125	2.115	3.4075
15	2.115	3.4075	3.22
16	3.4075	3.22	2.715
17	3.22	2.715	1.8025
18	2.715	1.8025	
19	1.8025		

Fig. 2. Representation of how time lags were created using time stepping. A dataset incorporating a 2 day time lag for a single variable is shown with the final data in grey. Time lags from 1 to 7 days were investigated.

the risk of overtraining, network training was carried out in two steps. The first step was to train the networks as above utilising all of the data and then using Olden scores (Olden and Jackson, 2002), to identify the percentage that each input feature contributed to the network. Olden scores allow the relative contribution of an explanatory variable to the network to be assessed against other variables (see Olden and Jackson, 2002). The second step was to reduce the number of features by selecting features in order of percent contribution until a total predetermined percent contribution was reached. Target total percent contribution was explored in steps of 10% with the network parameters being identified as above for each different total percent contribution. The accuracy of these networks was again assessed with Cohen's Kappa statistic to allow comparison between stages.

2.4. Versatile Quantum-inspired Evolutionary Algorithm

In this study we follow the wrapper approach for feature selection as introduced in (Kohavi and Sommerfield, 1995) and in detail discussed in (Kohavi and John, 1997). The wrapper methodology is a type of "black box" approach. In its core, it contains a general optimisation algorithm interacting with an induction or classification method. The optimisation task consists of a reliable identification of an optimal feature subset that maximises the classification accuracy determined by the inductor. Thus, the classification method provides a quality measure for a presented feature subset and hence, acts as the fitness function for a general evolutionary algorithm. Due to its low computational cost, we employed the Naïve Bayesian Classifier (NBC) in the wrapper context (Fig. 3). Although assumed to be less accurate than the Multilayer Perceptron (MLP) (Kotsiantis et al., 2006), NBC is often surprisingly competitive and can outperform some state-of-the-art algorithms on certain problems (Domingos and Pazzani, 1997; Kotsiantis, 2007).

We decided for the previously proposed Versatile Quantum-inspired Evolutionary Algorithm (vQEA) (Defoin-Platel et al., 2007) as the optimization algorithm in the wrapper due to its interesting properties in terms of solution quality and convergence speed. The method evolves in parallel a number of independent probability vectors, which interact at certain intervals with each other, forming

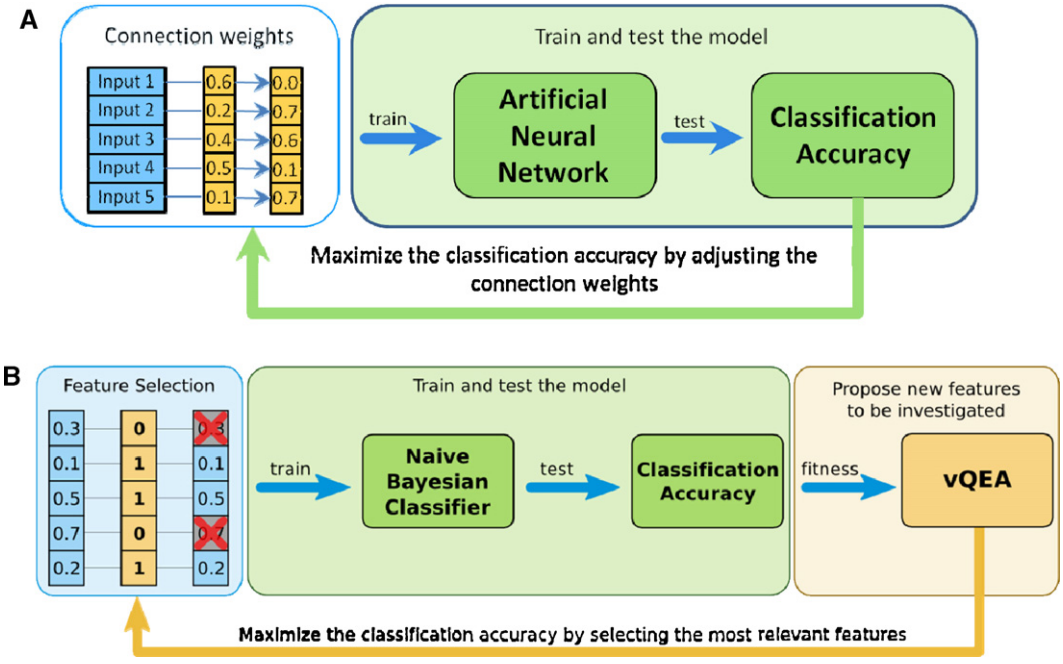

Fig. 3. Process overview of the model. (A) Artificial neural network; (B) Versatile Quantum-inspired Evolutionary Algorithm employing a Naïve Bayesian Classifier as a wrapper.

Table 2
Performance, parameters and number of features used to train Naïve Bayesian Classifier (NBC) associated with each region. * indicates significant increase ($p < 0.05$) compared with the best testing accuracy achieved by the MLP (Table 3), and ** a highly significant increase ($p < 0.001$) (T test).

Region	Lag	Generations	Parameter k	Final number of features selected	Overall Kappa	Test Kappa
West Auckland	4	3000	12	47	0.7912	0.6276**
Bay of Plenty	3	3000	10	42	0.8159	0.6347**
Taranaki	6	3000	7	94	0.9675	0.7034*
Wellington	6	1000	7	159	0.949	0.6961**
Canterbury	6	3000	9	87	0.8844	0.7082**

a multi-model Estimation of Distribution Algorithm (EDA) (Defoin-Platel et al., 2009). It has been shown that this approach performs well on epistatic problems, is very robust to noise, and needs only minimal fine-tuning of its parameters. In fact the standard setting for vQEA is suitable for a large range of different problem sizes and classes. Finally vQEA is a binary optimizer and fits well to the feature selection problem we want to apply it on.

The principle of the employed method is illustrated in Fig. 3. From each sample in the data set, features are selected using a binary mask. A "1"/"0" in this mask indicates selected/non-selected features of the data sample. Using k-fold cross-validation, the processed samples are then split into training and testing sets and passed to the classification method, i.e., the NBC. The learning process includes the presentation of all training samples. After the learning, the classification accuracy is determined on the set of test samples and results are averaged over all k folds. This accuracy provides a quality measure of the feature subset which in turn is passed to the optimisation algorithm, i.e., vQEA. Based on the quality, vQEA adapts the search strategy and generates new feature subsets to the NBC for evaluation. The whole process iterates until a termination criterion is met, i.e., a predefined classification accuracy is reached or the maximum number of iterations is exhausted.

2.5. Experimental setup

For vQEA, we chose a population structure of ten individuals organized in a single group which is globally synchronized every generation. This setting was reported to work well for a number of different binary optimization benchmarks (Defoin-Platel et al., 2009). The learning rate $\Delta\theta$ determines the convergence speed of vQEA and its setting has to be chosen in compliance with the problem at hand. Too large learning rates may result in premature convergence of the algorithm to non-optimal solutions, while too small learning rates will require the computation of potentially too many iterations. After some initial experiments, we chose $\Delta\theta = \pi/100$. The algorithm was allowed to evolve over a total of 3000 generations except for the Wellington region which evolved for 1000 generations. The reduced number of generations was a result of the Wellington region having a much higher number of features because the most oceanographic cells associated with it (Fig. 1). This significantly increased the time to evolve a generation and as the availability of the hardware needed to carry out the computations was limited, the number of generations was reduced. To guarantee statistical relevance, 30 independent runs were performed, using a different random number seed for each of them.

As mentioned above, the NBC was trained and tested using a k-fold cross-validation procedure. Parameter k was set for each dataset individually and are summarised in Table 2. The classification error was assessed by Cohen's Kappa statistic (Cohen, 1960) across both the entire dataset and the test dataset only. Features that were selected by the NBC in 90% of the runs were compared to the eight features that had the greatest contribution in the finalised MLP. As there was a high degree of correlation between some variables, if a model selected a feature and the other model selected another feature that was highly correlated with the initial feature, then that was considered a comparable selection between model types. Highly correlated features were not removed from the dataset because even though a feature is correlated, it still may provide significant performance improvement when analysed in conjunction with other features (Maier and Dandy, 2000; Guyon and Elisseeff, 2003).

3. Results

3.1. Feature selection

Features that had a large contribution to the MLP networks for each region are shown in Table 3. In general, the features that had a large contribution to all regions except Taranaki were wind and wave directions at varying time lags. Taranaki however, was most influenced by wind speed, again at varying time lags. The mean number of features identified by NBC across the 30 runs (Table 2) was higher in all regions than the corresponding features selected through percent contribution in the MLP (Table 4). When the top eight contributing features to the MLP are compared to features that the NBC selected in 90% of the runs it is clear that the models identified the same underlying pattern with an average of 4.8 (SEM 0.86) features the same, or highly correlated (great than 0.70% correlation) (Table 3). The features identified by both the MLP networks and NBC suggests that there are two separate oceanographic systems occurring around New Zealand that may influence *Physalia* presence. One system occurs in the Bay of Plenty region and a more complex system incorporates the West Auckland and Taranaki, regions (Fig. 4).

3.2. MLP accuracies

The best performing network models based on a combination of time lag and network parameters for each region are shown in Table 4. The optimum time lag over the regions ranged between 3 and 6 days. The large difference between the training and test results where the training kappa were between 0.97 and 0.99 compared with the test kappa of between 0.26 and 0.40 was an indication that overtraining had occurred.

The percentage contribution that determined the number of features selected was 50% in all regions except for Canterbury which was 40% (Table 4). In other words the network performed best when features selected from the largest contributing features represented 40–50% of the total contribution based on Olden scores (Olden and Jackson, 2002). Networks that were trained with a reduced number of features based on their contribution outperformed the corresponding full dataset. In all regions the training kappa decreased but there was a subsequent increase in test kappa ranging from 0.11 to 0.27 which was an indication that the reduction in the number of features had reduced overtraining in the networks (Table 4).

3.3. Naïve Bayesian Classifier accuracies

Because of the strong imbalance of the presence/absence data, the percent of correctly classified samples started at a high level >80%, for all regions, at the beginning of the evolutionary run.

Table 3
Features that made the greatest contribution, either positive or negative, to the MLP networks for each of the five regions. The letter and number after each feature indicate the oceanographic cell (Fig. 1) in which the feature was measured and the associated time lag. * indicates that the feature, or another highly correlated feature, was selected by the NBC.

Region	Positive features				Negative features			
	Rank	Cell-lag	Feature	Contribution	Rank	Cell-lag	Feature	Contribution
West Auckland	3	Cell 6-0	Wave direction*	58.31	1	Cell 3-1	Wave period*	−84.85
	8	Cell 5-0	Wave direction	43.67	2	Cell 5-0	Wind speed	−84.76
	9	Cell 1-3	Wave period	43.44	4	Cell 1-4	Wind speed*	−57.53
	10	Cell 4-2	Wind direction	41.84	5	Cell 5-1	Wind speed*	−55.91
Bay of Plenty	2	Cell 3-0	Wave direction*	90.32	1	Cell 6-0	Wind speed*	−100.29
	4	Cell 6-1	Wind direction*	78.55	3	Cell 2-1	Wave direction*	−79.57
	7	Cell 7-2	Wind speed*	73.35	5	Cell 4-1	Wave direction*	−77.47
	10	Cell 4-0	Wind direction*	58.33	6	Cell 6-3	Wind direction*	−75.38
Taranaki	5	Cell 6-3	Wind speed	30.28	1	Cell 6-6	Wind speed	−43.72
	6	Cell 2-1	Wave period	30.17	2	Cell 6-6	Wave period	−37.62
	7	Cell 7-2	Wind speed	28.53	3	Cell 2-1	Wind speed*	−34.69
	9	Cell 7-1	Wind speed*	26.37	4	Cell 5-2	Wave period*	−32.58
Wellington	1	Cell 10-1	Wind direction*	35.14	2	Cell 3-5	Wave direction*	−34.90
	5	Cell 10-5	Wind direction*	31.28	3	Cell 10-4	Wave period	−34.48
	6	Cell 7-1	Wave direction	31.14	4	Cell 8-5	Wave direction*	−32.36
	7	Cell 4-5	Wind speed	30.99	8	Cell 3-1	Wave period*	−30.76
Canterbury	2	Cell 1-4	Wave direction	55.83	1	Cell 3-6	Wave direction	−60.39
	3	Cell 2-3	Wind direction	52.03	4	Cell 3-6	Wind direction*	−51.77
	5	Cell 7-1	Wind direction*	50.93	6	Cell 8-2	Wind speed	−42.22
	8	Cell 7-6	Wind direction*	40.93	7	Cell 8-4	Wave direction*	−41.85

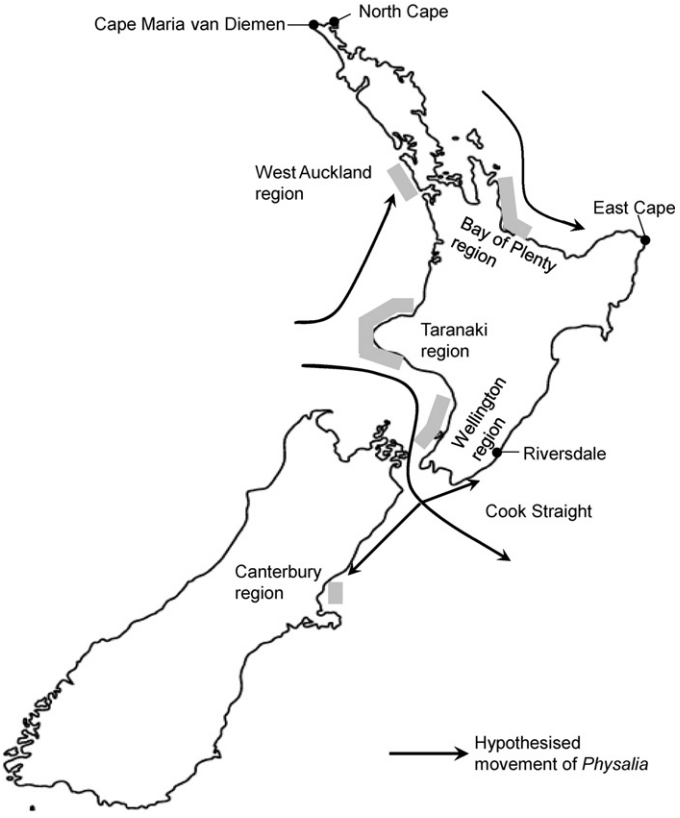

Fig. 4. Hypothesised representation of *Physalia* movement around New Zealand as indicated from both the ANN and NBC models.

Despite the high initial accuracy, improvements to classification accuracy were still possible in later generations (Fig. 5). Classification accuracies achieved on the test set were significantly higher than the MLP testing accuracies for all regions (Table 4). Despite the reduced number of generations over which the Wellington region was evolved over, this region achieved comparable classification accuracy with other regions, however, increased generations may have improved and clarified results further. As the vQEA was evolved, the accuracy for all regions steadily increased with a corresponding decrease in the number of features (Fig. 5).

4. Discussion

Network accuracy improved as a result of (1) reducing noise in the input data by excluding irrelevant information and, (2) improving the ratio of features to examples in the data set (Nath et al., 1997; Maier and Dandy, 2000). For example, test set predictions for the West Auckland region improved by 11% to a Kappa of 0.5191 compared with results obtained by (Pontin et al., 2009) in a study of *Physalia* occurrence in the West Auckland region where a kappa of 0.40 was obtained. Further to that study, more regions were included allowing generalisation over a wider area than achieved previously. As irrelevant features were eliminated, the ratio of features to examples decreased subsequently reducing the noise and increasing performance. A similar result was achieved with the NBC where performance increased as measured by the percentage of correctly classified events as features were discarded. The NBC outperformed the MLP classifying *Physalia* presence. Both models identified similar features as important in the marine system. Because similar features were identified independently by the different approaches they could be considered to be highly relevant

Table 4
Optimised training parameters used to train MLP networks and mean Cohen's Kappa statistic for the training, test and validation datasets associated with each region. "Neurons" indicates the number of hidden layer neurons. Numbers in brackets in the contribution column indicate total number of features included in the networks.

Region	Lag	Contribution	Neurons	Learning	Momentum	Epochs	Training	Test
West Auckland	4	50% (33)	8	0.2	0.1	900	0.9004	0.5191
Bay of Plenty	3	50% (31)	9	0.2	0.1	800	0.8824	0.4023
Taranaki	6	50% (54)	8	0.2	0.3	950	0.9691	0.6479
Wellington	6	50% (83)	4	0.6	0.6	400	0.9947	0.5807
Canterbury	6	40% (45)	9	0.6	0.3	800	0.9656	0.6279

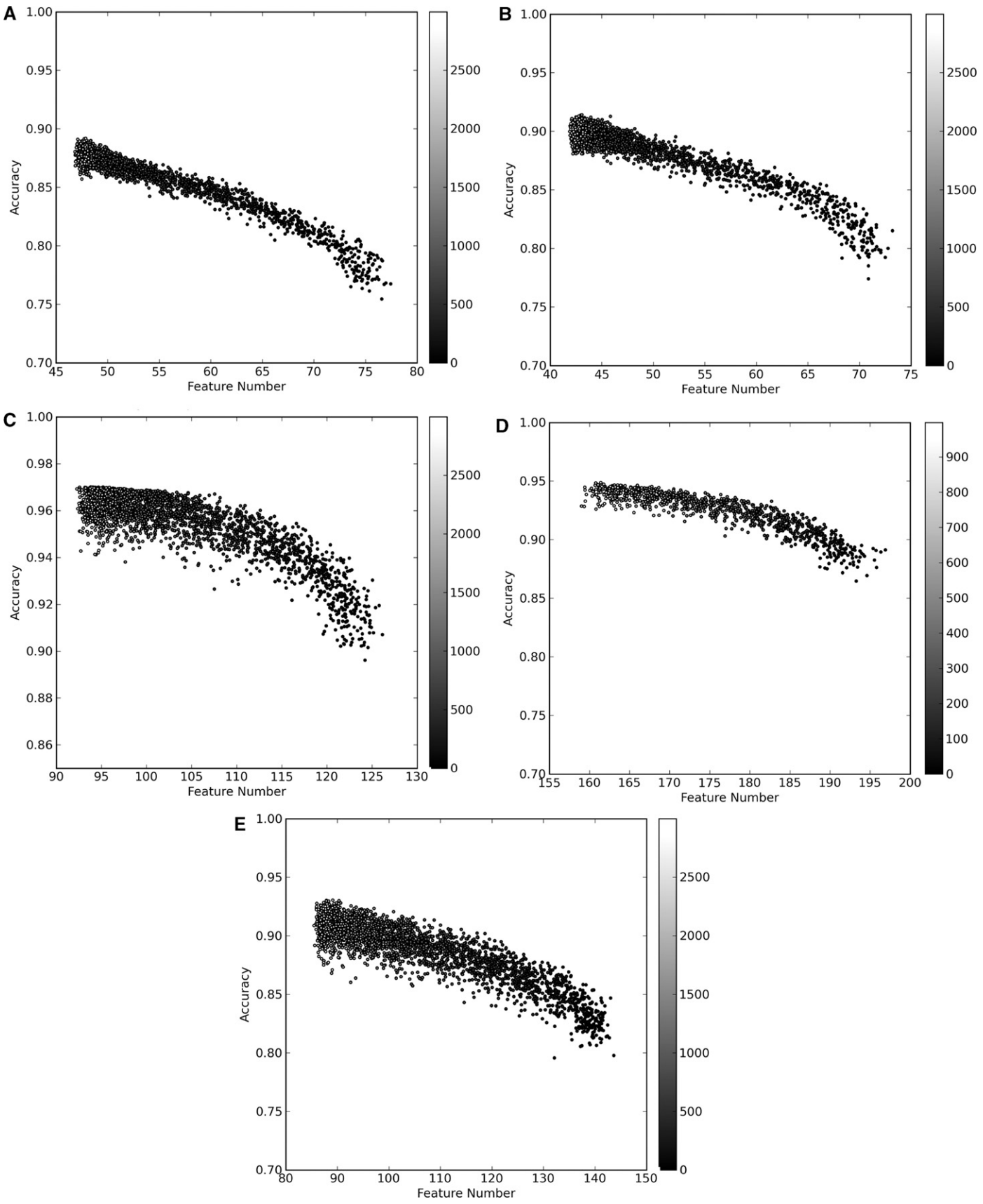

Fig. 5. Evolution of NBC for classifying *Physalia* presence in five New Zealand regions (A: West Auckland, B: Bay of Plenty, C: Taranaki, D: Wellington and E: Canterbury) in relation to the number of features incorporated in the model and classification accuracy (percentage correctly classified). The different grey levels correspond to the generation in which a given data point was obtained, the lighter the colour the later the generation. Note the Wellington region was only evolved over 1000 generations compared to 3000 generations for the other regions.

(Bowden et al., 2005; Muttil and Chau, 2007). Furthermore, by using the variables identified by both ANN and NBC as an ensemble it may be possible to produce a more accurate model. Several studies such as those by Araujo and New (2007) and Lankin-Vega et al. (2008) have shown that greater precision is often gained with ensembles of ecological models. When assessing the ecological role that a feature has within a given system, the MLP networks provide additional knowledge to that of an NBC as it is possible to determine whether the network responds either positively or negatively to a feature. That is not possible with a NBC.

The features identified by both the MLP networks and NBC suggests that there are two separate oceanographic systems occurring around New Zealand that may influence *Physalia* presence. One system occurs in the Bay of Plenty region and a more complex system incorporates the West Auckland and Taranaki, regions (Fig. 4). In the Bay of Plenty it appears that oceanographic conditions to the north of the region play an important role determining *Physalia* occurrence as indicated by the MLP networks. Whereas, for the other regions wind and wave directions that are from the west to northwest promote the presence of *Physalia*.

Brodie (1960) reported the release of over 10,000 float cards to assess surface ocean currents around New Zealand. It is reasonable to assume that a passive disperser such as *Physalia* would display similar movement patterns to the cards. Cards released from the North Cape drifted down the east coast of the North Island through the Bay of Plenty to East Cape. Cards released from East Cape were not recovered but the East Cape current moves south until it meets the Canterbury current around the Cook Straight area (Gardner, 1961). The pattern recorded with float cards corresponds with the hypothesis that the general direction of *Physalia* movement in the Bay of Plenty is from the north. Cards released west of Cook Straight between the top of the South Island and Taranaki indicated that general movement was towards and through Cook Strait. Although Dell (1952) noted that it was possible for drift bottles released in the Southern Ocean to be stranded between Taranaki and Cape Maria van Diemen at the top of the North Island. These drift patterns support the suggested findings in this study that the West Auckland, Taranaki and Wellington regions are linked. The models suggested that the Canterbury region is also linked with this system with northwest winds moving individuals through Cook Strait and down the east coast of the South Island. For the Canterbury region, this observation contradicts the drift data for the region that indicates a steady current from the bottom of the South Island up the east coast of the island to the Southern convergence zone (Brodie, 1960). As there are only 5 records of people being stung south of Canterbury over the study period it is unlikely that individuals are being transported from the south as similar or increased incidence rate would be expected compared with Canterbury.

More interesting is that molecular data taken from *Physalia* collected around New Zealand and Australia also supports the existence of two circulatory systems indicated by the models (Pontin, 2010). In a study of New Zealand *Physalia* phylogenetics Pontin (2010) identified that a single clade exists in the Bay of Plenty extending down most of the east coast of the North Island to Riversdale. In all the other modelled regions excluding the Bay of Plenty, the molecular results were not quite as clear because a complex, of potentially two separate clades were detected. However, it was clear from this analysis that genetic information is being shared between all regions except for the Bay of Plenty indicating that either individuals or gametes are capable of moving between those regions.

The lack of available surface current information for inclusion in the models has limited the scope to generalise the importance of the selected features. However, because wind and wave conditions either enhance or suppress current strength (Stanton et al., 1997), the lack of current data should not influence model ability to detect the two circulatory systems. The models account for the influence of the currents indirectly through detecting oceanographic conditions that can enhance or suppress the occurrence of *Physalia* in the modelled region. Also limiting the model was the daily observation data as this necessitated the averaging of the explanatory data preventing the exploration of diurnal pattern of wind forcing and other short duration phenomena which may provide additional information to aid in the prediction of *Physalia* occurrence and identification of key ecological drivers.

5. Conclusion

This study has provided a foundation to increased understanding of how a passive disperser, represented by *Physalia* is circulated around New Zealand on which further model development can be based. The identification of two independent circulatory systems by the models will require further study to confirm that similar patterns occur in other New Zealand passive drifting species. More generally, this study demonstrates the ability of machine learning and data mining techniques to generate interesting hypotheses from noisy and complex data, typical of ecological systems about which there is little knowledge.

References

Araujo, M.B., New, M., 2007. Ensemble forecasting of species distributions. Trends in Ecology & Evolution 22, 42–47.

Baluja, S., 1994. Population-based incremental learning. In: A Method for Integrating Genetic Search Based Function Optimization and Competitive Learning. Carnegie Mellon University, Pittsburgh, PA.

Bolker, B.M., 2008. Ecological Models and Data in R. Princeton University Press, Princeton, NJ.

Bowden, G.J., Dandy, G.C., Maier, H.R., 2005. Input determination for neural network models in water resources applications. Part 1. Background and methodology. Journal of Hydrology 301, 75–92.

Bradbury, R.I., Laurel, B., Snelgrove, P.V.R., Bentzen, P., Campana, S.E., 2008. Global patterns in marine dispersal estimates, the influence of geography, taxonomic category and life history. Proceedings of the Royal Society B – Biological Sciences 275, 1803–1809.

Brodie, J.W., 1960. Coastal surface currents around New Zealand. New Zealand Journal of Geology and Geophysics 3, 235–252.

Brosse, S., Guegan, J.-F., Tourenq, J.-N., Lek, S., 1999. The use of artificial neural network to assess fish abundance and spatial occupancy in the littoral zone of a mesotrophic lake. Ecological Modelling 120, 299–311.

Burnham, K.P., Anderson, D.R., 2002. Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach. Springer, New York.

Chen, M.C., Chen, L.S., Hsu, C.C., Zeng, W.R., 2008. Information granulation based data mining approach for classifying imbalanced data. Information Science 178, 3214–3227.

Cohen, J., 1960. A coefficient of agreement for nominal scales. Educational and Psychological Measurement 20, 37–46.

Collins, A.G., 2002. Phylogeny of Medusozoa and the evolution of cnidarian life cycles. Journal of Evolutionary Biology 15, 418–432.

Cortes, C., Vapnik, V., 1995. Support-vector networks. Machine Learning 20, 273–297.

Cowen, R.K., Lwiza, K.M.M., Sponaugle, S., Paris, C.B., Olson, D.B., 2000. Connectivity of marine populations: open or closed? Science 287, 857–859.

Cowen, R.K., Paris, C.B., Srinivasan, A., 2006. Scaling of connectivity in marine populations. Science 311, 522–527.

Dawson, M.N., Jacobs, D.K., 2001. Molecular evidence for cryptic species of *Aurelia aurita* (Cnidaria, Scyphozoa). Biological Bulletin 200, 92–96.

Defoin-Platel, M.D., Schliebs, S., Kasabov, N., 2009. Quantum-inspired evolutionary algorithm: a multimodel EDA. Transactions on Evolutionary Computation 13, 1218–1232.

Defoin-Platel, M.D., Schliebs, S., Kasabov, N., 2007. A versatile quantum-inspired evolutionary algorithm. In: IEEE Congress on Evolutionary Computation , Singapore, pp. 423–430.

Dell, R.K., 1952. Ocean current affecting New Zealand. New Zealand Journal of Science and Technology B 34, 86–91.

Domingos, P., Pazzani, M., 1997. On the optimality of the simple Bayesian classifier under zero-one loss. Machine Learning 29, 103–130.

Elith, J., Graham, C.H., Anderson, R.P., Dudik, M., Ferrier, S., Guisan, A., Hijmans, R.J., Huettmann, F., Leathwick, J.R., Lehmann, A., Li, J., Lohmann, L.G., Loiselle, B.A., Manion, G., Moritz, C., Nakamura, M., Nakazawa, Y., Overton, J.M., Peterson, A.T., Phillips, S.J., Richardson, K., Scachetti-Pereira, R., Schapire, R.E., Soberon, J., Williams, S., Wisz, M.S., Zimmermann, N.E., 2006. Novel methods improve prediction of species' distributions from occurrence data. Ecography 29, 129–151.

Fisher, N.I., 1995. Statistical analysis of circular data. Cambridge University Press, Cambridge.
Flexer, A., 1996. Statistical Evaluation of Neural Network Experiments, Minimum Requirements and Current Practice. In: Trappl, R. (Ed.), Cybernetics and Systems '96, Proceedings of the 13th European Meeting on Cybernetics and Systems Research. Austrian Society for Cybernetic Studies, pp. 1005–1008.
Fogel, G.B., 2008. Computational intelligence approaches for pattern discovery in biological systems. Briefings in Bioinformatics 9, 307–316.
Friedman, N., Geiger, D., Goldszmidt, M., 1997. Bayesian network classifiers. Machine Learning 29, 131–163.
Gardner, D.M., 1961. Hydrology of New Zealand coastal waters. Bulletin of the New Zealand Department of Scientific and Industrial Research 138, 1–84.
Guyon, I., Elisseeff, A., 2003. An introduction to variable and feature selection. Journal of Machine Learning Research 3, 1157–1182.
Han, K.H., Kim, J.H., 2004. Quantum-inspired evolutionary algorithms with a new termination criterion. Hε gate, and two phase scheme. IEEE Transactions on Evolutionary Computation 8, 156–169.
Hochachka, W.M., Caruana, R., Fink, D., Munson, A., Riedewald, M., Sorokina, D., Kelling, S., 2007. Data-mining discovery of pattern and process in ecological systems. Journal of Wildlife Management 71, 2427–2437.
Joy, M.K., Death, R.G., 2004. Predictive modelling and spatial mapping of freshwater fish and decapod assemblages using GIS and neural networks. Freshwater Biology 49, 1036–1052.
Kinlan, B.P., Gaines, S.D., 2003. Propagule dispersal in marine and terrestrial environments: a community perspective. Ecology 84, 2007–2020.
Knowlton, N., 2000. Molecular genetic analyses of species boundaries in the sea. Hydrobiologia 420, 73–90.
Kohavi, R., Sommerfield, D., 1995. Feature subset selection using the wrapper method: overfitting and dynamic search space topology. In: The First International Conference on Knowledge Discovery and Data Mining , pp. 192–197.
Kohavi, R., John, G.H., 1997. Wrappers for feature subset selection. Artificial Intelligence 97, 273–324.
Kotsiantis, S.B., 2007. Supervised machine learning. A review of classification techniques. Informatica 31, 249–268.
Kotsiantis, S.B., Zaharakis, I.D., Pintelas, P.E., 2006. Machine learning: a review of classification and combining techniques. Artificial Intelligence Review 26, 159–190.
Lane, C.E., 1960. The Portuguese Man-of-War. Scientific America 2002, 158–168.
Lankin-Vega, G., Worner, S.P., Teulon, D.A.J., 2008. An ensemble model for predicting *Rhopalosiphum padi* abundance. Entomologia Experimentalis et Applicata 129, 308–315.
Lek, S., Delacoste, M., Baran, P., Dimopoulos, I., Lauga, J., Aulagnier, S., 1996. Application of neural networks to modelling nonlinear relationships in ecology. Ecological Modelling 90, 39–52.
Levin, L.A., 2006. Recent progress in understanding larval dispersal, new directions and digressions. Integrative and Comparative Biology 46, 282–297.
Liu, J., Hu, Q., Yu, D., 2008. A weighted rough set based method developed for class imbalance learning. Information Sciences 178, 1235–1256.
Maier, H.R., Dandy, G.C., 2000. Neural networks for the prediction and forecasting of water resources variables: a review of modelling issues and applications. Environmental Modelling and Software 15, 101–124.
Moon, J.H., Pang, I.C., Yang, J.Y., Yoon, W.D., 2010. Behavior of the giant jellyfish *Nemopilema nomurai* in the East China Sea and East/Japan Sea during the summer of 2005. A numerical model approach using a particle-tracking experiment. Journal of Marine Systems 80, 101–114.
Murthy, S.K., 1998. Automatic construction of decision trees from data. A multidisciplinary survey. Data Mining and Knowledge Discovery 2, 345–389.
Mutanga, O., Skidmore, A.K., 2004. Integrating imaging spectroscopy and neural networks to map grass quality in the Kruger National Park, South Africa. Remote Sensing of Environment 90, 104–115.
Muttil, N., Lee, J.H.W., 2005. Genetic programming for analysis and real-time prediction of coastal algal blooms. Ecological Modelling 189, 363–376.
Muttil, N., Chau, K.W., 2007. Machine-learning paradigms for selecting ecologically significant input variables. Engineering Applications of Artificial Intelligence 20, 735–744.
Nath, R., Rajagopalan, B., Ryker, R., 1997. Determining the saliency of input variables in neural network classifiers. Computers & Operations Research 24, 767–773.
Olden, J.D., Jackson, D.A., 2002. Illuminating the "black box": a randomization approach for understanding variable contributions in artificial neural networks. Ecological Modelling 154, 135–150.
Olden, J.D., Joy, M.K., Death, R.G., 2004. An accurate comparison of methods for quantifying variable importance in artificial neural networks using simulated data. Ecological Modelling 178, 389–397.
Palumbi, S.R., 1992. Marine speciation on a small planet. Trends in Ecology & Evolution 7, 114–118.
Pontin, D.R., 2010. Factors Influencing the Occurrence and Abundance of Stinging Jellyfish (*Physalia* spp.) at New Zealand Beaches. Lincoln University, Lincoln.
Pontin, D.R., Watts, M.J., Worner, S.P., 2008. Using multi-layer perceptrons to predict the presence of jellyfish of the genus *Physalia* at New Zealand beaches. In: International Joint Conference on Neural Networks , Hong Kong, pp. 1171–1176.
Pontin, D.R., Worner, S.P., Watts, M.J., 2009. Using time lagged input data to improve prediction of stinging jellyfish occurrence at New Zealand beaches by multi-layer perceptrons. Lecture Notes in Computer Science 5506, 907–914.
Prechelt, L., 1996. A quantitative study of experimental evaluations of neural network learning algorithms: current research practice. Neural Networks 9, 457–462.
Ready, J., Kaschner, K., South, A.B., Eastwood, P.D., Rees, T., Rius, J., Agbayani, E., Kullander, S., Froese, R., 2010. Predicting the distributions of marine organisms at the global scale. Ecological Modelling 221, 467–478.
Rumelhart, D.E., Hinton, G.E., Williams, R.J., 1986. Learning representations by back-propagation errors. Nature 323, 533–536.
Segurado, P., Araujo, M.B., 2004. An evaluation of methods for modelling species distributions. Journal of Biogeography 31, 1555–1568.
Schliebs, S., Defoin-Platel, M.D., Worner, S.P., Kasabov, N., 2009. Integrated feature and parameter optimization for an evolving spiking neural network: exploring heterogeneous probabilistic models. Neural Networks 22, 623–632.
Slaughter, R.J., Beasley, M.G.D., Lambie, B.S., Schep, L.J., 2009. New Zealand's venomous creatures. The New Zealand Medical Journal 122, 83–97.
Stanton, B.R., Sutton, P.J.H., Chiswell, S.M., 1997. The East Auckland Current, 1994–95. New Zealand Journal of Marine and Freshwater Research 31, 537–549.
Stockwell, D.R.P., Peters, D.P., 1999. The GARP modelling system: problems and solutions to automated spatial prediction. International Journal of Geographic Information Systems 13, 143–158.
Tolman, H.L., 1998. Validation of a new global wave forecast system at NCEP. In: Edge, B.L., Helmsley, J.M. (Eds.), Ocean Wave Measurements and Analysis. ASCE, pp. 777–786.
Virkkala, R., Marmion, M., Heikkinen, R.K., Thuiller, W., Luoto, M., 2010. Predicting range shifts of northern bird species: influence of modelling technique and topography. Acta Oecologica 36, 269–281.
White, C., Selkoe, K.A., Watson, J., Siegel, D.A., Zacherl, D.C., Toonen, R.J., 2010. Ocean currents help explain population genetic structure. Proceedings of the Royal Society B – Biological Sciences 277, 1685–1694.
Xu, L., Chow, M.-Y., 2006. A classification approach for power distribution systems fault cause identification. IEEE Transactions on Power Systems 21, 53–60.
Zhang, S., Zhang, C., Yang, Q., 2003. Data preparation for data mining. Applied Artificial Intelligence 17, 375–381.