Factors influencing the occurrence of stinging jellyfish (*Physalia* spp.) at New Zealand beaches

---

A thesis
submitted in partial fulfilment
of the requirements for the Degree of
Doctor of Philosophy

at

Lincoln University

by

David R. Pontin

Lincoln University

2009

Abstract

Abstract

Individuals of the cnidarian genus *Physalia* are a common sight at New Zealand beaches and are the primary cause of jellyfish stings to beachgoers each year. The identity of the species and the environmental factors that determine its presence are unknown. Lack of knowledge of many marine species is not unusual, as pelagic invertebrates often lack detailed taxonomic descriptions as well as information about their dispersal mechanisms such that meaningful patterns of distribution and dispersal are almost impossible to determine. Molecular systematics has proven to be a powerful tool for species identification and for determining geographical distributions. However, other techniques are needed to indicate the causal mechanisms that may result in a particular species distribution. The aim of this study was to apply molecular techniques to the cnidarian genus *Physalia* to establish which species occur in coastal New Zealand, and to apply models to attempt to forecast its occurrence and infer some mechanisms of dispersal.

Physalia specimens were collected from New Zealand, Australia and Hawaii and sequenced for Cytochrome c oxidase I (COI) and the Internal transcribed spacer 1 (ITS1). Three clans were found: a Pacific-wide clan, an Australasian clan and New Zealand endemic clan with a distribution confined to the Bay of Plenty and the East Coast of the North Island. Forecasting *Physalia* occurrence directly from presence data using artificial neural networks (ANN) proved unsuccessful and it was necessary to pre-process the presence data using a variable sliding window to reduce noise and improve accuracy. This modelling approach outperformed the time lagged based networks giving improved forecasts in both regions that were assessed. The ANN models were able to indicated significant trends in the data but would require more

data at higher resolution to give more accurate forecasts of *Physalia* occurrence suitable for decision making on New Zealand beaches.

To determine possible causal mechanisms of recorded occurrences and to identify possible origins of *Physalia* the presence and absence of *Physalia* on swimming beaches throughout the summer season was modelled using ANN and Naïve Bayesian Classifier (NBC). Both models were trained on the same data consisting of oceanographic variables. The modelling carried out in this study detected two dynamic systems, which matched the distribution of the molecular clans. One system was centralised in the Bay of Plenty matching the New Zealand endemic clan. The other involved a dynamic system that encompassed four other regions on both coasts of the country that matched the distribution of the other clans. By combining the results it was possible to propose a framework for *Physalia* distribution including a mechanism that has driven clan divergence. Moreover, potential blooming areas that are notoriously hard to establish for jellyfish were hypothesised for further study and/or validation.

Contents

List of Tables

List of Figures

Chapter 1: Introduction

1.1 Introduction

Modelling of marine populations has primarily focused on commercial fish populations with many models and model types developed to assess population levels and predict future yields (McAllister & Kirkwood 1998; Cotter *et al.* 2004; Pelletier & Mahevas 2005). Although in recent years with the advent of genetic studies depicting how populations are linked at a genetic level there is a growing interest in how marine larvae disperse to determine and understand theses linkages (Cowen *et al.* 2000). Compared to other marine taxa, jellyfish have been neglected in marine research for the past 100 years (Haddock 2004). More recently, there has been an increasing awareness of the role jellyfish have within the marine ecosystem, particularly with respect to the threat of climate change and higher rates of eutrophication within ocean systems. Both have been linked with increases in jellyfish populations (Mills, 2001; Purcell and Arai, 2001; Parsons and Lalli, 2002) with consequent environmental problems such as blocked nets in aquacultural enterprises, and high populations that can affect fish stocks by severely depleting eggs, small larvae and plankton (Purcell *et al.* 2007). As a result, there is renewed interest in identifying environmental factors that influence jellyfish abundance and dispersal to assess any potential impacts. However, there is considerable difficulty defining a jellyfish populations as there are potential issues with species taxonomy and few datasets of appropriate population data exist (Purcell 2005), therefore the ability to establish base population levels for species in specific geographic regions is restricted. Despite this, there is considerable opportunity to utilise molecular techniques to define populations so that were possible datasets can be related to the populations, providing that it is possible to identify shortcomings in the data and it is treated appropriately (Elith *et al.* 2006).

1.2 Cnidaria in New Zealand

The phylum Cnidaria encompasses all true corals and their relatives (eg hydroids, sea anemones and sea fans). The phylum is highly diverse with many species displaying brilliant colouring which when combined with the radial symmetry that is characteristic of cnidarians (Barnes 1980) creates a beauty that is surpassed by few other animals. Eight hundred and ninety five species of Cnidaria have been recorded in New Zealand waters, representing nearly 10% of the world's diversity (Smith & Gordon 2003) of these 210 are unnamed. Species range from the endemic tree-like black coral (*Antipathes fiordensis*) through to the brilliantly coloured sea anemone, *Corynactis australis* and include 26 species of jellyfish. The phylum Cnidaria has traditionally been split into three classes (Anthozoa, Hydrozoa and Scyphozoa) but because of recent interest in genetic analysis of the phylum this is subject to debate. The class Cubozoa was added to the phylum in 1975 and consists of species that have a free swimming planula larva that settles and develops into a sessile polyp (Collins 2002). Because of further genetic information Marques & Collins (2004) proposed that a new class, Staurozoa, could be formed from the Scyphozoans giving a possible five classes in the phylum (Anthozoa, Hydrozoa, Scyphozoa, Staurozoa and Cubozoa, Figure 1.1). In spite of the rapid progress made in the higher-level classification of Cnidaria there has been little change within the class Hydrozoa to which the genus *Physalia* belongs, as this appears to be a clade well supported by genetic analyses (Collins 2002).

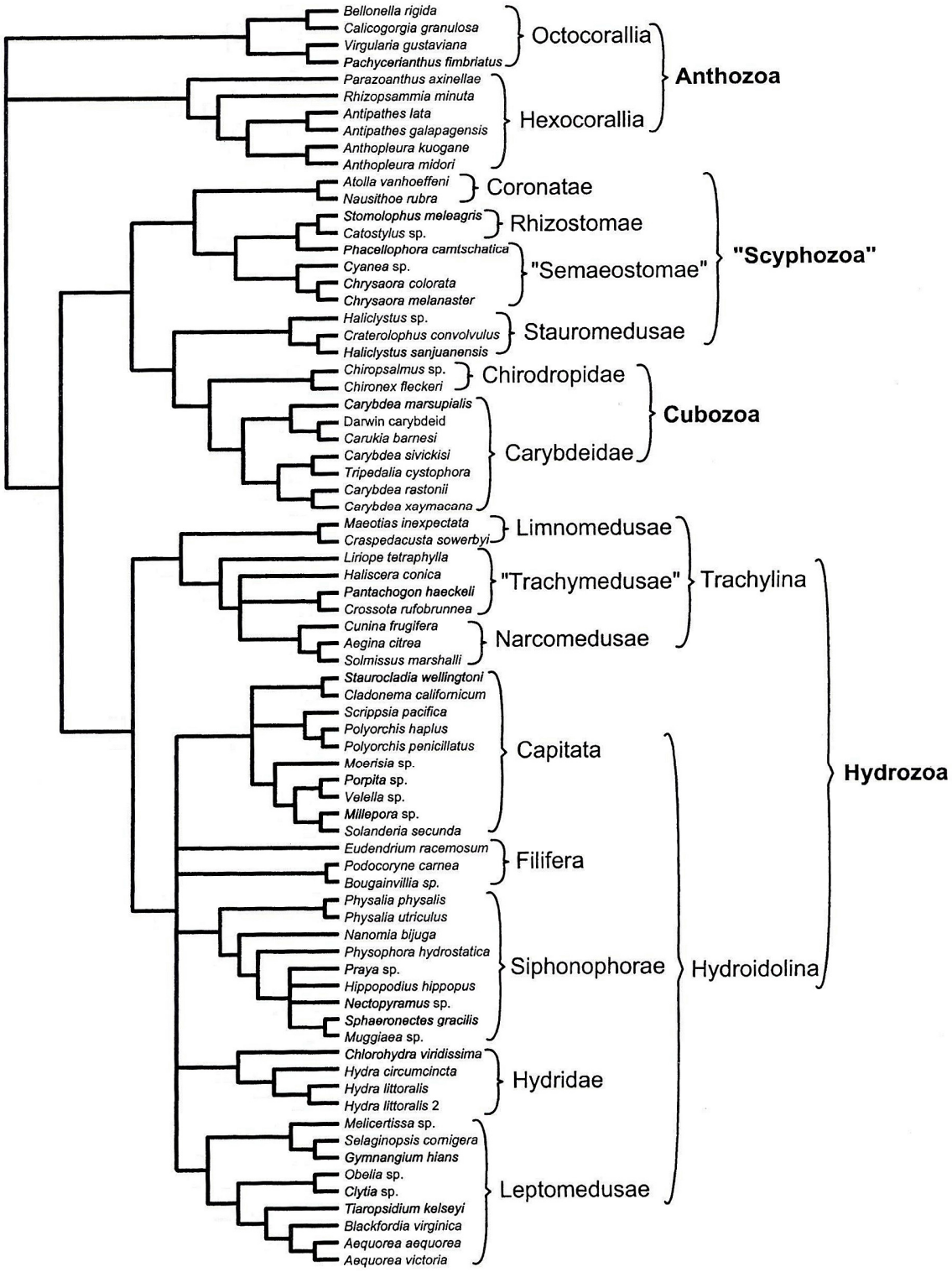


Figure 1.1: Taxonomic tree of the Hydrozoa as proposed by Collins (2002).

The class Hydrozoa contains 3702 nominal species (Bouillon *et al.* 2006) of which 134 are considered to be present in New Zealand (Bouillon & Barnett 1999). Hydrozoans are often not considered to be true jellyfish as the medusa is not the dominant form for many species (Barnes 1980). The order Siphonophora, to which the genus *Physalia* belongs, is considered a

highly specialised Hydrozoan order that has evolved from an essentially sessile benthic existence to become pelagic. Members exist as large pelagic colonies composed of modified polypoid and medusoid individuals. Siphonophorans are classified into three groups based on the presence or absence of two features that are associated with pelagic life. Species that possess swimming bells are classified as physonect species These species are considered to have evolved from the cystonects, which possess a float but in turn evolved into the calycophores, which are characterised by having both features (Collins 2002). The presence of a float made it possible for the early siphonophores to evolve from a benthic to pelagic existence as the float enables the individual to move around the ocean through the passive use of winds and currents. Some species such as the pacific *Nanomia bijuga* have been recorded at depths between 10m and 800m (Robinson *et al.* 1998) as they are capable of regulating the gas content of the float. The development of the swimming bell, as in *Lensia conoidea,* is an adaptation for self-locomotion within the pelagic realm rather than relying on external factors.

The bluebottle is a common jellyfish in New Zealand that is readily identified by beachgoers, predominantly because the species is responsible for most of the stings inflicted on these people each year. There is considerable taxonomic confusion about which species are found in coastal New Zealand. The literature suggests the species most likely to be present are the *Physalia physalis* (Portuguese man-of-war) and *Physalia utriculus*, although neither has been categorically confirmed as present in New Zealand (Bouillon & Barnett 1999; Collins 2002; Bouillon *et al.* 2006). Furthermore, the existence of *P. utriculus* as a species is highly debateable as according to Totton (1960) and Bouillon *et al.* (2006) only *P. physalis* exist as a species and this view is supported by the Integrated Taxonomic Information System (http://www.itis.usda.gov) however, Collins (2002), Mandojana (1990), Yanagihara *et al.* (2002) and Alam *et al.* (2002) all use the name *P. utriculus* in their papers highlighting the taxonomic issues already present within this genus.

Compounding this confusion, molecular data have revealed that another jellyfish species, *Aurelia aurita* (the moon jellyfish), historically considered to be a cosmopolitan species with little geographic morphological variation is actually a complex of seven closely related species (Dawson & Jacobs 2001). Like *A. aurita, Physalia* inhabits all the world's oceans with little morphological variation (Totton 1960). The last taxonomic review of *Physalia* was conducted by Totton in 1960 and little detailed taxonomic study of the genus has been undertaken since. This suggests that *Physalia* may also turn out to be a species complex, initially identifiable through molecular techniques and then by a thorough revision of the taxonomy as part of an "integrated taxonomy" approach, as suggested by Dayrat (2005).

1.3 Surf Lifesaving New Zealand and *Physalia*

For the past 99 years New Zealand beaches have been patrolled by volunteer surf lifeguards who take care of many incidents that range from complex water rescues to providing simple elements of first aid. Currently, there are 72 surf lifesaving clubs in New Zealand that patrol from approximately late October until mid March. Because the beaches are patrolled by volunteers, clubs have developed a method of accurately recording all aspects of their work, primarily for funding purposes. The result is a highly professional organisation that has detailed records of every patrol that has been carried out stretching back at least 10 years. Each report records basic environmental conditions (wind direction and intensity, wave height and surf conditions) as well as detailing anything that has happened during the patrol. Within this data there are records of incidents where surf lifeguards have treated members of the public for jellyfish stings. These records can be considered a proxy for the presence of *Physalia* as this is the only genus of jellyfish that is known to be capable of stinging people in New Zealand (Slaughter *et al.* 2009). As with most data that have not been collected for scientific purposes there are several issues concerning its accuracy but because there are few

long term datasets that contain records of jellyfish incidences (Mills 2001) the Surf Lifesaving New Zealand dataset presents an opportunity for further study.

1.4 Artificial Neural Networks

Current knowledge about patterns of marine larvae dispersal patterns is limited particularly for species that have a wider dispersal pattern other than near-coastal (Shanks *et al.* 2003; Kinlan *et al.* 2005). For instance Shanks *et al.* (2003) when suggesting an optimum distance been marine reserves for connectivity could only find 32 taxa that had the necessary dispersal information from which to make a recommendation. From the findings it was noted that all but one taxon had a larvae dispersal distance of less than 400 km. A key determinate of the distance that a species could disperse was if the larvae were able to feed (Kinlan et al. 2005). With species that had non-feeding larvae dispersal, estimates were approximately 30km, whereas for feeding larvae, it was approximately 100km. Because of the lack of information on pelagic species with large scale dispersal capability the majority of attempts to model marine larvae dispersal has focused on coastal species.

Models that have been used to determine marine larvae dispersal include those based on Eulerian and Lagrangian flow models (Cowen *et al.* 2000; Siegel *et al.* 2003; Daewel *et al.* 2008), differential equation models (Eckman 1996) and/or qualitative models based on direct observation (Olson 1985). A new development in modelling marine larvae dispersal has been to use high-resolution ocean circulation models to generate individual-based models (Cowen *et al.* 2006). The accuracy and realism of such models depends on the accuracy of the input data particularly the biological information that is available. The study by Cowen et al. (2006) was able to define pelagic larval duration, larval behaviour in terms of vertical and horizontal swimming capabilities, and adult spawning strategies within the model increasing the relevance and accuracy of the findings. In cases where little is known about the biology

and/or dispersal patterns of the target species it may be necessary to use other data-driven methods to identify parameters that can be then be used into an individual-based model. *Physalia* inhabits a complex environment that is highly variable so it is necessary to consider a large number of variables to determine those most likely to affect the populations. Furthermore, each variable potentially includes significant noise masking true patterns. Artificial Neural Networks (ANN) have shown considerable promise in their ability to model data, especially noisy and incomplete ecological data (Lek *et al.* 1996; Olden & Jackson 2002b; Joy & Death 2004; Cocu *et al.* 2005) however, their use in ecology is still not widely accepted despite having been shown to out-perform conventional approaches (Lek *et al.* 1996; Brosse *et al.* 1999; Mutanga & Skidmore 2004).

ANNs were designed as simple models of the human brain mimicking the way it can tackle complex problems (Crick 1989). The brain of a human consists of over 100 billion neurons that are interconnected to form a complex network that is able to organise and recognise complex sensory inputs. Furthermore, when the brain repeatedly receives similar sensory inputs it is able to recall previous responses to the given stimuli and over time optimise its response through learning. An ANN is a computer algorithm that mimics that process.

There are many types of neural networks, each with their own particular purpose. For example, self organising maps (SOM) are clustering and vector quantisation algorithms often used to identify key variables and can be compared to principal component analysis (Kohonen 1990; Brosse *et al.* 2001). One of the more common ANN, often used for prediction, is the multilayer perceptron (MLP) (Rumelhart & McClelland 1986; Lek *et al.* 1996). An MLP is made up of three primary layers or groups of artificial neurons (input, hidden and output) (Figure 1.2). The initial layer is termed the input layer. The input layer represents the data entering the network. Each variable that is input into the network is assigned a node or neuron

within the input layer. The next layer is the hidden layer; this may be a single layer or group of layers depending on the complexity of the network. The hidden layer also has neurons embedded within it. The number of neurons is decided upon by the researcher. Each individual neuron in the input layer is connected to every neuron in the hidden layer. Each neuron processes information through the use of mathematical functions linked within a network (Batchelor 1998) (Figure 1.3). As in the animal neuron, each processing element receives input (μ_i in Figure 1.3) from many other neurons. Each input, μ_i, is multiplied by its associated weight or coefficient that dictates the strength or size of the input. The modified inputs are summed and the result (a) is further modified by a transfer function f(a) that is either fed forward to other neurons or becomes the network's output. The data that each input neuron contains is processed through the hidden layer neurons where the data is usually summed and a nonlinear function (activation function) is applied to the sum resulting in a value being obtained for each variable. The information that is produced in the hidden layer is then transferred to the output layer. The output layer normally consists of a single neuron representing the desired output from the model. If more than one output is desired then additional output neurons can be added.

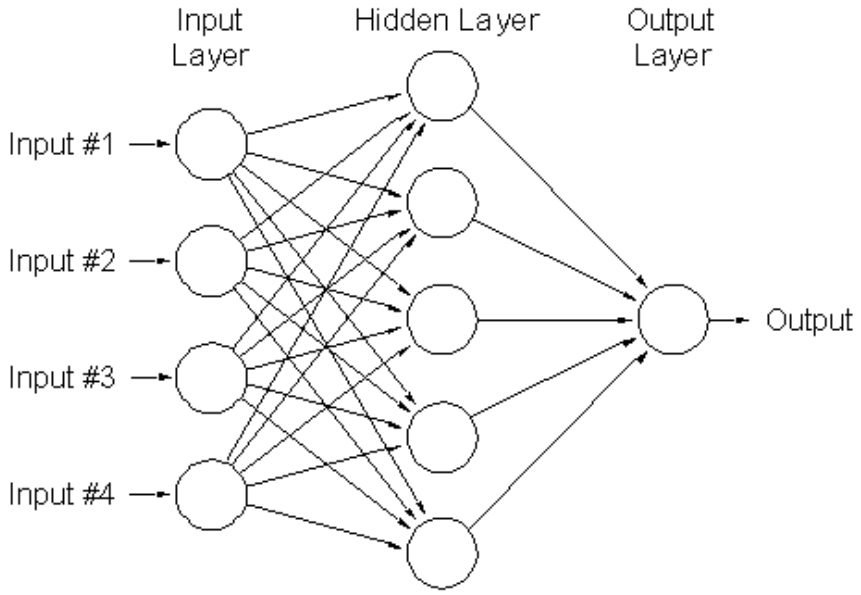


Figure 1.2: Visual representation of a multilayer perceptron neural network consisting of an input layer one hidden layer and an output layer.

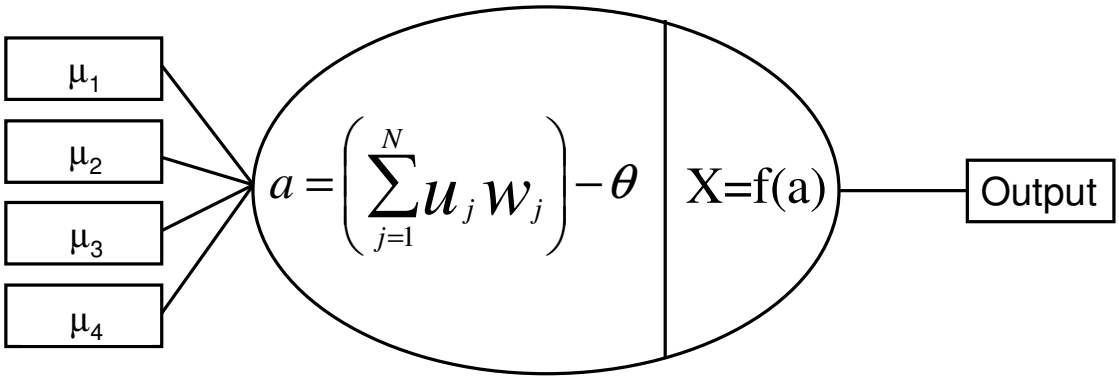


Figure 1.3: Diagrammatic representation of a single artificial neuron. The symbols $\mu_{1\text{-}4}$ represent input variables.

Before the data is processed by the network it is randomly broken into two or three subsets. The largest of the subsets is referred to as the training subset as this is the data that is repeatedly processed by the network in the training stage. The training stage refers to the process by which the network learns patterns in the data. To improve the accuracy of the model, the backpropagation algorithm was developed (Rumelhart *et al.* 1986). The backpropagation technique is a type of supervised training. During training the network algorithm adjusts the weights of the network connections over many hundreds of iterations to minimise by least squares, the difference between the network output (prediction) and the observed data. To further calibrate the network, output is checked using the least squares method against the second subset, the test set, and the connection weight associated with each neuron is adjusted in series of iterations to optimise the model output. In other words, the weights are adjusted to minimise the error between the network or modelled output and the observed response. This process is commonly referred to as training the network. While the least squares method is used to fit standard statistical models, the difference with ANN models is that there is greater complexity. ANN models learn by iteration and can model any non-linearity in the data. The production or validation subset is a subset of the data that is set aside from the training process. Once the network is trained then the network predictions are

compared with the observed data in the production set to test the network's ability to generalise and adapt to new data.

The speed that a backpropagation network learns is controlled by the leaning rate parameter. As a neuron passes a value from one layer of the network to the next, the value is modified by the connection weight. Each time the data is processed by the network the weighting of each connection between neurons is adjusted at a rate determined by the leaning rate parameter to either positively or negatively reinforce the connection, changing the importance of each variable to change within the model (Olden & Jackson 2002b). In other words, variables that have large connection weights, either positive or negative, within the model have a larger influence on how that variable affects the output than variables with smaller connection weights. The goal of training the network is to adjust the weights until a global minimum error is reached so that the network is considered optimised. However, if a network converges to a local minimum then misinterpretation of variable contribution and reduced network accuracy may result (Olden & Jackson 2002b). To reduce the problem of convergence to a local minimum and smoothing the transverse of the error surface, momentum has been used (Lek *et al.* 1996; Brosse *et al.* 1999; Olden & Jackson 2002b; Joy & Death 2004). The learning rate that is set to regulate the magnitude of the change in weights and the momentum adds a proportion of the pervious iteration's change to the contention weights to the current iteration's change. The result of training a network with an appropriate learning rate and momentum is that there is a high degree of probability of model convergence on the global instead of a local minimum (Olden & Jackson 2002b).

Perhaps the greatest criticism of using ANNs to model ecological data is the lack of explanatory information that the models give in relation to importance of individual variables and interactions that occur between them. This has lead to ANNs being labelled as 'black

boxes' (Olden & Jackson 2002b). The lack of explanatory power is of some concern, as it is difficult to interpret and therefore comment on the relationships that are occurring within the studied system. There have been several approaches proposed to overcome the problem, ranging from the development of the neural network diagrams such as a Hinton diagram which is able to show the strength of the connection weights (Hinton *et al.* 1986) to a method of using the connection weights method proposed by Olden & Jackson (2002b) to rank the average importance of the variables relative to each other after the model was run many times. For a full review of methods used to determine variable importance see Olden & Jackson (2002b), Gevery *et al.* (2003) or Olden *et al.* (2004).

Because of the rapid development with regard to network types and architecture there are no established or recognised guidelines to follow unlike standard statistical modelling techniques. For instance Lek *et al.* (1996) used a simple network containing a single hidden layer that used a sigmoid activation function to model fish populations. Six years latter Visen *et al.* (2002) trialled six network architectures with different levels of complexity and different activation functions to classify grain seed. While the rapid development of ANNs has resulted in a trial and error approach to determine if a network type or particular network architecture is appropriate, most take an experimental approach to the optimisation of the number of neurons. Many studies train networks with different combinations of neurons and then choose the highest performing combinations (Ozesmi & Ozesmi 1999; Visen *et al.* 2002; Liang *et al.* 2003). In contrast, the more traditional statistical techniques have clearly defined criteria for when a particular technique is appropriate and when it is not. It is important to point out that ANNs are still under development with regard to model selection and optimisation. As time progresses and as theoretical considerations develop, general rules and guiding criteria will emerge in the same way as for standard statistical modelling techniques.

The application of molecular techniques to establish phylogenetic relations and identify potential cryptic species is now widespread in ecology (Head *et al.* 1998; Ballard & Whitlock 2004; Baker & Bradley 2006). Additionally, the use of models to answer important questions in ecology is a long established discipline. However, it is unusual that both techniques are combined to answer scientific questions or that one approach is used to validate the information gained from the other. While there are many benefits from using both approaches in any study, for example, independent validation of the identification and possible explanation of patterns in ecological data, the two techniques are very dissimilar and practitioners have widely different skills, training and interests and have limited understanding of the other discipline. As a result, the benefits of a multidisciplinary approach may be lost. In ecology in particular, the greatest benefits of a combined approach will be the earlier identification of important processes and relationships particularly with species or systems that are relatively unstudied.

1.5 Aims

The overall aim of this thesis was to investigate which species of *Physalia* are present in New Zealand and to use ANN modelling to identify and understand the key processes that drive their occurrence on beaches used by the public for swimming. Molecular techniques were used to indicate whether *Physalia* in New Zealand belongs to a complex of two or more species and to identify their distribution. The Surf Lifesaving New Zealand dataset of jellyfish stings (a proxy for *Physalia* presence) was used to generate a predictive ANN model to identify key factors that influence the incidence of *Physalia* at New Zealand beaches and forecast periods of high risk of jellyfish incidence so that surf lifeguards can warn members of the public. By combining both techniques it was expected that a clearer interpretation of *Physalia* distribution in New Zealand coastal waters could be achieved.

The specific objectives were to:

- Determine which species of Physalia are present in New Zealand and map their distribution around the coastline using molecular techniques.
- Identify climate and oceanographic variables that may influence the presence of jellyfish at New Zealand beaches.
- Develop a predictive model to forecast *Physalia* occurrence in New Zealand.

1.6 Thesis structure

This thesis comprises five chapters (Chapters 2-6) written in manuscript format.

Chapter 2 details a molecular analysis of *Physalia* specimens collected from New Zealand, Australia and Hawaii the purpose of which was to determine which species of *Physalia* occur in New Zealand waters. Also detailed is how the results of this study affect the choice of molecular markers for Cnidarians particularly the use of cytochrome c oxidase I (COI) for species identification.

Chapter 3 describes a pilot study that explores the efficacy of using an artificial neural network to model the presence of *Physalia* based on unprocessed oceanographic data

Chapter 4 integrates the knowledge gained in Chapter 3 to improve model accuracy by applying a time lag function to the input data at a local scale and to begin to determine important features in the study system.

Chapter 5 applies the novel use of a modified sliding window analysis to pre-process input data to forecast periods of high probability of *Physalia* occurrence in two regions of New Zealand using ANN. An additional goal of this chapter was to assess the potential for the development of a *Physalia* risk assessment index.

In Chapter 6 the model is expanded to encompass the majority of the New Zealand coastline so that variables likely to influence *Physalia* presence can be determined and compared among regions. This chapter also compares important contributory variables identified by the ANN's against those identified by a Naïve Bayesian Classifier.

Chapter 7 comprises an overall discussion that synthesises the research presented in previous Chapters and presents recommendations for future research.

Chapter 2: Molecular systematics of the genus *Physalia* (Cnidaria: Siphonophora) in New Zealand

2.1 Abstract

Physalia physalis (Portuguese man-of-war) and *P. utriculus* (bluebottle) have both been described as present in New Zealand waters, however reports are often conflicting and the presence of neither species has ever been confirmed. The utility of mitochondrial cytochrome c oxidase I (COI) DNA barcodes for identification of cnidarian species is debatable and has yielded mixed results due to an unusually slow rate of evolution in anthozoans. I seek to clarify which species of *Physalia* are present in New Zealand and to establish whether COI can be used for species identification. Fifty four specimens collected from 13 locations around New Zealand and Australia were sequenced for both COI and the first internal transcribed spacer (ITS1) of the nuclear ribosomal gene cluster. Sequences were analysed using maximum likelihood and split decomposition neighbour networks to determine conflict between clans (the unrooted analog of clades). Three clans were identified from both the COI and ITS sequences, none of which correspond to *P. physalis*. It appears that COI can be used as a species identification tool for non-anthozoan cnidarians and slow evolutionary rates may be confined to the Anthozoa. The results are complex and it is possible that hybridisation has occurred as clans are not consistent between the two genes. Nevertheless, it seems that there are at least three species of *Physalia* present in New Zealand and only one of these is likely to be a named species (*P. utriculus*).

2.2 Introduction

The diversity of jellyfish has often been underestimated due to their morphological simplicity and the historical belief that oceans provide little barrier to gene flow so there are few opportunities for allopatric divergence (Palumbi 1992; Knowlton 2000; Dawson and Jacobs 2001). As a consequence there are believed to be a number of widespread cosmopolitan species in ocean ecosystems. With the growing use of molecular techniques, studies are now showing that species once thought to be cosmopolitan often consist of many cryptic species. For example, Dawson and Jacobs (2001) showed that *Aurelia aurita* (the moon jellyfish), commonly believed to be cosmopolitan with little or no geographic variation, can be reclassified into seven species based on molecular information. Another cosmopolitan species is *Physalia physalis* (L.) (Portuguese man-of-war), which is thought to inhabit all the world's oceans (Lane 1960; Yanagihara *et al.* 2002). There is considerable potential for the genus *Physalia* to contain hidden cryptic diversity as the last taxonomic review was over 50 years ago (Totton 1960). The primary issue that needs to be addressed within this genus is the number and identity of species and the extent of their geographical distributions, as there is significant debate over this.

According to the taxonomy of Cnidaria proposed by Collins (2002), *Physalia* is placed in the Siphonophora and considered a sister taxon to all other siphonophores. *Physalia* taxonomy has been revised many times from Lamark's early revision (Lamark 1801) through to a revision by Totton (1960), however it is still unclear exactly how many species are in the genus, with two species commonly named *P. physalis* and *P. utriculus*. The existence of *P. utriculus* as a distinct species is highly debateable. According to Totton (1960) and Bouillon *et al.* (2006) only *P. physalis* should be recognised as a species and this view is supported by the Integrated Taxonomic Information System (http://www.itis.usda.gov), however Collins

(2002), Mandojana (1990), Yanagihara *et al.* (2002) and Alam *et al.* (2002) all use the name *P. utriculus* in their papers. Driving this debate are overlapping distributions and similar morphological descriptions. *Physalia physalis* is considered to have a global distribution (Totton 1960; Pages and Gili 1992; Bouillon *et al.* 2006) whereas *P. utriculus* is regarded as confined to the Pacific (Yanagihara *et al.* 2002). The morphological characteristics used to differentiate the species are not precise and subject to interpretation. For example, the descriptions of both *P. physalis* and *P. utriculus* are very similar, with two features commonly used to differentiate them; (1) *P. utriculus* has a single main fishing tentacle while *P. physalis* has multiple main tentacles, and (2) *P. physalis* has a larger float (10-25 cm in length) than *P. utriculus* (4-8 cm) (Fenner 1997). The possibility that medusae classified as *P. utriculus* may just be juveniles of *P. physalis* has not been ruled out. However, Totton (1960), after examining individuals from around the world, noted that although there was variation, in his opinion this was not sufficient to indicate additional species, highlighting that identification of individuals is difficult. Molecular techniques, as the initial part of an "integrated taxonomy" approach (Dayrat 2005), provide a possible tool for resolving the taxonomic ambiguity of *Physalia* and determining how many species are present in this genus.

An integrated taxonomic approach to determining whether there are cryptic species in the genus *Physalia* has potential problems, especially with respect to choosing which genes to analyse. To detect and determine potential species boundaries it is important to use genes that evolve rapidly enough to detect such boundaries, furthermore a voucher specimen needs to be deposited for morphological exploration of any molecular results. Le Goff-Vitry *et al.* (2004), Collins *et al.* (2005), Dunn *et al.* (2005), Govindarajan *et al.* (2005), Collins *et al.* (2008) and Moura *et al.* (2008) have all successfully used the mitochondrial gene, 16S rDNA, to examine hydrozoan species boundaries. An alternative approach is the use of mitochondrial gene cytochrome c oxidase I (COI) as a DNA barcode and therefore a universal way of identifying

species as proposed by Hebert, Cywinska *et al.* (2003). However, it has been suggested that COI evolves at a much slower rate in Cnidaria than in other taxa (Hebert, Ratnasingham *et al.* 2003). This view has been based particularly on data from the Anthozoa (Shearer *et al.* 2002; Shearer and Coffroth 2008) where it appears that rates of mitochondrial evolution are up to 20 times slower than in other taxa (Shearer *et al.* 2002). However, recent research by Huang *et al.* (2008) and Govindarajan *et al.* (2005) indicated that the Hydrozoa display normal patterns of evolution. Dawson and Jacobs (2001) suggested that the substitution rate is higher in the scyphozoan, *Aurelia aurita*, and that COI can be used to discriminate potential species. The choice between the use of 16S or COI is compounded by a lack of published sequences to guide selection and validate results. There is one published sequence for *P. physalis* for both 16S (AY935284) and COI (AY937374) in Genbank. These sequences are from the same specimen. There is also one 16S sequence for *P. utriculus* (AY512511). Because COI is capable of determining species boundaries in Hydrozoa and is recognised as a primary gene choice in species identification it is appropriate to use COI alone to identify potential species boundaries in *Physalia*.

An issue with using mtDNA for species identification is that mtDNA diversity is not always correlated with nuclear gene diversity as processes such as greater dispersal by males than females can cause mtDNA to diverge while nuclear genes do not (Moritz 1994). It is therefore is important to use both sources of genetic information when assessing evolutionarily distinct populations (Cronin 1993). The internal transcribed spacer (ITS) is a commonly used region for this purpose as it evolves rapidly and can differentiate between closely related species (Hills and Davis 1986; Tang *et al.* 1996; Collins *et al.* 2000). Moreover ITS has been used for this purpose within the Scyphozoa (Dawson 2003), Anthozoa (Goulet and Coffroth 2003) and Hydrozoa (Zhang *et al.* 2009) to good affect.

P. physalis and *P. utriculus* have both been described as present in New Zealand waters (Carson 1965; Wesrerskov and Probert 1981; Bouillon and Barnett 1999), however reports are conflicting and the presence of neither species has been confirmed. Using DNA sequences from the mitochondrial gene COI and the first internal transcribed spacer (ITS1) of the nuclear ribosomal gene cluster, we seek to determine the diversity and identity of *Physalia* species in New Zealand waters, and in particular whether there is evidence for the presence of *Physalia utriculus*.

2.3 Materials and Methods

2.3.1 Sampling

A total of 54 specimens were collected from 13 locations around New Zealand and Australia (Figure 1). Specimens were either collected directly from the ocean or from the beach as they washed ashore. Excess sea water was removed from each specimen by blotting with a paper towel before being placed in 100% ethanol and stored at -20ºC. A brief morphological analysis was conducted to establish if there were any specimens that conformed to the *P utriculus* morphology, as described by (Fenner 1997), with a predominant fishing tentacle of up to 1m in length and pneumatophore, up to 50mm in length. Further COI sequences were sourced from Dr Brenden Holland (University of Hawaii) for Brisbane (3), Hawaii (2) and Midway (1), one of which Dr Holland identified as *P. utriculus* but was assigned the code Midway 1 throughout the analysis because of the uncertain taxonomic status of *P. utriculus*. One COI sequence labelled *P. physalis* (Atlantic Ocean) was obtained from GenBank (AY937374).

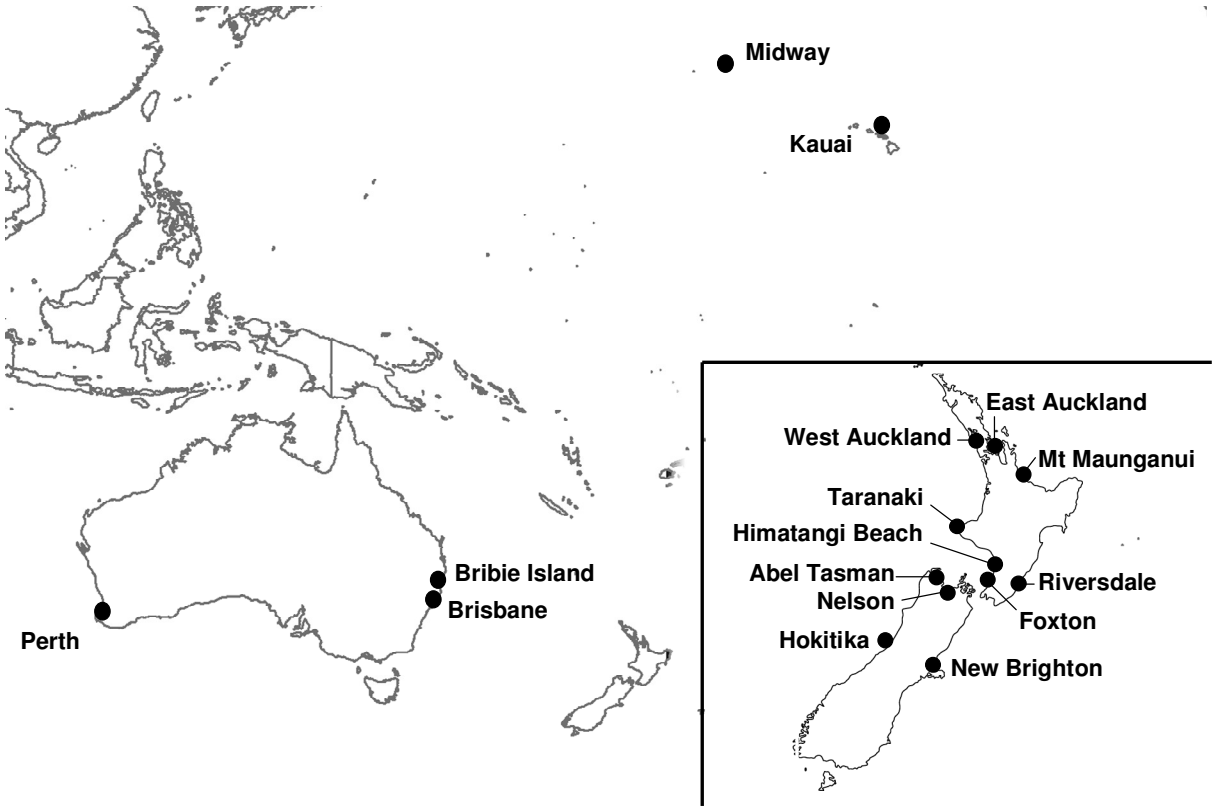


Figure 2.1: Locations of *Physalia* collection sites

2.3.2 DNA extraction and sequencing

Tissue from the oral arms or main fishing tentacles was cut from each specimen and total genomic DNA was extracted from the cut tissue using either the DNeasy® Tissue Kit (Qiagen) or the AxyPrep Multisource Genomic DNA Miniprep Kit (Axygen) following the manufacturers' protocols. 2.5µl of extracted DNA was amplified by polymerase chain reaction (PCR) in a total volume of 25µl with 2.5µl 10 x PCR buffer (Qiagen), 2.5µl 8mM dNTPs (i.e. 2mM of each, New England Biolabs), 1.2µl of each 10µM primer (Invitrogen) and 0.2µl (1 unit) of Taq polymerase (5 unit/µl; Qiagen). The PCR reaction began with two minutes denaturation at 94ºC, followed by 33 cycles of 40 seconds at 92ºC, 40 seconds at 45ºC and 90 seconds at 72ºC, finishing with a five minute extension at 72ºC and cooling to 4ºC. Mitochondrial cytochrome c oxidase subunit 1 (CO1) was amplified using the primers

HCO2198 (5'-TAAACTTCAGGGTGACCAAAAAATCA-3') and LCO1490 (5'-GGTCAACAAATCATAAAGATATTGG-3') (Folmer *et al.* 1994). The internal transcribed spacer 1 (ITS1) region of the nuclear ribosomal gene cluster was amplified using the primers Cas18sF1 (5'-TACACACCGCCCGTCGCTACTA-3') and Cas5p8sB1d (5'-ATGTGCGTTCRAAATGTCGATGTTCA-3') (Ji *et al.* 2003). Amplifications were confirmed by electrophoresis. PCR products (2-3µl) were sequenced (10µl total volume) using 0.5µl BigDye™ (Applied Biosystems), 2µl sequencing buffer and 0.8µl of 10mM primer. The sequencing reaction began with a 1 minute denaturation at 96ºC, followed by 25 cycles of 10 seconds at 96ºC, 5 seconds at 50ºC and 4 minutes at 60ºC, then cooling to 10ºC. Sequencing products were purified by ethanol precipitation, air dried and run on an ABI-PRISM® 377 automated sequencer (Applied Biosystems) according to manufacturer's instructions. Both strands were sequenced to improve accuracy. Sequences were obtained for both genes for all specimens except four specimens New Brighton 103, Auckland East 14 and Nelson 1 where only COI was obtained and Taranki 1 where only ITS1 was obtained.

2.3.3 Phylogenetic analysis

Sequences were aligned using PRANK (Loytynoja and Goldman 2005) with default parameters and in four cases of obvious misalignment of large sections of sequence, were adjusted by eye. The total length of the alignment was 566 bp for COI and 456 bp for ITS1. Phylogenetic trees were constructed using maximum likelihood (ML) as implemented in the program PAUP*4 (Swofford 2002) using a heuristic search with starting trees obtained by stepwise addition and tree bisection-reconnection (TBR) branch swapping. Trees were also constructed using distance methods and maximum parsimony but as these were broadly congruent with the ML trees only the likelihood trees are presented. Due to the unavailability of close outgroups, trees were left unrooted. Parameters for the ML model were determined by MODELTEST using AIC (Posada and Crandall 1998). One thousand bootstrap replicates

were performed to estimate clan support. As the trees presented in this paper are all unrooted we prefer to use the term "clan" (*sensu* Wilkinson *et al.* (2007)) rather than "clade" to denote the unrooted analog of a monophyletic group. Split decomposition neighbour networks were calculated for both genes using SplitsTree4 (Hudson and Bryant 2006) to assess potential conflict within the datasets. For genes where areas of conflict were identified a sliding window was implemented across the gene. Window lengths starting at 50 base pair (bp) and increasing by 50bp up to 75% of the length of the gene were assessed. For each window 1000 parsimony bootstrap replicates were performed in PAUP*4 using the fast stepwise addition option. For each window the bootstrap support for each node was extracted using a MATLAB 7.6.0 script and conflicting signals were isolated.

2.4 Results

Results of the morphological analysis are displayed in Table 2.1, but the species identifications to not match the clans identified by either COI or ITS (see below). The best-fit models of nucleotide sequence substitution, estimated by MODELTEST, are for COI the K81uf+I+Γ model (nucleotide frequencies = A: 0.3839, C: 0.1416, G: 0.1683, T: 0.3062 , proportion of invariant sites = 0.2100, gamma shape parameter = 0.7202, rA-C = 1.000, rA-G = 9.8019, rA-T = 3.7476, rC-G = 3.7476, rC-T = 9.8019, rG-T = 1.000), and for ITS1 the TVM+G model (nucleotide frequencies = A: 0.2699, C: 0.2089, G: 0.2385, T: 0.2828, proportion of invariant sites = 0, gamma shape parameter = 0.3396 rA-C = 1.0423, rA-G = 0.7742, rA-T = 1.3478, rC-G = 0.2027, rC-T = 0.7742, rG-T = 1.000).

Table 2.1: Specimen morphology and corresponding COI and ITS clans and specimen name associated with collected specimens. For number of tentacles, "?" indicates that the original tentacle number is unknown. For length of bell, "short" = <50mm and long = >50mm.

Specimen (code)	Collection date	COI clan	ITS clan	Number of tentacles	Length of bell
New Brighton 2 (NB2)	13/2/05	1	I	multiple	short
New Brighton 3 (NB3)	13/12/05	1	I	multiple	short
New Brighton 4 (NB4)	13/12/05	1	I	multiple	short
New Brighton 5 (NB5)	13/12/05	1	I	multiple	short
New Brighton 7 (NB7)	13/12/05	1	I	multiple	short
New Brighton 8 (NB8)	13/12/05	1	I	multiple	short
New Brighton 9 (NB9)	13/12/05	1	I	multiple	short
New Brighton 10 (NB10)	13/12/05	1	I	multiple	short
New Brighton 22 (NB22)	6/1/08	1	I	multiple	short
New Brighton 32 (NB32)	6/1/08	1	I	multiple	short
New Brighton 103	16/1/06	2	-	multiple	short
New Brighton 109	16/1/06	2	I	multiple	short
Foxton 1 (F1)	19/12/06	1	I	multiple	short
Foxton 2 (F2)	19/12/06	1	I	multiple	short
Foxton 3 (F3)	19/12/06	1	I	multiple	short
Foxton 5 (F5)	19/12/06	1	I	multiple	short
Himatangi Beach 1	10/12/06	1	I	multiple	short
Himatangi Beach 2	10/12/06	1	I	multiple	short
Himatangi Beach 3	10/12/06	1	I	multiple	short
Himatangi Beach 5	10/12/06	1	I	multiple	short
Taranaki 1 (T1)	2/2/06	-	I	multiple	short
Taranaki 2 (T2)	2/2/06	1	I	multiple	short
Taranaki 3 (T3)	2/2/06	1	I	multiple	short
Taranaki 4 (T4)	2/2/06	1	I	multiple	short
Riversdale 1 (R1)	19/2/06	3	III	single	short
Riversdale 2 (R2)	19/2/06	1	I	multiple	short
Riversdale 3 (R3)	19/2/06	1	I	multiple	short
Riversdale 4 (R4)	19/2/06	3	III	multiple	long
Riversdale 13 (R13)	29/11/06	3	III	multiple	short
Riversdale 18 (R18)	29/11/06	3	III	single	short
Riversdale 20 (R20)	29/11/06	1	I	multiple	short
Hokitika 1 (H1)	26/11/05	1	I	single	short
Hokitika 2 (H2)	26/11/05	1	I	multiple	short
Hokitika 3 (H3)	26/11/05	1	I	single	short
Hokitika 4 (H4)	26/11/05	1	I	?	short
Hokitika 53 (H53)	3/3/06	1	I	single	short
Mt Maunganui 1 (MM1)	20/3/06	3	III	multiple	long
Mt Maunganui 2 (MM2)	20/3/06	3	III	multiple	long
Mt Maunganui 3 (MM3)	20/3/06	3	III	multiple	long

Mt Maunganui 4 (MM4)	20/3/06	3	III	multiple	long
Western Australia 1	16/2/07	2	II	single	short
Western Australia 2	16/2/07	1	I	single	short
Western Australia 3	16/2/07	1	I	single	short
Western Australia 6	16/2/07	1	I	single	short
Brisbane 7 (B7)	24/11/07	2	II	single	short
Brisbane 9 (B9)	24/11/07	2	II	single	short
Brisbane 10 (B10)	24/11/07	2	II	single	short
Nelson 1 (N1)	18/12/07	2	I	multiple	short
Nelson 2 (N1)	18/12/07	2	I	multiple	short
Nelson 3 (N1)	18/12/07	2	-	multiple	short
Abel Tasman 2 (AT2)	7/3/08	1	I	single	short
East Auckland 10 (AE10)	7/12/07	1	I	multiple	short
East Auckland 14 (AE14)	5/2/08	1	-	?	short
East Auckland 15 (AE15)	9/3/08	1	I	multiple	short
West Auckland 2 (AW2)	15/11/07	1	I	multiple	short

2.4.1 Cytochrome c oxidase I

Three clans are identified from the COI sequences (Figure. 2.2) with a minimum bootstrap support of 82%, with the Genbank voucher sequence for *P. physalis* not grouping with any of the clans and forming an isolated branch. The specimen (Midway 1) identified by Brenden Holland (University of Hawaii) as *P. utriculus* groups with a number of specimens from the entire geographic range sampled (clan 2). Specimens from Riversdale and Mt Maunganui form a distinct clan (clan 3) that has a high degree of internal structure when compared to the other clans. This clan also has the smallest geographic range of all the clans and is only found in the northeastern and eastern areas of the North Island of New Zealand. The remaining specimens formed the final clan (clan 1) which had specimens from all Australasian locations except Mount Maunganui. Pairwise genetic distances between clans varied between 7.3% and 12.6% (Table 2.2). Pairwise genetic distances within clans were all <1.5% except for clan 3, which had a mean pairwise distance of 6.1%, despite having the smallest geographical range. Moreover, there was evidence of a barcoding gap (i.e. a distinct gap between the distributions of within-clan and between-clan distances (Meyer & Paulay 2005)) for clans 1 and 2 but not clan 3 (Figure 2.3).

Table 2.2: Mean pairwise genetic distances within and between clans of *Physalia*

COI		ITS	
Clan	Sequence divergence (%)	Clan	Sequence divergence (%)
Clan 1	1.1	Clan I	0.8
Clan 2	1.3	Clan II	1.2
Clan 3	6.1	Clan III	0.9
Clan 1 and 2	9.4	Clan I and II	2.0
Clan 1 and 3	11.6	Clan I and III	1.8
Clan 2 and 3	11.6	Clan II and III	2.2
Clan 1 and *P. physalis*	11.6		
Clan 2 and *P. physalis*	7.3		
Clan 3 and *P. physalis*	12.6		

2.4.2 Internal transcribed spacer

The ITS1 tree (Figure 2.4) is similar to the CO1 tree but there are significant differences. Three clans are identified with a minimum bootstrap support of 55%, however it was not possible to obtain voucher ITS1 sequences for *P. physalis* or *P. utriculus*, so this information is missing. To avoid confusion between COI and ITS clans, ITS clans are referred to by Roman numerals. The only clan that is identical to a COI clan and has strong bootstrap support (79%) is clan 3/III. ITS clan I contains specimens from both COI clans 1 and 2. The specimens from COI clan 2 found in ITS clan I were collected from Nelson and New Brighton and are the only specimens from COI clan 2 found in New Zealand. The remained of COI clan 2 with individuals from Western Australia and Brisbane form clan II but there is minimal support for this clan (55%) and the node has been collapsed. The pairwise genetic distances between and within ITS clans were similar, with mean pairwise distances within clans ranging from 0.8% to 1.2% and mean pairwise distances between clans ranging from 1.8% to 2.2% (Table 2.2) and significant overlap between the distributions (i.e. no barcoding gap) (Figure 2.5).

Split decomposition neighbour networks (Figures 2.6 and 2.7) support the ML findings and highlight the clan structure found in the ML analysis. Moreover, the networks show that there is some conflict between ITS clans II and III particularly regarding the placing of the WA1 (Western Australia) sequence. Results of the sliding window analysis indicate that a window length of 200bp of is optimal to assess potential conflict within the ITS sequences. Percentage bootstrap support for the inclusion of WA1 in ITS clan II or ITS clan III is shown in Figure 8. This analysis shows clear conflict for the inclusion of WA1 in either clan II or clan II. Bootstrap support for the inclusion of WA1 in ITS clan II decreased below 50% for windows starting at bp72 to bp132 with a corresponding rise above 50% in support for the inclusion of WA1 in ITS clan III, i.e the middle third of this sequence supports the placement of WA1 in ITS clan II, while both ends support the placement of WA1 in ITS clan III. This conflict is likely to be the cause of the overall low bootstrap support for ITS clan II (Figure 4) because when WA1 is removed bootstrap for ITS clan II increase to 91%.

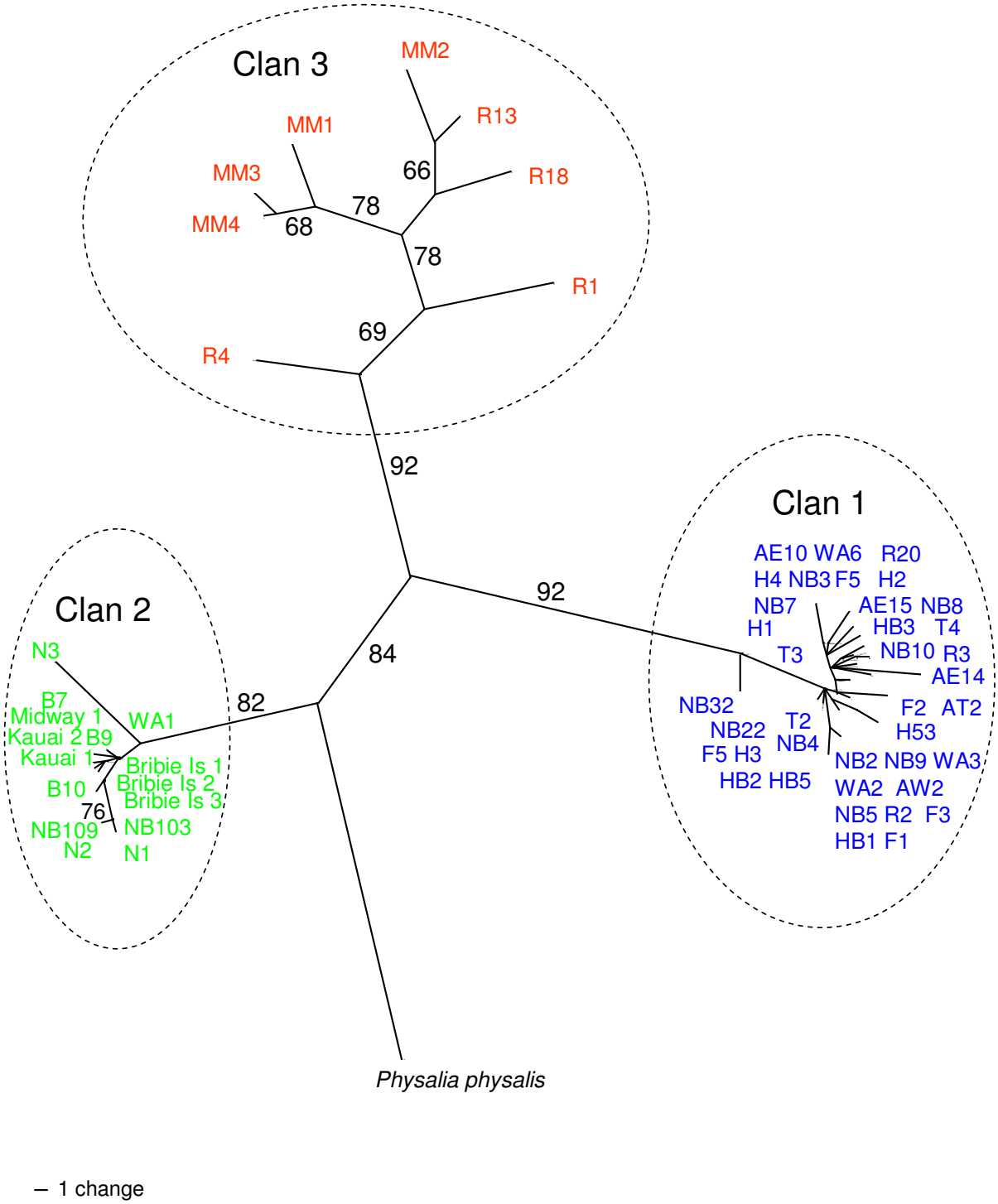


Figure 2.2: Unrooted maximum likelihood tree for COI. Numbers on branches indicate bootstrap support (1000 replicates).

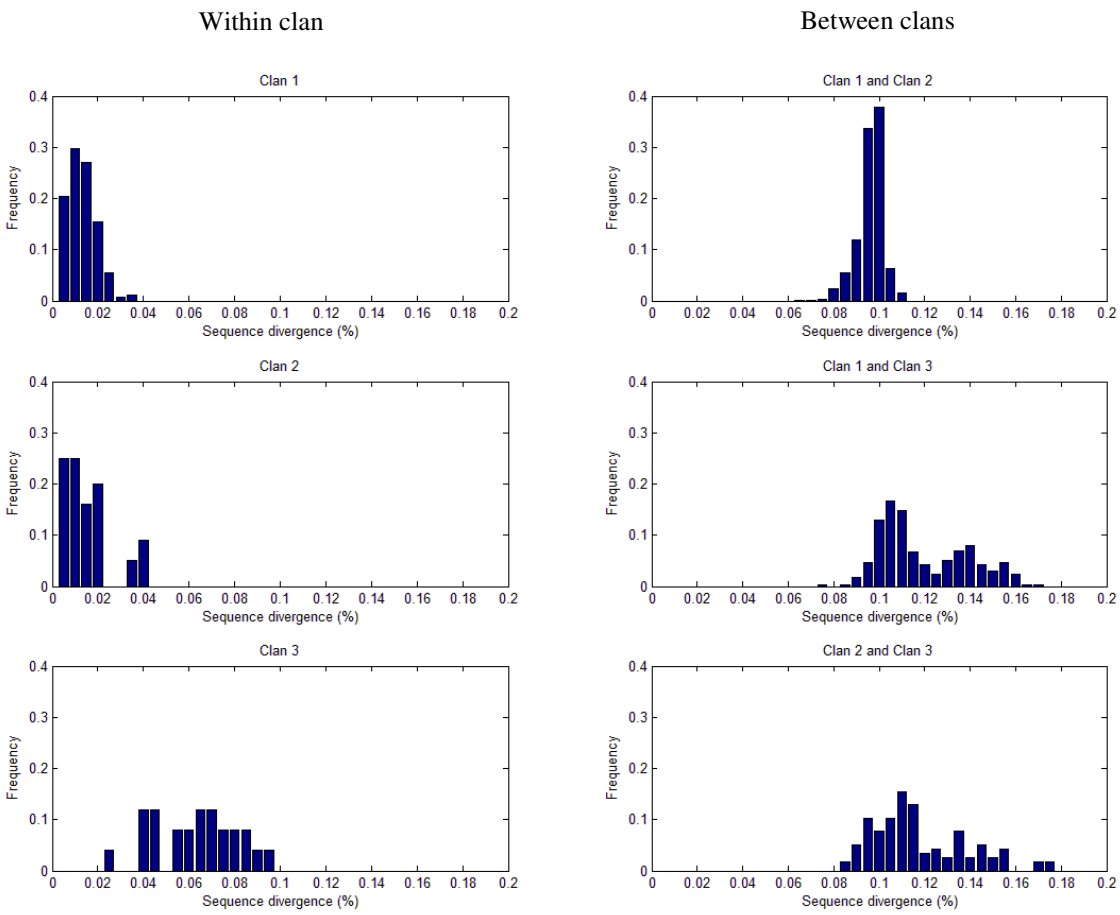


Figure 2.3: Distributions of within-clan and between-clan pairwise genetic distances for COI

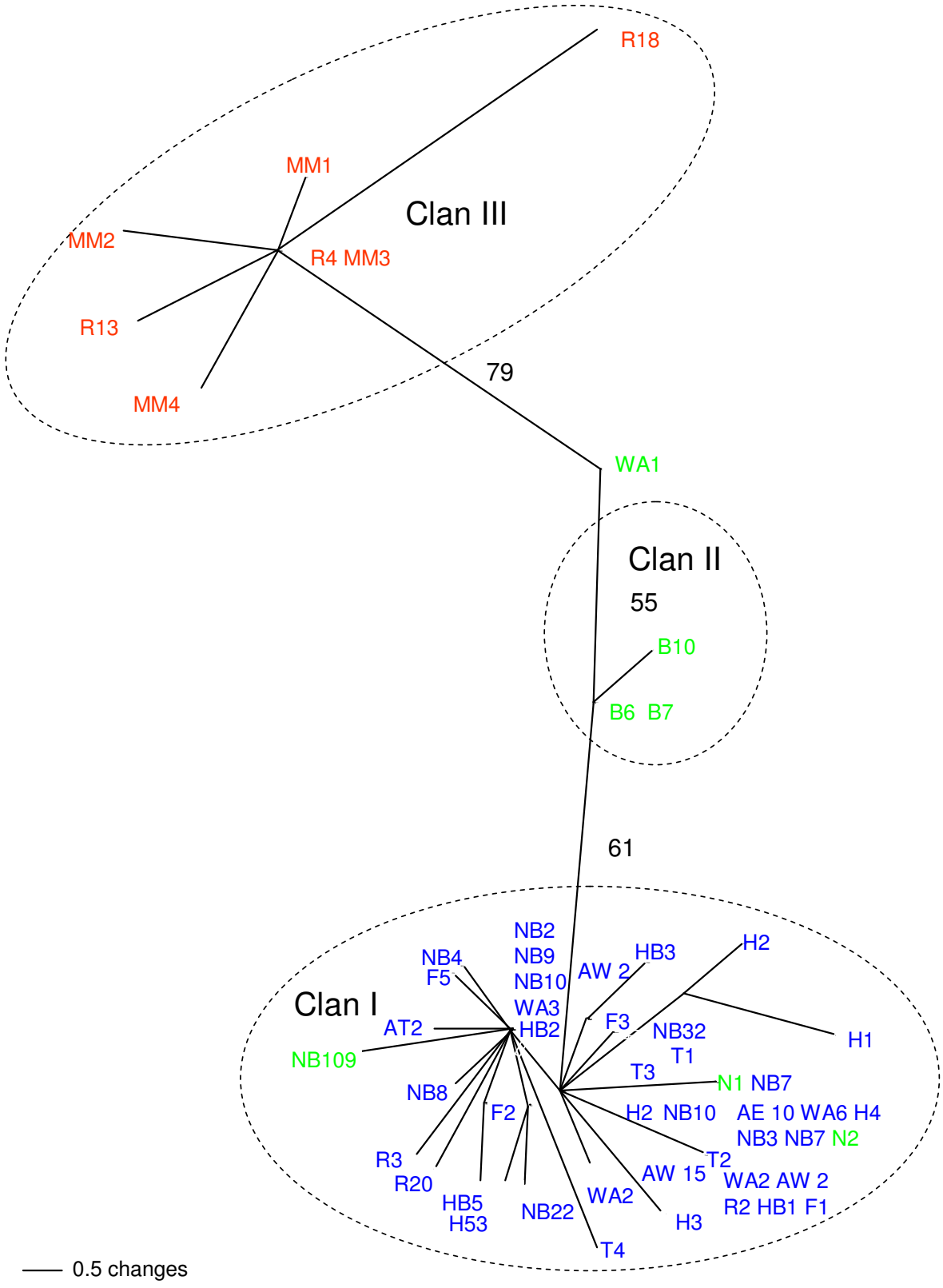


Figure 2.4: Unrooted maximum likelihood tree for ITS1. Numbers on branches indicate bootstrap support (1000 replicates).

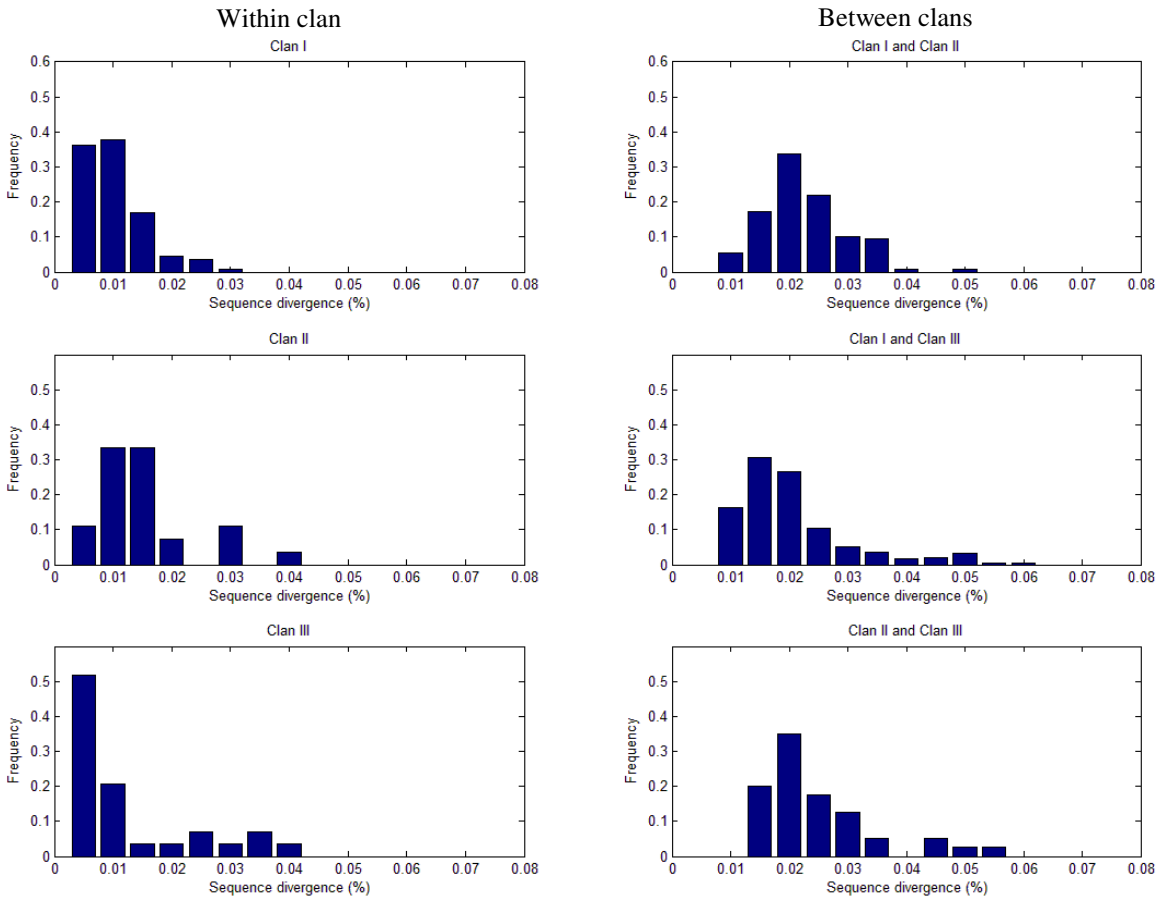


Figure 2.5: Distributions of within-clan and between-clan pairwise genetic distances for ITS1.

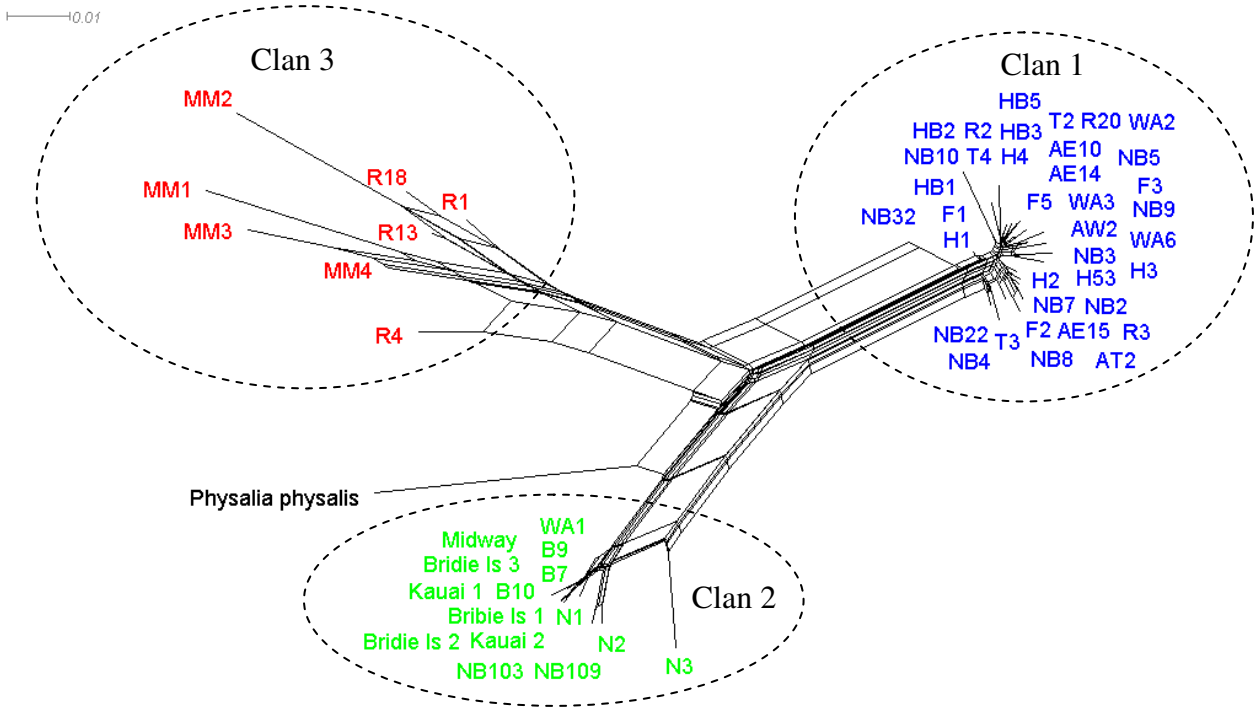


Figure 2.6: Split decomposition neighbour network for COI.

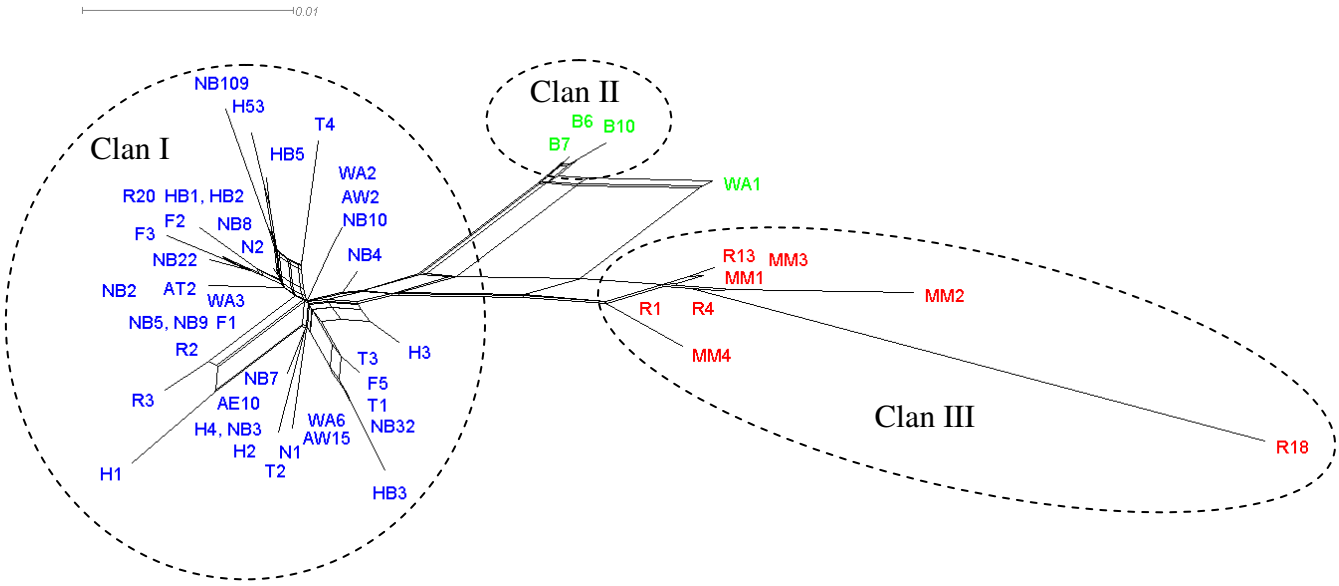


Figure 2.7: Split decomposition neighbour network for ITS1.

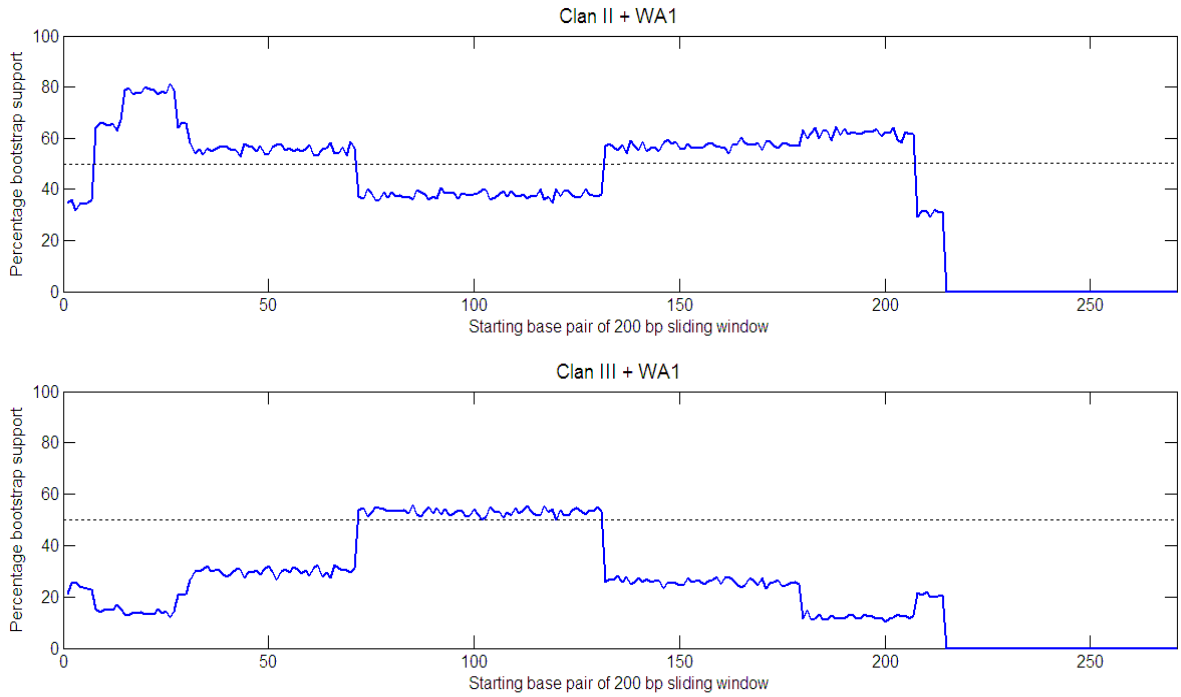


Figure 2.8: Percentage bootstrap support across *Physalia* ITS sequences for conflicting nodes using a 200 base pair sliding window (1000 replicates per window).

2.5 Discussion

As for many species with ambiguous taxonomy, molecular techniques are ideal for investigating the phylogenetic structure of *Physalia* in New Zealand waters. However, in this study they have generated a surprisingly complex taxonomic picture. Nevertheless, the results provide a solid base for further molecular and morphological investigation of the genus at both local and global scales. This is necessary to resolve the taxonomic structure of the genus.

2.5.1 New Zealand species

It appears that *Physalia* that inhabit New Zealand coastal waters are a complex, for which it is impossible to determine the exact number of species or the corresponding species names. Because of the strong molecular evidence for a species complex, a taxonomic revision of *Physalia* is necessary. Morphological revision is necessary as specimens that conform to both

P. physalis and *P. utriculus* morphology group together in clans 1/I and 2/II. Clan 3/III specimens only conform to *P. physalis* morphology (Table 2), but this may be because of the small number of individuals examined. The lack of specimens from clans 1/I and 2/II conforming to a distinct morphology highlights the ambiguity of the characteristics used to identify species, however, a more in-depth analysis is beyond the skill of this author and the scope of the study.

An integrated taxonomy approach requires the use of multiple taxon identification techniques to differentiate species. From the molecular results obtained here it is hypothesised that the *Physalia* complex that inhabits New Zealand coastal waters consists of at least two species because cytonuclear discordance between clans 1/I and 2/II creates uncertainly for these clans as separate species. Clan 3/III is well supported for both genes and unlike other clans shows substantial internal structure, and the concordance between COI and ITS indicates that it has not recombined with other clans. Clan 3/III was the only one found in the northeast of the North Island, a locality that raises the possibility of a separate source area. Brodie (1960) released over 10,000 float cards to assess the surface ocean currents around New Zealand and it is reasonable to assume that *Physalia* would display similar movement patterns to these cards as they are likely to be influenced by wind and currents in a similar way. Cards released from the North Cape drifted down the east coast of the North Island and via the East Auckland Current to the East Cape. Past the East Cape the East Current continues south until it meets the Canterbury Current in the south of the North Island (Gardner 1961). This pattern of the currents may explain why clan 3/III was found in East Auckland, and Riversdale had both clans present. Moreover, there are two permanent eddies to the north of the Bay of Plenty that have the ability to act as a recycling barrier, as described for New Zealand rock lobster (*Lasus edwardsii*) larvae (Chiswell and Roemmich 1998). Rock lobster larvae that were spawned in the area were shown to be caught in the eddy system for as long as three

years, allowing them to be recruited as juvenile lobsters. If populations of *Physalia* have become trapped in the eddy system then this could explain why the clan found there shows so much internal structure, as subpopulations could have been isolated from each other for significant periods of time. Exploration and formal identification of this clan is required to establish whether it represents a new species. The possibility that the clan is a cryptic species (or species complex) is strengthened by the observation that cryptic species have been found in other common jellyfish such as *Aurelia aurita* (Dawson and Jacobs 2001) and the upside-down jellyfish, *Cassiopea* (Holland *et al.* 2004).

2.5.2 Cytonuclear discordance

The current global initiative to use COI as a stand-alone species discriminator has been predominately successful across most taxa attempted (Ward *et al.* 2005 ; Kerr *et al.* 2007), but it is recognised by the barcoding community that there will be exceptions. In particular, groups that exhibit hybridisation, ancestral polymorphism and pseudogenes pose potential problems for barcoding (Bensasson *et al.* 2001; Funk and Omland 2003). Hybridisation has influenced the evolution of the Anthozoa, particularly in genera such as *Acropora* and *Alcyonium* (van Oppen *et al.* 2000; McFadden and Hutchinson 2004). Our results suggest that hybridisation may also have occurred in *Physalia*, as three individuals from Nelson and New Brighton, whose mtDNA places them in COI clan 2, have nuclear DNA that places them in ITS clan I, and the WA1 ITS sequence shows conflicting signal with different parts of the sequence placing it in two different clans, suggesting that it may be of hybrid origin. (McFadden 1999) suggested that *Octocorallia* possess many traits that predispose them to hybridisation including closely related, morphologically similar species with overlapping ranges and reproductive periods. It appears that *Physalia* share the majority of these traits. Clans have significant geographical overlap and Totton (1965) concluded that although there was morphological variation, it was not enough in his opinion to indicate multiple species

within the genus. There also appears the possibility of reproductive overlap as gonodendra (reproductive structures) are produced continuously on new individuals and therefore it is assumed that there is a steady gamete supply. Moreover, individual specimens were collected from all clans throughout the sampling period indicating that gametes from different clans could be present at the same time.

It seems reasonable to suppose that COI clans 1 and 2 may have originated in the Tasman Sea and based on the dominant west to east wind flow over New Zealand (Heath 1985) were transported through Cook Strait and down the east coast of the South Island. As both COI clans are present on the Australian east coast, the most likely hybrid zone is located somewhere in the Tasman Sea. As little is known about the developmental rates of individuals and their rate of movement under wind and current conditions it is impossible to determine a localised area for further targeted sampling. It may be possible to use computer modelling to establish how such factors effect the dispersal of *Physalia* to give a clearer indication of where potential hybrid zones may be located.

An alternative explanation for the cytonuclear discordance is that ancestral polymorphism has been maintained in some lineages within the genus. For this to have occurred, the common ancestor of COI clans 1 and 2 would need to have had two mitochondrial alleles, one of which became fixed in clan 2 (or the other is not represented in our sample) while clan 1 retained both. Although this is a plausible scenario and cannot be discounted, we consider it less likely than hybridisation. Hybridisation is prevalent in the Anthozoa and is thought to have an important role in speciation (Willis *et al.* 2006). Hydrozoa display similar traits to corals when in the polyp stage of their lifecycle and the possibility that a similar mechanism has occurred in *Physalia* seems likely, although without further study this is impossible to verify.

With the results indicating that hybridisation could have occurred, this highlights the need to include multiple genes, particularly from the nuclear genome, along with COI to gain an overall view of the phylogeny of the group before ascertaining whether COI can be used as a species discriminator. From these results it appears that COI alone is not an appropriate species discriminator within *Physalia*.

2.6 Conclusion

This study attempts to clarify which species of *Physalia* occur in New Zealand waters using molecular techniques. The results indicate that there are at least two and potentially three or more species present, and only one of these is likely to be a named species (*P. utriculus*). Furthermore, the results raise questions about the taxonomy of the genus to the point that it may be necessary to undertake a global review of the taxon to establish the correct relationships. The use of COI for barcoding was also assessed and although mtDNA evolution occurs at rate where species identification is possible, COI should not be used as a stand-alone species identifier until the issues with the taxonomy are resolved.

Chapter 3: Using Multi-Layer Perceptrons to predict the presence of jellyfish of the genus *Physalia* at New Zealand beaches

3.1 Abstract

The apparent increase in number and magnitude of jellyfish blooms in the oceans of the world has lead to concerns over potential disruption and harm to global fishery stocks. Additionally, jellyfish causes problems for bathers on swimming beaches throughout the world, as their sting cause discomfort or even death. Because of the potential harm that jellyfish populations can cause to humans and economic activity and to avoid impact it would be helpful to model jellyfish populations so that species presence or absence can be predicted. Data on the presence or absence of jellyfish of the genus *Physalia* was modelled using Multi-Layer Perceptrons (MLP) based on oceanographic data. Results indicated that MLP are capable of predicting the presence or absence of *Physalia* in two regions in New Zealand and of identifying significant biological variables.

3.2 Introduction

Jellyfish blooms have the potential to change the species composition in an ecosystem through altering the availability of food resources, and therefore, threatening fisheries (Graham *et al.* 2001). Furthermore, it has been reported that jellyfish populations are increasing in both the intensity and frequency of blooms (Mills 2001; Purcell 2005). As well jellyfish often cause beach closures because of the risk to public (Purcell *et al.* 2007). To begin to understand potential impacts to marine ecosystems the investigation of factors that contribute to the formation of a bloom and that determine jellyfish movement is necessary. By understanding the factors that influence the movement and distribution of individuals there is

potential to develop models to predict where and when jellyfish are likely to occur. The ability to predict jellyfish occurrence will allow warnings to be issued to safeguard fisheries and mitigate the threat of jellyfish stings on swimmers at beaches in coastal regions.

The genus *Physalia* is one of the most commonly found jellyfish on New Zealand beaches, and is the most commonly found stinging jellyfish. *Physalia* is considered to be one of the more primitive living jellyfish as it lacks many of the morphological characteristics associated with species that evolved later (Collins 2002; Dunn *et al.* 2005). In particular *Physalia* only have a pneumatophore (float) and lack a swimming bell (Collins 2002) causing them to permanently inhabit the surface of the ocean (Lane 1960). Also the lack of any swimming mechanisms means that *Physalia* is completely dependant on ocean winds and currents for movement. The only adaptation for movement *Physalia* possess is the float, in that there are two morphs one with a left hand sail and one with a right hand sail, allowing individuals to move at slightly different angles in the same wind condition (Barnes 1980). These characteristics mean that potentially any *Physalia* population movements can be modelled based on wind, current and swell information. For this reason *Physalia* are an ideal target species to investigate the problem of predicting the occurrence of jellyfish populations based on oceanographic data. Because detailed scientific datasets on jellyfish are virtually non-existent we used a data set that has been collected for non-scientific purposes. The dataset was sourced from Surf Lifesaving New Zealand (SLSNZ). Surf Lifesaving New Zealand is a volunteer organisation that provides surf lifeguards on beaches throughout New Zealand.

Because SLSNZ is a volunteer organisation it is reliant on community funding to operate and subsequently has developed sophisticated recording systems to document all aspects of their service to the community. The result is that there are detailed records in electronic format of every patrol that has occurred on the 72 patrolled beaches in New Zealand from the 200/2001

season. The unique aspect of this dataset is that incidents involving jellyfish stings have been recorded. Based on investigation of the data and the fact that *Physalia* is the only stinging species regularly recorded we regarded the data held by SLSNZ as a proxy presence/absence dataset for *Physalia* in New Zealand.

Clearly such data is noisy with non-linear patterns. Artificial Neural Networks (ANN), and Multi-Layer Perceptrons (MLP) in particular have shown great promise in their application to identify factors that influence biological populations, particularly in a complex environment (Lek *et al.* 1996; Olden & Jackson 2002b; Joy & Death 2004) however, their use for this purpose in ecology is still not widely accepted despite having been shown to outperform more conventional techniques (Lek *et al.* 1996; Brosse *et al.* 2001; Mutanga & Skidmore 2004). The combination of high model performance and the ability to determine variable contributions to the model makes ANN a valuable tool for understanding the underlying factors that drive the presence of *Physalia* at New Zealand beaches.

The aim of this study was to investigate the potential of an ANN model to predict the presence of *Physalia* on New Zealand beaches based on oceanographic data and to use the model to determine factors that may cause or inhibit the occurrence of *Physalia*.

3.3 Method

3.3.1 Data

As the goal of this work was to predict the presence of *Physalia* jellyfish on New Zealand beaches from oceanographic data, two data sets were sourced and combined into the final modelling data set. These sets were oceanographic data and data from Surf Lifesaving New Zealand (SLSNZ)

3.3.2 Oceanographic data

Oceanographic data was sourced from the National Institute of Water and Atmosphere (NIWA). The data contained time series outputs from NOAA/NCEP Wavewatch III model hindcast (Tolman 1998) representing 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})). MATLAB® was used to transform and manipulate the files so that they were able to be incorporated in the models. 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 file contained daily data for significant wave height (m), peak period (s),peak direction (°N), wind velocity (ms^{-1}) and wind direction(°N). For each region, data from a cell was included if the cell was less than 250km distant from the centre of the region. For this work the oceanographic data for two regions in New Zealand were extracted, West Auckland and the Bay of Plenty. The oceanographic cells associated with each of these regions are shown in Figures 3.1 and 3.2, respectively.

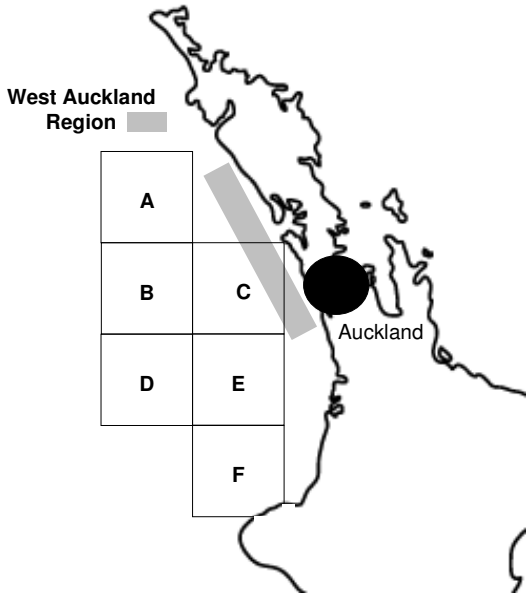


Figure 3.1: Oceanic cells associated with the West Auckland region

Figure 3.2: Oceanic cells associated with the Bay of Plenty region

3.3.3 Surf Lifesaving data

Data concerning jellyfish incidents was sourced from Surf Lifesaving New Zealand (SLSNZ). SLSNZ maintains an electronic database of all patrol records. We accessed the records of patrols carried out from the 2000/2001 season to the 2004/2005 season. The database recorded all incidents of jellyfish stings. In addition all patrol records carried out during this time period were also extracted. Records that showed a beach headcount of zero (that is, there

were no people on the beach) were excluded, as clearly there will be no jellyfish incidents if no one is swimming at the time. The use of the SLSNZ data restricted the study to dates late southern hemisphere spring to early autumn as this is the time when lifeguards patrol the beaches. It must be noted that this data is not continuous but is dependant on when patrols where carried out, which were primarily in the weekends but for between mid December and the end of January daily patrols are carried out.

3.3.4 Final Data Sets

The West Auckland data set contained 434 data points of which 100 (23%) represented the presence of *Physalia*. The West Auckland data set contained 36 variables from six ocean cells, with five continuous variables each, and six single month periods in binary format. The Bay of Plenty data set contained 411 data points of which 79 (19%) represented the presence of *Physalia*. The Bay of Plenty data set contained 51 variables from nine ocean cells of five variables each, and six single month periods. Months in both regional data sets were represented using an orthogonal binary encoding of six digits, for example January was represented by the code 001000.

3.3.5 Training and Evaluation of MLP

Standard three neuron-layer MLP were used in these experiments, and the learning algorithm used was unmodified back-propagation with momentum. The MLP used was coded by Mike Watts in C^{++}. 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* at the region on that particular day.

The method of training and evaluating the MLP (and also selecting the parameters) was similar to that suggested in Flexer (1996) and Prechelt (1996). To determine optimum

parameters a total of 64 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 1000 trials. For each trial, the training and test data set was randomly divided into a training set, consisting of two-thirds of the available data, and a test set consisting of the remaining one-third. A MLP was then created with randomly initialised connection weights and trained over the training data set. The accuracy of the MLP over the training set was then evaluated to determine how well the network had learned the training data. The accuracy of the MLP was then evaluated over the testing data set to determine how well the network generalised. Accuracy was measured as both the percentage of examples correctly classified and using Cohen's Kappa statistic (Cohen 1960). Whereas percentage accuracy is easily interpreted, it is also easily biased by unbalanced numbers of classes. That is, percentage correct may be misleadingly high when the data set in question has only a small number of examples from one class. The Kappa statistic takes the number of examples of each class into account and thus yields a less biased measure of accuracy than percentages (Cohen 1960).

For each trial the contributions of each input neuron to the output of the network was also determined, using the method of Olden and Jackson as described in Olden & Jackson (2002b). This method has been experimentally determined to give the least-biased estimate of the contribution of each input neuron (Olden *et al.* 2004) and has been used previously in ecological modelling applications (Joy & Death 2004).

At the completion of the 64 runs, the run with the highest mean kappa over the testing sets was selected as the winner for that region. The accuracy of the networks within this run was then evaluated over the validation data set. A sensitivity analysis was also performed over the significant continuous input variables of the best-generalising network within that run. That is, a sensitivity analysis was performed over each non-binary variable of the MLP with the

highest testing Kappa of the winning run. This was to illustrate the response of the network to variations in these variables so that the influence of strongly contributing inputs (as determined above) could be investigated.

3.4 Results and Discussion

3.4.1 Training Parameters

The optimal training parameters for each region, as determined by generalisation accuracy, are presented in Table 3.1. The number of hidden neurons and amount of training required for the Bay of Plenty region was substantially greater than that required for the West Auckland region. Although a general rule of thumb for determining the architecture of MLP is that the number of connections should be less than the number of training examples (data points and associated input variables). However, with the networks for the Bay of Plenty, reducing the number of hidden neurons so that this rule was observed meant that the performance of the networks was unacceptably low.

Table 3.1: Optimal training parameters by region, "Neurons" refers to the number of hidden layer neurons.

Region	Neurons	Epochs	Learning rate	Momentum
West Auckland	5	200	0.05	0.1
Bay of Plenty	15	500	0.1	0.1

3.4.2 Accuracies

The accuracies of the MLP for each region are presented in Table 3.2 as both overall percentage correct and as Cohen's Kappa statistic. It is apparent that the networks for both regions were able to generalise reasonably well. For the West Auckland region, the validation accuracies were the highest accuracies recorded for both regions. While the results for the Bay

of Plenty region would seem to indicate that overtraining has occurred, as could be expected from the size of the networks, the high validation accuracy shows that the networks were none the less still able to generalise beyond the training data.

There was a relatively large gap between the percentage accuracies and Kappa values over the test data sets. This indicates that a relatively large number of test presence examples were falsely classified as absences. A large number of false negatives could be expected to yield a high validation accuracy if the number of presences in the validation set is very low. However analysis of the validation data showed that the distribution of occurrences in the validation data set was equal to that of the training and testing set. Also, a large number of false negatives would adversely affect the Kappa statistic for the validation data set, which plainly did not happen.

Table 3.2: Mean and standard deviation of accuracies per region. "Train" is the accuracy over the training data sets, "Test" is the accuracy of the test data set and "Validate" is the accuracy over the independent validation data set.

Region		Train	Test	Validate
West Auckland	%	80.88/1.82	77.79/3.26	82.0/1.96
	κ	0.35/0.07	0.25/0.08	0.37/0.07
Bay of Plenty	%	95.13/1.77	75.15/3.99	81.89/4.02
	κ	0.83/0.06	0.19/0.09	0.45/0.10

3.4.3. Most Contributing Variables

The four variables that positively contributed the most to the networks for each region are presented in Table 3.3, and the four variables that negatively contributed the most for each region are presented in Table 3.4. It is immediately apparent from both of these tables that the contributions of the inputs for the Bay of Plenty region networks were much larger than for the West Auckland region networks. This is almost certainly because of the greater amount of

training that the Bay of Plenty region networks received through an increased number of epochs: as the method of Olden and Jackson (2002) is a decompositional, weight-based method, a larger amount of training meant that the magnitudes of the connection weights were able to grow larger than was the case with the West Auckland region networks. Therefore, the contributions were correspondingly higher.

Table 3.3: The most positively contributing variables to both regions networks

Region	Variable Name	Contribution
West Auckland	January	6.35/1.18
	December	5.67/1.27
	Wave period C	4.37/1.92
	Wave period F	2.79/1.43
Bay of Plenty	January	42.85/6.64
	December	35.58/7.91
	Wind direction G	19.40/12.83
	Wave height J	19.19/6.67

The months of January and December are significant positive variables for both regions. In other words, there was a greater probability of *Physalia* being present in these regions during these months than at other months examined. This is considered biologically plausible as December and January are both warm months (Greig *et al.* 1988). This means that there is potentially more food present for the jellyfish during these months and the increase in sea surface temperature allows for more rapid growth and reproduction. Moreover, there is a corresponding increase in people swimming at the beach increasing the possibly of a sting occurring. Wave period is also significant for the West Auckland region. An increase in wave period denotes that the waves have been generated further away (Toba *et al.* 1990) indicting that there had been sustained conditions that would transport the jellyfish into the region and hence increase their probability of occurring. A large wind direction was found to be significant for oceanic cell G in the Bay of Plenty region. If spawning grounds exist to the North of the region then wind from this direction is more likely to blow jellyfish into the Bay of Plenty area with local conditions influencing their occurrence at beaches. A larger wave

height, especially in combination with wind direction, enables the jellyfish to travel further, faster, increasing the probability of arrival in the region.

Table 3.4: Most negatively contributing variables

Region	Variable Name	Contribution
West Auckland	April	-4.79/1.18
	Wind direction F	-4.75/1.67
	Wind direction E	-3.78/2.60
	March	-3.59/1.64
Bay of Plenty	Wave period L	-34.46/15.0
	Wind direction H	-29.36/13.72
	Wave direction K	-29.10/10.05
	Wind speed G	-24.87/13.83

The months of April and March had a significant negative contribution for the West Auckland region. That is, there was a lower probability of *Physalia* being present in this region during these months than in other months. This is also considered to be biologically plausible as the temperatures during this time decrease significantly (Greig *et al.* 1988). Increases in wind direction in oceanic cells E and F also decreases the probability of *Physalia* being present. As can be seen in Figure 1, as wind direction becomes more northerly, jellyfish may be blown past the West Auckland region or this result may indicate where a *Physalia* spawning ground is located. Chapter 2 suggests that there is a possibility of a spawning ground in the Tasman Sea to the southwest of Auckland which supports the model assumption that more northerly winds decrease the probability of *Physalia* presence.

For the Bay of Plenty region, the wind direction in oceanic cells H and K makes a significant negative contribution as shown in Table 3.4. In other words, as wind direction in these cells becomes more northerly, the probability of *Physalia* presence decreases. This contradicts the interpretation of what happens in oceanic cell G but is reasonable as both oceanic cells H and K are located further away from the coast and only winds from the north-east would cause jellyfish to be pushed towards the bay. The situation with wave period for the Bay of Plenty

region is the exact opposite to the West Auckland region. This result indicates that local conditions are more important for the occurrence of *Physalia* in the Bay of Plenty region.

3.4.4 Sensitivity Analysis

Sensitivity analysis is a way to visualise how an ANN responds to the variation of a single variable. To perform a sensitivity analysis over variable n, all other input variables are set to their mean values, while the values of n are varied across the range of n, and the output of the ANN recorded. The advantage of a sensitivity analysis is that it allows for a more detailed investigation of the importance of a particular variable. Whereas an analysis of the importance of each input will yield a single overall value for the contribution of each input, a sensitivity analysis shows how the network reacts to that variable across its range. Results of the sensitivity analysis are shown in Figure 3.3 for Auckland and Figure 3.4 for the Bay of Plenty. Variables analysed from the West Auckland region showed that the networks response to variation from all variables examined was close to linear. The variables analysed from the Bay of Plenty region were nonlinear, as would be expected from the increased amount of training and subsequent greater contributions of the variables to the network. In particular wind direction from cell H strongly indicated that winds greater than 180° were not conducive to the presence of *Physalia*. Sensitivity analyses were not performed over binary variables, as this was not appropriate. Therefore, even though months such as January and December were found by contribution analysis to bevery significant for the West Auckland Region, no sensitivity analysis was performed for these variables.

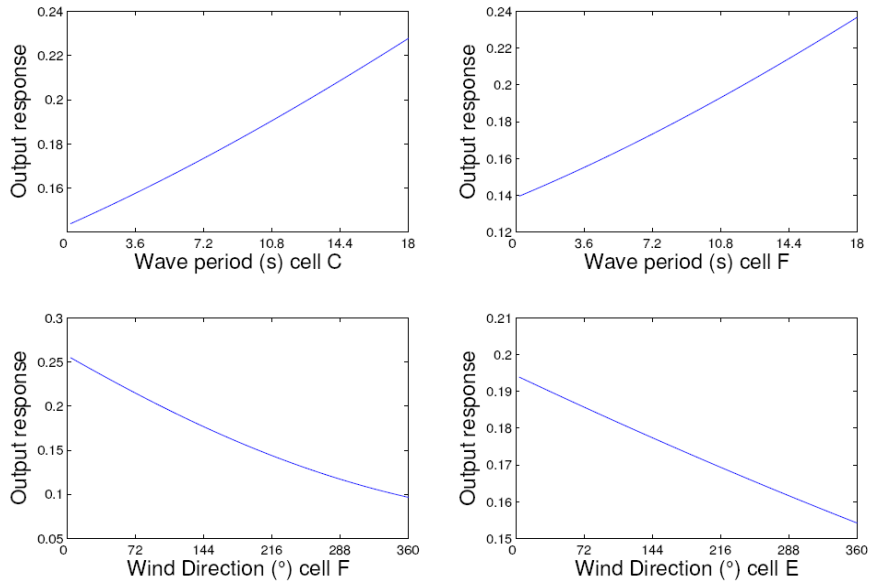


Figure 3.3: Sensitivity analysis of the most significant continuous variables for the West Auckland region

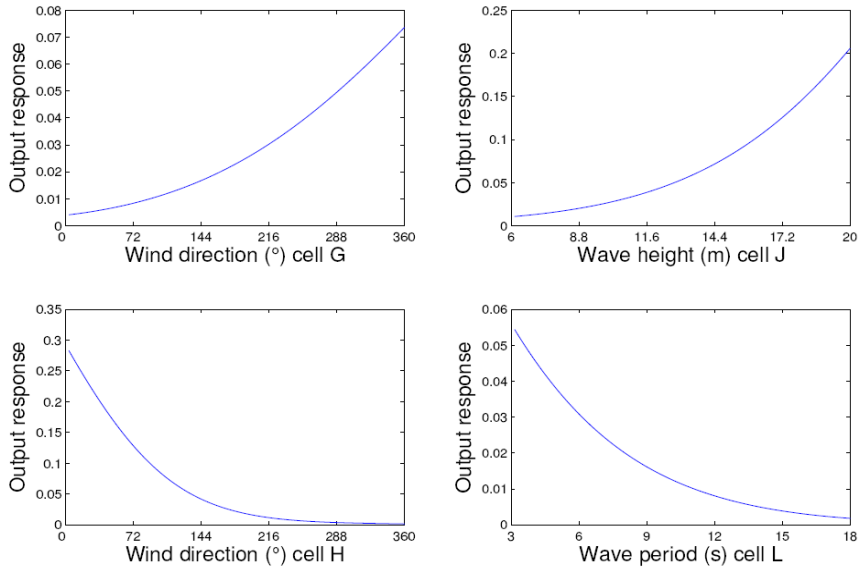


Figure 3.4: Sensitivity analysis of the most significant continuous variables for the Bay of Plenty region

3.4.5. Issues and Improvements

As is often the case with ecological data sets, the data used in this study is very noisy. This is because the presence and absence of jellyfish were inferred from reported jellyfish stings of swimmers. This results in several potential gaps in the data set: firstly, because it is clearly

possible for jellyfish to be present and not sting people; secondly, it is possible that some stings were not reported; thirdly, because beaches are not uniformly patronised over time, as there are many more swimmers during the weekend and public holidays than there are during the working week. While a multi-year survey of *Physalia* populations would be ideal for this study, the data used here was all that was available.

3.5 Conclusion

The study investigated the potential for using an MLP model to predict the presence or absence of jellyfish of the genus *Physalia* at the beaches in two regions of New Zealand. The results of input variable contribution analysis of the resulting networks are also presented. The results have shown that MLP models are capable of identifying patterns in the presence of *Physalia* in the two target regions from oceanographic data. Although the contribution analysis allows for further optimisation to generate and investigate additional hypotheses concerning *Physalia* presence and absence based on oceanographic data further exploration and refinement is necessary to draw meaningful conclusions.

As MLP are able to model *Physalia* occurrence from simplified data that lacks ecological relevance a more though exploration of data can occur. Work presented in Chapter 4 will explore methods of increasing accuracy and ecological relevance through the use of time lags within the input data. Whereas Chapter 6 will expand the study to other regions of New Zealand, and identify and clarify oceanographic features and time frames that influence *Physalia* dispersal around the New Zealand shoreline.

Chapter 4: Using time lagged input data to improve prediction of stinging jellyfish occurrence at New Zealand beaches by Multi-Layer Perceptrons

4.1 Abstract

Environmental changes in oceanic conditions have the potential to cause jellyfish populations to rapidly expand leading to ecosystem level repercussions. To predict potential changes it is necessary to understand how such populations are influenced by oceanographic conditions. Data recording the presence or absence of jellyfish of the genus *Physalia* at beaches in the West Auckland region of New Zealand were modelled using Multi-Layer Perceptrons (MLP) using time lagged oceanographic data as input data. Based on the Kappa statistic MLP models are shown to generalise well and give improved predictions of the presence or absence of *Physalia* compared to previous studies. Moreover, an analysis of the network contributions indicated that an interaction between wave and wind variables at different time intervals can promote or inhibit the occurrence of *Physalia*.

4.2 Introduction

The oceans of the world are undergoing fundamental shifts in their environments, primarily through anthropogenic influences (Arai 2001). Factors such as changing temperatures and currents may change the distribution of many pelagic marine species such as jellyfish and those changes may have ecosystem level repercussions. Jellyfish species are well known to cause problems for swimmers around beaches, block nets used in aquaculture enterprises such

as salmon farms and large numbers severely reduce fish stocks by severely depleting eggs, small larvae and plankton (Purcell *et al.* 2007). Several recent studies have noted that there is circumstantial evidence that jellyfish (Cnidaria) populations are increasing (Mills 2001; Purcell 2005) and that the changing marine environment is the main cause. To determine the true effect of a changing marine environment on pelagic species large scale population movements of such species need to be understood. Population movements of pelagic species are difficult to determine because few large scale datasets exist and consequently little work has been done in this area. One of the few studies that have attempted to model pelagic jellyfish population dynamics over a large area was carried out by Johnson *et al.* (2001) who modelled *Chrysaora quinquecirrha* in the Gulf of Mexico. Their aim was to estimate possible suitable locations for an intermediate life-stage of this species using information from the Gulf of Mexico circulation model (Johnson & Perry 1999). Their model predicted potential areas in which the intermediate life-stage could settle and complete their lifecycle validated against known areas of polyp settlement. There is a growing need for this type of information for other species of jellyfish so that bloom patterns, population dispersion and species ranges can be understood and predicted.

In an initial study Pontin *et al.* (2008) (Chapter 3) we explored the potential for an MLP to model the presence or absence of *Physalia* using data sourced from Surf Lifesaving New Zealand (SLSNZ). Despite that a simple approach was used, the study demonstrated that it there was potential for the MLP to identify patterns in the data which when refined may lead to the ability to forecast *Physalia* presence. The aim of the present study was to explore the use of a time lag in the input data to better represent the relationship between the data and *Physalia* presence at beaches. At better representation of the relationship increase the potential to predict *Physalia* presence on New Zealand beach can be assessed.

4.3 Methods

4.3.1 Oceanographic Data

Oceanographic data was sourced and processed as in Chapter 3. Time lags were created by time-stepping the data from one to six days. In other words data from one to six days prior was included into the final datasets. For this study the oceanographic data from the West Auckland region in New Zealand was extracted (Figure 4.1). The West Auckland datasets contained variables from six ocean cells, with five variables each for each day that was time-stepped, and an index for the date measured from the 1st of October for each year (Table 4.1). Furthermore, a dataset was created for window length of one in which the date index was removed to determine if the addition of a date index of this nature was appropriate.

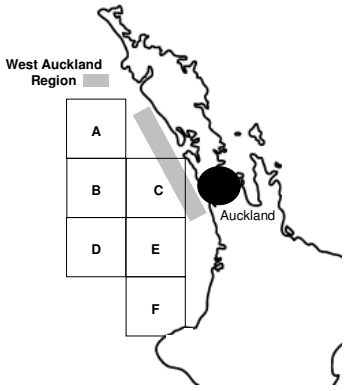


Figure 4.1: Oceanic cells associated with the West Auckland region

4.3.2 Surf Lifesaving Data.

Data recording jellyfish incidents was sourced from Surf Lifesaving New Zealand (SLSNZ). SLSNZ maintains an electronic database of all patrol records from the 2000/2001 season including all incidents of jellyfish stings. We accessed the records of patrols carried out from the 2000/2001 season to the 2004/2005 season. Data from the West Auckland region was used in this study. The West Auckland region comprises 8 patrolled beaches with an average patronage of 290 people present on each beach (SEM 18.96). Records that showed a beach headcount of zero, in other words when nobody was swimming, were excluded from the

dataset. The result was a minimum of 56 beach patrons recorded in the region at one time. An occurrence corresponded to one or more of the beaches in the region recording a jellyfish sting. It is acknowledged that there will be correlation between the numbers of swimmers present and the number of stings but using presence/absence data across the region will minimise the effect of either a large or small number of swimmers present. The use of the SLSNZ data restricted the study to dates from late southern hemisphere spring to early autumn as this is the time when lifeguards patrol the beaches. The West Auckland dataset comprised 432 data points of which 100 (23%) represented the presence of *Physalia.*

4.3.3 Training and Evaluation of MLP

Standard three neuron-layer MLP were used to model the data, and the learning algorithm used was unmodified back-propagation with momentum. The MLP used, was coded by Mike Watts in C^{++}. The method of training networks was as in Chapters 3 except each run consisted of 100 trials as apposed to 1000 as no loss of accuracy was noted with the small number of trials. MLP accuracy was measured as both the percentage of examples correctly classified and using Cohen's Kappa statistic (Cohen 1960). Whereas percentage accuracy is easily interpreted, it is also easily biased by unbalanced classes. In other words, percentage correct may be misleadingly high when the dataset in question has only a small number of examples from one class. The Kappa statistic takes the number of examples of each class into account and thus yields a less biased measure of accuracy than percentages.

For each trial the contributions of each input neuron to the output of the network was also determined, using the method of Olden and Jackson as described in Olden & Jackson (2002b). This method has been experimentally determined to give the least-biased estimate of the contribution of each input neuron (Olden *et al.* 2004) and has been used previously in ecological modelling applications (Joy & Death 2004).

At the completion of the runs, the run with the highest mean kappa over the test sets was selected as the winner for the particular lag length for that region. The accuracy of the networks within this run was then evaluated over the validation dataset. The results were then compared and the best dataset and hence time lag was selected. A sensitivity analysis was then performed over the significant input variables of the best-generalising network within that run. A Sensitivity analysis was performed on the four variables that made the highest contribution to the next as indicated by the Olden score. Sensitivity analysis was carried out to determine the response of the network to variations in the input variables so that the influence of strongly contributing inputs could be investigated.

4.4 Results and Discussion

4.4.1 Training Parameters

The optimal training parameters for each dataset, as determined by testing accuracy, are presented in Table 1. In general was would be expected, datasets with smaller time lags optimised with fewer neurons and epochs than larger time lags. An exception was a time lag of 6 days which required fewer epochs than for a lag of 5 days.

4.4.1 Accuracy

The accuracies of MLPs for each region are presented in Table 2. Overall percentage correct prediction and Cohen's Kappa statistic were calculated. The inclusion of a date index over the season significantly improved testing accuracy (κ 0.63) compared with the same data set without the date index (κ 0.22) ($p = <0.001$, T test). Furthermore the accuracies for both the train and test set for the time lagged data were better than those found by Pontin *et al.* (2008) (κ 0.35/.025), where time was represented by months using orthogonal binary encoding.

Table 4.1: The number of the most strongly contributing variables and optimal training parameters by region for each time-lagged dataset; * represents the dataset without a date index. The number of hidden layer neurons and learning paramters are also shown.

Window length	Number of variables	neurons	learning	momentum	epochs
1*	30	3	0.15	0.7	500
1	31	3	0.15	0.2	500
2	61	6	0.1	0.05	500
3	91	7	0.1	0.05	600
4	121	7	0.1	0.05	1000
5	151	7	0.1	0.05	1000
6	181	7	0.1	0.1	600

The dataset that produced the highest testing and validation Kappa for jellyfish presence had a time lag of one day and for this study contained oceanographic data that only occurred on the day of the event. This dataset was not considered for further analysis as it lacked ecological relevance because conditions that occurred in distant cells would not influence what was happening at beaches. The dataset that had the best test accuracy and had ecological relevance was the dataset that had a five day window (κ 0.40). This model however, lacked the ability to generalise (κ 0.27) compared to the dataset with a six day window (κ 0.38) which had a similar testing accuracy (κ 0.39). This possibly indicates that overtraining has occurred and further exploration of parameters surrounding the current optimum parameters may rectify this issue.

Table 4.2: Mean and standard deviation of performance criteria for the training, test and validation datasets. The performance criteria are overall percentage accuracy (%) and Cohen's Kappa statistic (κ). Note * represents the dataset without a date index.

Window length		Train	Test	Validation
1*	%	85.1/2.57	74.2/4.41	77.0/5.75
	κ	0.54/.11	0.22/.10	0.50/0.06
1	%	89.0/1.97	81.6/3.25	76.3/3.12
	κ	0.77/0.04	0.63/0.07	0.30/0.13
2	%	82.0/2.25	79.0/4.06	59.6/1.53
	κ	0.73/0.08	0.32/0.09	0.38/0.09
3	%	96.9/1.16	77.8/4.21	76.7/4.19
	κ	0.91/0.04	0.38/0.10	0.33/0.08
4	%	99.4/0.76	76.2/4.22	78.7/3.78
	κ	0.98/0.02	0.36/0.10	0.38/0.09
5	%	99.2/0.60	79.3/3.17	71.2/3.42
	κ	0.97/0.02	0.40/0.09	0.27/0.08
6	%	99.3/0.76	78.5/4.25	78.2/2.92
	κ	0.98/0.02	0.39/0.10	0.38/0.08

4.4.2 Contributing Variables

For the dataset which had a five day window, the top eight variables that contributed to network prediction are presented in Table 4.3. Wind speed in cell E both on the day of the event and the day prior to the event had significant negative contributions. That means when there is wind in the cell the probability of *Physalia* occurring decreases. The interpretation of this result is counterintuitive but is reasonable when the result of the sensitivity analysis is considered (Figure 4.2b). The sensitivity analysis indicates that the output response of the network is an inverse logarithmic relationship with light winds, between 0 ms^{-1} and 3.2ms^{-1}, decreasing the probability of *Physalia* occurring proportionally higher than winds greater than 3.2ms^{-1}. Iosilevskii & Weihs (2009) calculated that sailing speed in winds of less than 3ms^{-1} would be less than 0.2ms^{-1} or 17.3 km per day. As the West Auckland region is located adjacent to the southern limit of the region light winds would not provide enough propulsion to transport individuals on to the beach on the day as the centre of the cell is ca. 115km away from the region. Wind direction in cell C on the day prior to an event and in cell F three days prior to the appearance of jellyfish also made a significant negative contribution. In other

words, as wind direction in these cells becomes more northerly, the probability of *Physalia* presence decreases as a northerly wind would blow jellyfish past the beach towards Taranaki's coastline and outside the region.

Table 4.3: Variables identified as the most influential variables contributing to the activation of the output. The letter and number after each variable indicates the oceanographic cell (Figure 4.1) in which the variable was measured followed by the time lag.

	Variable name	Contribution
Negative variables	Wind speed E-0	-40.7/12.2
	Wind direction C-1	-37.5/14.2
	Wind direction F-3	-33.9/15.2
	Wind speed E-1	-30.5/9.80
Positive variables	Wave period A-3	42.6/13.5
	Wave height F-3	34.6/12.2
	Wave period B-3	31.6/11.0
	Wind direction D-2	30.7/14.3

Wave period in cells A and B three days prior to the event had significant positive contribution. An increase in wave period denotes that the waves have been generated further out to sea (Toba *et al.* 1990) indicating that there has been sustained conditions that are likely to transport the jellyfish into the region and increase their probability of an incident at the beach. The time lag is also highly relevant as it takes time for a swarm to travel from the outer cells to the beaches and corresponds well to reality. An increase in wave height in cell F three days prior causes an increase in the likelihood of *Physalia* occurrence. That could be through either waves being generated from wind forcing within the cell or a large swell travelling through the cell both of which would create conditions that would facilitate the transportation of *Physalia* into the region. Similarly, as wind direction tended towards the north that would cause a swarm to be blown towards the coastline. That may account for the significant positive contribution of wind direction in Cell D two days prior to an event. One important aspect the contributions indicated that wave conditions in the outer cells of the region two to three day prior to an event had a positive contribution to jellyfish incidence whereas wind

conditions on the day or the day prior to an event in the inner cells could reduce the presence of *Physalia*. These results are interesting as they seem to show the particular oceanographic conditions necessary to transport swarms into the region but that wind conditions close to the shore can dictate whether or not a swarm reaches close to the beach.

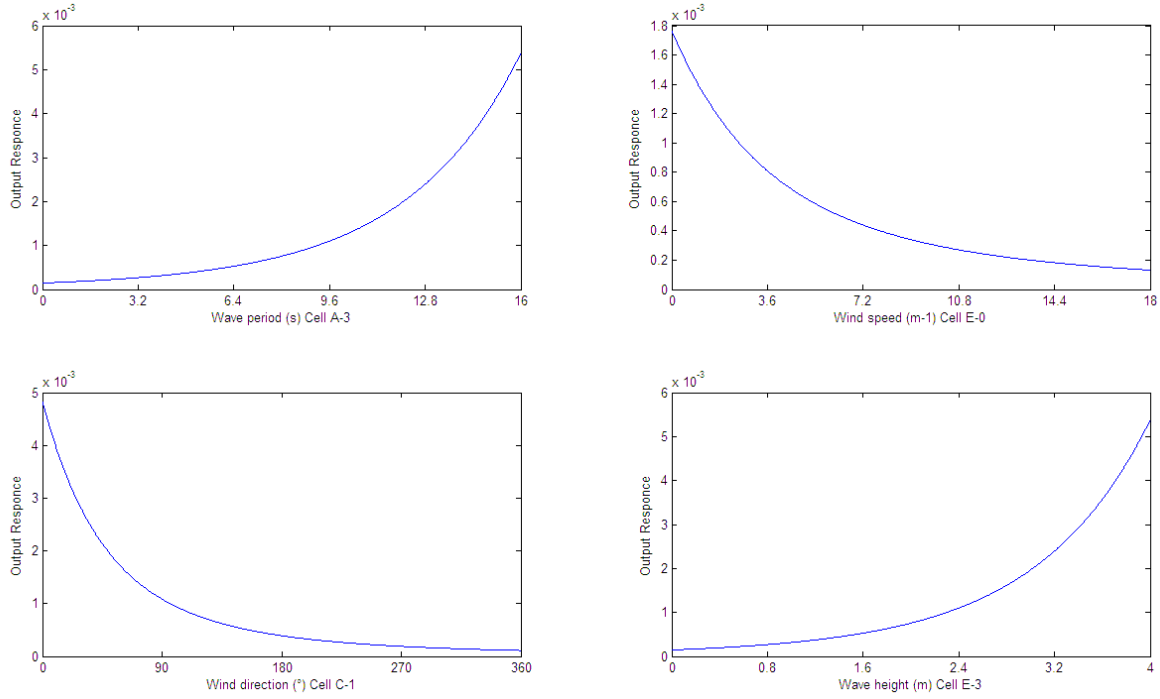


Figure 4.2**:** Sensitivity analysis of the most significant contributing variables for the West Auckland region.

4.5 Conclusion

This study clearly shows the potential for using MLP to predict the presence or absence of *Physalia* at beaches around New Zealand and that the incorporation of time lags within the data are important for increased prediction accuracy. The results showed that MLP can learn to predict to a limited degree of accuracy the presence of *Physalia* in a target region using time lagged oceanographic data. Furthermore, the contribution analysis generated valuable ecological information to assist interpretation.

Future work will expand the study to other regions of New Zealand, and will investigate methods to improve performance, such as reducing the number of input variables by removing those that are highly correlated. We will also develop a better likelihood value for the presence of *Physalia* so that the models can be assessed as a possible warning mechanism for the general public at beaches.

Chapter 5: Forecasting *Physalia* occurrence on New Zealand beaches

5.1 Abstract

Every year over 100 are people are stung by *Physalia* on New Zealand beaches. With the potential for life threatening reactions to the stings, the feasibility of forecasting *Physalia* occurrence with artificial neural networks (ANN) as a potential basis for a warning system was investigated. Previous work has indicated that modelling *Physalia* presence based on oceanographic data is imprecise because of noise and ambiguities inherent to the data. To reduce noise a variable sliding window based on a modified cascading temporal correlation analysis was used to pre-process the data from the West Auckland and Canterbury regions in an attempt to improve ANN models. Networks based on variable sliding window pre-processing outperformed the time lagged based networks giving improved forecasts in both regions. The time lagged networks gave predictions little better than chance. The models based on a variable sliding window indicated significant trends in the data but lacked the necessary resolution to accurately forecast *Physalia* occurrence for a real world application. The improved forecasts provided by the variable sliding window pre-processed models could be attributed to noise reduction by eliminating data that is not ecologically meaningful. This study further suggests the use of a modified cascading temporal correlation analysis as a noise reduction technique in ecological time series data.

5.2 Introduction

Presence only modelling is being increasingly utilised to predict and understand distributions of species, primarily based on increasingly available historical records (Graham *et al.* 2004; Elith *et al.* 2006). There are several issues surrounding the use of historical records of species distributions that includes the lack of accurate absence data and the fact that data is drawn from unstructured sampling (Zaniewski *et al.* 2002; Elith *et al.* 2006). However if such constraints are accounted for resulting models are capable of providing a relative index for occurrence. Species distribution or habitat suitability maps generated from such data have allowed ecologists to investigate a range of ecological questions with a relative degree of success (Guisan & Thuiller 2005). However, species in such distribution models tend to be sessile or persistent in the environment over a period of time and little work has been carried out with transient species or highly mobile species.

Artificial neural networks (ANN) are recognised to have the ability to make accurate forecasts from data that is noisy and nonlinear (Lek *et al.* 1996; Adya & Collopy 1998; Zhang *et al.* 1998) which has resulted in their growing use in ecological modelling (Olden *et al.* 2008). Moreover, ANN tend to outperform standard statistical techniques when applied to noisy and nonlinear data (Lek *et al.* 1996; Brosse *et al.* 1999; Mutanga & Skidmore 2004). To date ANN have been successfully used to assess fish abundance in lakes (Brosse *et al.* 1999), soybean (*Glycine wightii*) phenology (Elizondo *et al.* 1994), to predict the maturity date of spring wheat (*Triticum aestivum*) (Hill *et al.* 2002) and to predict aphid (*Rhopalosiphum padi*) abundance (Worner *et al.* 2002; Lankin-Vega *et al.* 2008). ANN have been used in species distribution modelling to good effect despite the fact the networks give a continuous result between 0 and 1, which has necessitated the use of thresholds to assess presence (Manel *et al.* 1999; Olden & Jackson 2002a; Gevrey & Worner 2006). Despite this, Marmion *et al.* (2008)

suggested model output can also be used as an index to gauge the possibility of species being present.

Each year over 100 people are recorded as being stung by a species of *Physalia* jellyfish (Surf Life Saving New Zealand unpublished data). The exact species is currently unknown because the taxonomy of this genus is uncertain (Chapter 2), however the characteristics of the sting and subsequent treatment are the same for all species in the genus (Slaughter *et al.* 2009). *Physalia* stings primarily cause localised pain but severe stings can produce nausea, vomiting, breathing difficulties and cardiovascular collapse, leading to possible death (Slaughter *et al.* 2009). Because *Physalia* is the only genus of jellyfish that is capable of stinging people in New Zealand (Slaughter *et al.* 2009) and Surf Life Saving New Zealand (SLSNZ) has recorded when people have been treated by lifeguards for jellyfish stings, such records create a proxy presence dataset for *Physalia* near swimming beaches around New Zealand.

Physalia is dispersed by ocean winds and currents by means of a pneumatophore (float) for passive movement (Collins 2002) and as a result the species in this Genus solely inhabits the surface of the ocean (Lane 1960). Because of this characteristic *Physalia* population movements have potential to be modelled based on wind, current and swell information. Previous work has shown a potential to model *Physalia* populations for the purpose of identifying environmental factors, (Pontin *et al.* 2008; Pontin *et al.* 2009) (Chapters 3, 4) leading to the possibility of forecasting likely periods of occurrence. A difficulty arises with a transient species, such as *Physalia* whose presence or absence is influenced by biotic and abotic factors that are not stable over time. For instance *Physalia* is present in New Zealand coastal waters but may or may not be present in specific localities, presenting several challenges for forecasting its presence in a noisy environment with a great deal of temporal instability.

Worner *et al.* (2002) applied a modified cascading temporal correlation analysis (Thomas *et al.* 1983) to weather data when modelling *R. padi* migrations with the aim of reducing network overtraining through the network fitting the noise. A temporal cascade takes the target output and correlates it with the mean value of the target output variable over a defined time period in a series of iterations for each observed data point. The result was a significant improvement in network accuracy. It is reasonable to expect that there will be a high level of noise in the oceanographic data, so a temporal cascade could be utilised as a noise reduction technique. This study aims to assess the potential to forecast *Physalia* occurrence in two regions of New Zealand using ANN based on a modified temporal cascading correlation termed this chapter as a variable sliding window.

5.3 Methods

5.3.1 ANN based on a variable sliding window.

Data from the combined SLSNZ and oceanographic dataset was extracted for the West Auckland and Canterbury regions for the 2000/2001 season to the 2004/2005 season (Figure 5.1) as described in Pontin *et al.* (2009), as these regions were able to be validated during the 2008/2009 season. To implement a variable sliding window it was necessary to calculate a likelihood index value of *Physalia* occurrence for each window size investigated. The likelihood index value was calculated as a moving average of *Physalia* occurrence based on the window size investigated. A variable sliding window based on cascading temporal correlation analysis (Thomas *et al.* 1983) was programmed in MATLAB 7.6.0 and applied to both regions (MathWorks 2008). Pontin *et al.* (2009) highlighted the importance of accounting for variable changes over windows of time. Therefore, correlations between abiotic variables and *Physalia* occurrence over 2 to 14 day time periods were investigated. For

each index value and variable a matrix was created consisting of the index value and the number of index values immediately before, as determined by the time period that ranged between 2-14 days, with corresponding target oceanographic variable values (Figure 5.2). For each matrix the correlation coefficient was obtained between the index values and the oceanographic variable values. The correlation coefficients for each matrix analysed were collated and highly correlated periods (< -0.7 or > 0.7) were identified for each oceanographic variable and time period. An arbitrary decision based on length of selected periods, strength of correlation and number of times selected was made to determine which time period to use for a given *Physalia* likelihood index point with the selected time periods ranging from 3 to 10 days prior for both regions. To gain the final input values for the oceanographic variables the mean of all oceanographic values over the selected time period for each *Physalia* likelihood index point was calculated and combined with the corresponding likelihood index value and date index value. At the conclusion of the pre-processing the West Auckland data set contained 111 *Physalia* likelihood index points and 31 variables and Canterbury had 75 *Physalia* likelihood index points and 41 variables.

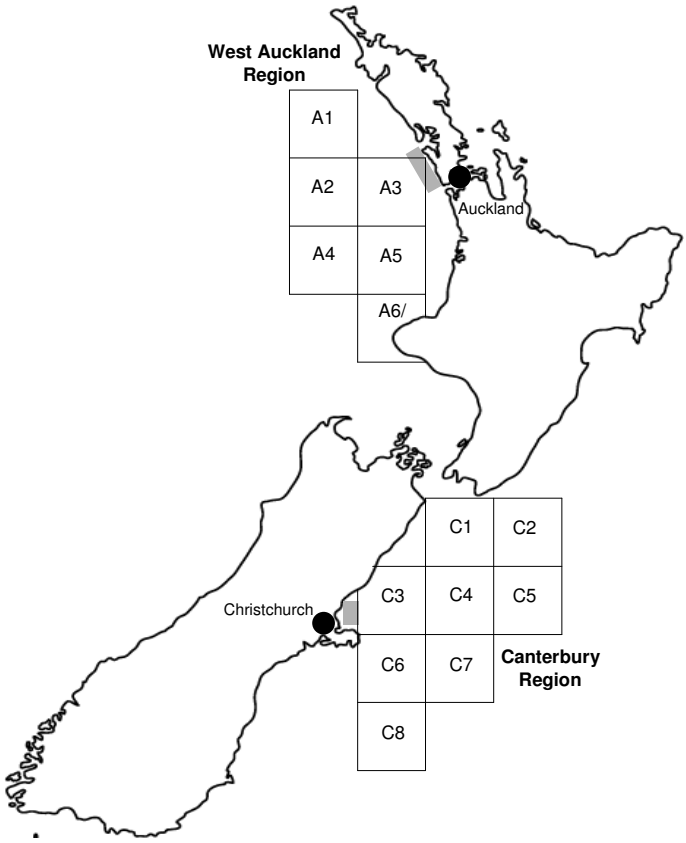


Figure 5.1: Oceanic cells associated with the West Auckland region and Canterbury regions. Regions are shown by gray shaded area.

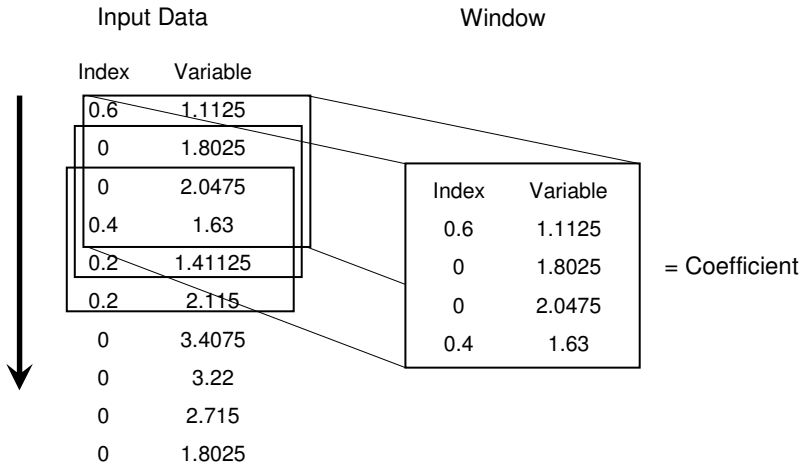


Figure 5.2: Representation of a sliding window where the average value of the variable in the window is correlated with a moving average of *Physalia* occurrence (likelihood index). Only a sliding window of 4 days only is shown. Moving windows from 2 to 14 days were also investigated.

5.3.2 ANN based on simple time lagged data

Data from the combined SLSNZ and oceanographic dataset was extracted for the West Auckland and Canterbury regions for the 2000/2001 season to the 2004/2005 season as described in Pontin *et al.* (2009). Time lags were created by time-stepping the data from one to six days as in Chapter 4 (Figure 5.3).

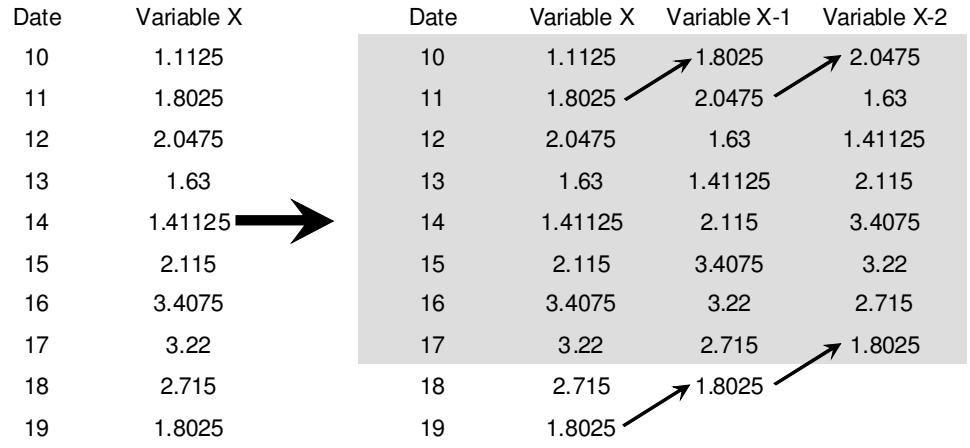


Figure 5.3: Representation of how time lags were created. A dataset incorporating a 2 day time lag for a single variable is shown with the final data in grey. Time lags from 1 to 7 days were investigated.

5.3.3 Training and Evaluation of MLP

Standard three neuron-layer MLP were used to model the data, and the learning algorithm used was unmodified back-propagation with momentum. The MLP used, was coded in C^{++}. The method of training networks is described in Chapters 3 and 4. Variable sliding window network accuracy was measured by assessing the mean, mean squared error (MSE) and mean absolute error (MAE), as suggested by Stanski *et al.* (1989) and Sheiner & Beal (1981) on both the training and test subsets. Whereas for the time lagged networks with simple presence and absences percentage accuracy, Cohen's Kappa statistic (Cohen 1960) and receiver operating characteristic (ROC) curves (Hanley & McNeil 1982) were calculated as performance criteria. The 100 networks trained from the best network parameter combination were selected to be validated for each data type.

For the time lagged data the proportion of variables to examples was high especially with the larger time lags, to reduce the risk of overtraining training of the networks was carried out in two steps. The first step was to train the networks as above, utilising all of the data and then using the Olden scores (Olden & Jackson 2002b), to identify the percentage that each input variable contributed to the network. The second step was to reduce the number of variables by selecting variables 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 re-found as above for each different total percent contribution. The accuracy of these networks was assessed as described above.

5.3.4 Validation

The validation dataset was comprised of additional *Physalia* occurrence data sourced from SLSNZ during the period from mid December 2008 to the start of February 2009 season. As the data was not used in training the networks it was used as independent data to evaluate the networks ability to generalise. This time period was chosen as lifeguards are on the beaches daily at a time of high public usage. Forecast data for oceanographic variables was extracted from the NOAA/NCEP Wavewatch III model (Tolman 1998). The Wavewatch III model is run daily, generating a seven day forecast for significant wave height (m), peak wave period (s), peak wave direction (ºN) and U and V wind variables which are wind direction and intensity as vector components (ms^{-1}). Twice weekly, data was extracted for this study so that the maximum forecast range was four days. Because of data collection error, forecasts were not extracted for the 21/12/08 to 3/1/09 period. Forecast data was processed for both the variable sliding window and presence based networks for each region as described above. *Physalia* likelihood index points varied between regions (West Auckland 51 and Canterbury 47) because the period of continuous patrols varied between regions. Whereas, for the time lagged validation data the West Auckland region had 42 *Physalia* occurrence points and

Canterbury 36 respectively. The same threshold for presence (0.5) was applied to the time lagged networks as in Chapter 3 and 4 (Pontin *et al.* 2009) to determine classification of the network output.

5.4 Results

5.4.1 Forecasts based on time lagged data

The forecasts generated from the time lagged networks in general gave predictions accuracies within the bounds of pervious studies with the Canterbury region having lower accuracies (κ=-0.17 AUC=0.483) than the Auckland region (κ=0.38, AUC=0.616). The forecasts when individual networks were compared tended to be highly variable with large ranges in the predicted value as shown by the density plot in Figure 5.4. ROC plots further highlight the lack of precision showing predictions little better than chance.

5.4.2 Networks based on variable sliding windows

The optimal training parameters of the variable sliding window networks for each region, as determined by test set accuracy, are presented in Table 5.1 along with the corresponding MSE and MAE for the training and test subsets. Variable contribution to each of the regions is shown in Table 5.2. Wind direction in cells A5, A3 and A6 all had a negative contribution to the model for the West Auckland region indicating that as the wind bearing increased tending towards the Northwest, the likelihood index of occurrence decreased. Whereas, wave period in cells A6, A1 and A3 had a positive influence on the likelihood index, indicating a longer time between waves promoted occurrence. The Canterbury region wind speed in cells C4, C5, C7 and C8 had a negative influence on the networks, where increased wind speed decreased the index of occurrence. Wind direction in cells C7 and C6, wave height in C1 and wave period in C2 all had a positive influence on the networks.

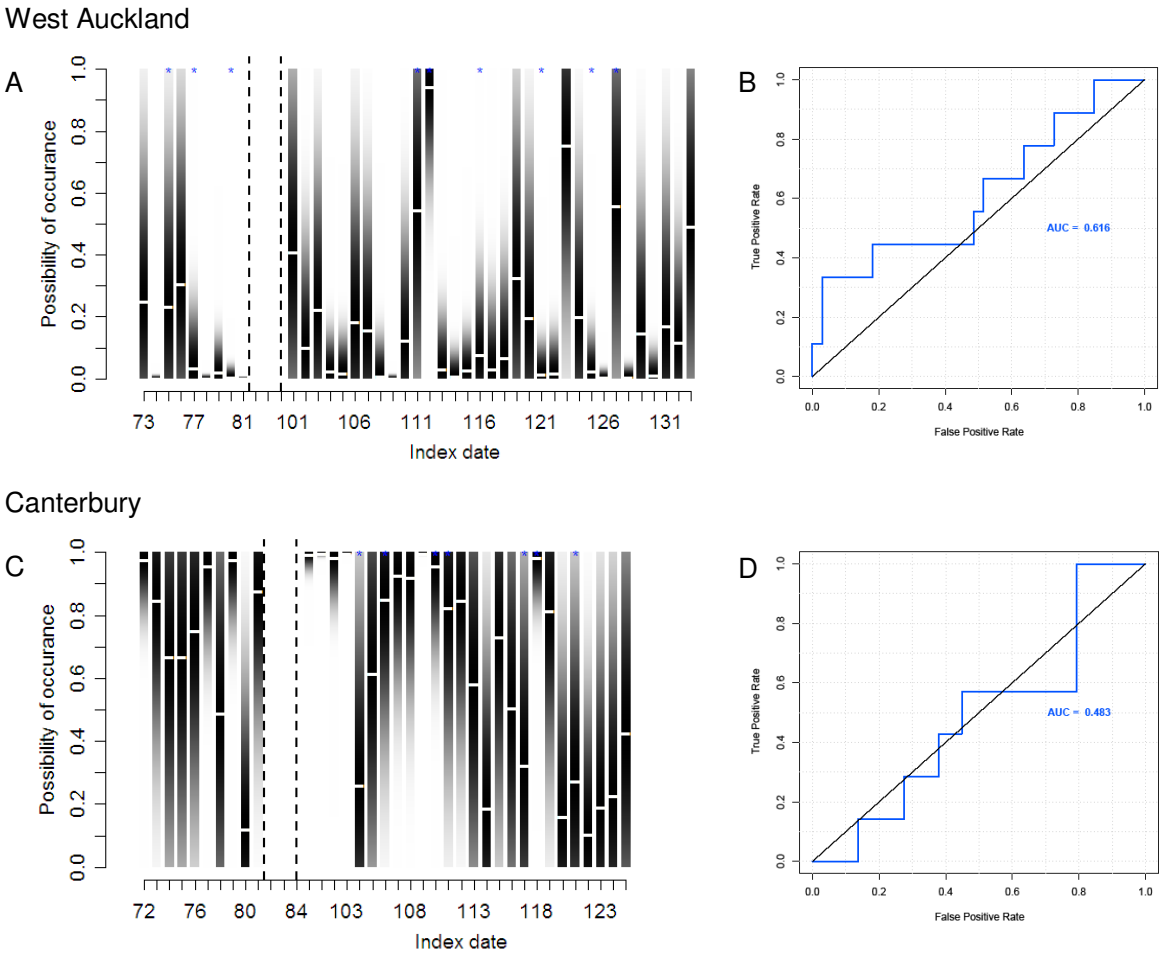


Figure 5.4: Density plot (A and C) and corresponding ROC curve (B and D) of forecast *Physalia* occurrence in both the West Auckland and Canterbury region during the 08/09 season from time lagged networks. Each bar is a graphical representation of the distribution of forecasts generated from 100 neural networks. The mean forecast is represented by the horizontal white bar. The asterisks represent actual occurrences of *Physalia* with missing data indicated by the gap between the dotted lines. Index date is from 1st October 2008.

Table 5.1: Optimal training parameters for each region for variable sliding window networks. "Neurons" is the number of hidden layer neurons. The MSE and MAE values are the mean of a 100 trained networks.

Region	Neurons	Learning	Momentum	Epochs	Train MSE	Train MAE	Test MSE	Test MAE
West Auckland	6	0.1	0.05	450	0.032	0.148	0.028	0.139
Canterbury	7	0.05	0.05	250	0.174	0.59	0.240	0.22

Table 5.2: Variables identified as contributing most to the activation of the output in the West Auckland and Canterbury regions for the variable sliding window networks. The cell letter and number indicate the oceanographic cell (Figure 1) in which the variable was measured.

	Negative Variables		Positive Variables	
Region	Variable	Contribution	Variable	Contribution
West Auckland	Cell A5 wind direction	-8.41	Cell A6 wave period	3.84
	Cell A3 wind direction	-6.45	Cell A1 wave period	3.66
	Cell A6 wind direction	-4.12	Cell A5 wave direction	3.24
	Cell A1 wave height	-3.29	Cell A3 wave period	3.14
Region	Variable	Contribution	Variable	Contribution
Canterbury	Cell C8 wind speed	-7.89	Cell C7 wind direction	4.23
	Cell C5 wind speed	-5.75	Cell C1 wave height	4.10
	Cell C7 wind speed	-5.39	Cell C6 wind direction	3.74
	Cell C4 wind speed	-3.36	Cell C2 wave period	3.73

5.4.3 Variable sliding window based forecasts

Forecasted likelihood index of occurrence for *Physalia* in both the West Auckland and Canterbury regions are shown in Figures 5.5 and 5.6. The range of forecasts generated by the 100 networks for each day varies greatly from day to day. In general, closer to the extremes there was less variability between the forecasts compared with periods where the forecasts swing from one extreme to the other. Variability between forecasts was more pronounced in the West Auckland region. By assessing thresholds that are optimised to maximise correct forecasts and minimise false positives for each region (Table 5.2) it is possible to assess model performance. Networks from the Canterbury region correctly classified 85% of *Physalia* occurrences compared to 67% for West Auckland although both had high false positive rates (42% and 42% respectively). The variable sliding window networks

significantly outperformed the presence based networks as they could only forecast correct occurrence 40% (Canterbury, false positive rate 92%) and 33% (Auckland, false positive rate 25%).

Table 5.3: Percentage of correct presence forecasts and false positives and negatives for *Physalia* occurrence for variable sliding widow networks in both the West Auckland and Canterbury region during the 08/09 season under different thresholds. A threshold was considered reached if the 95% confidence of the mean forecast (n=100) encapsulated the threshold. Numbers of actual *Physalia* occurrences are shown in brackets for the correct presence forecasts. Optimum thresholds are in **bold**.

West Auckland region			
Threshold	Correct presence forecast (%) (12)	False absence (%)	False presence (%)
0.3	**66.67**	**33.33**	**42.50**
0.35	58.33	41.67	37.50
0.4	50.00	50.00	32.50
0.45	50.00	50.00	30.00
0.5	41.67	58.33	27.50
0.55	41.67	58.33	22.50
0.6	8.33	91.67	20.00
Canterbury region			
Threshold	Correct presence forecast (%) (7)	False absence (%)	False presence (%)
0.3	85.71	14.29	60.00
0.35	85.71	14.29	55.00
0.4	**85.71**	**14.29**	**42.50**
0.45	71.43	28.57	42.50
0.5	71.43	28.57	42.50
0.55	57.14	42.86	42.50
0.6	57.14	42.86	40.00

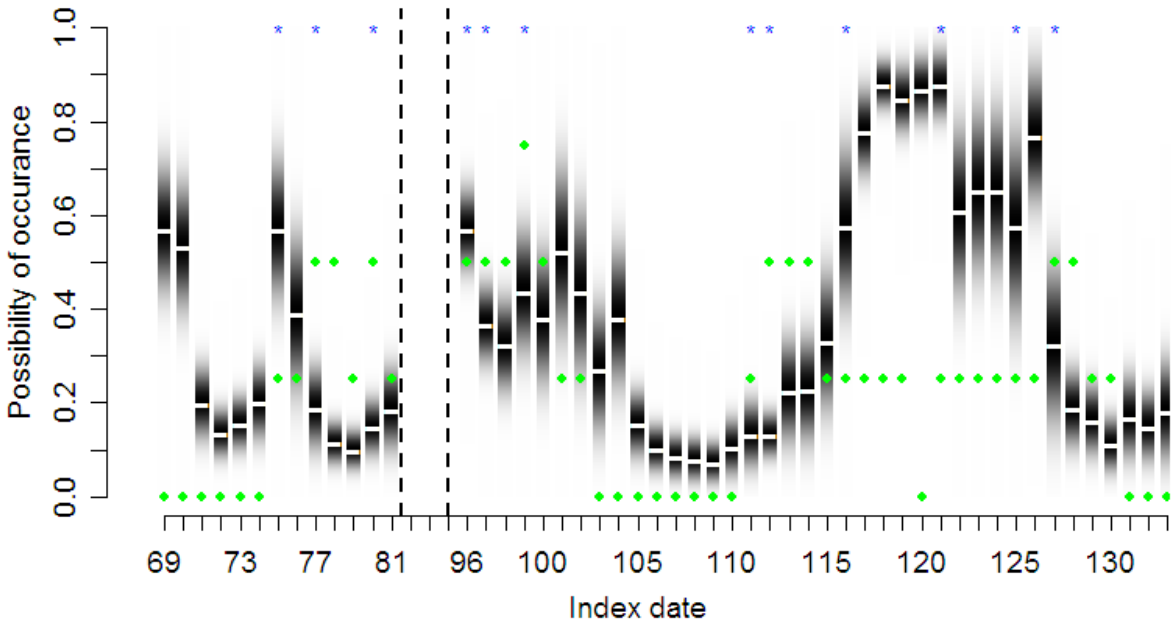


Figure 5.5: Density plot of forecast of the likelihood index of *Physalia* occurrence in the West Auckland region during the 08/09 season. Each bar is a graphical representation of the distribution of forecasts generated from 100 neural networks. The mean forecast is represented by the horizontal white bar. The blue asterisks represent observed occurrences of *Physalia* whereas the small green diamond represents the running average of occurrence. Missing data is indicated by the gap between the dotted lines. Index date is from 1st October 2008.

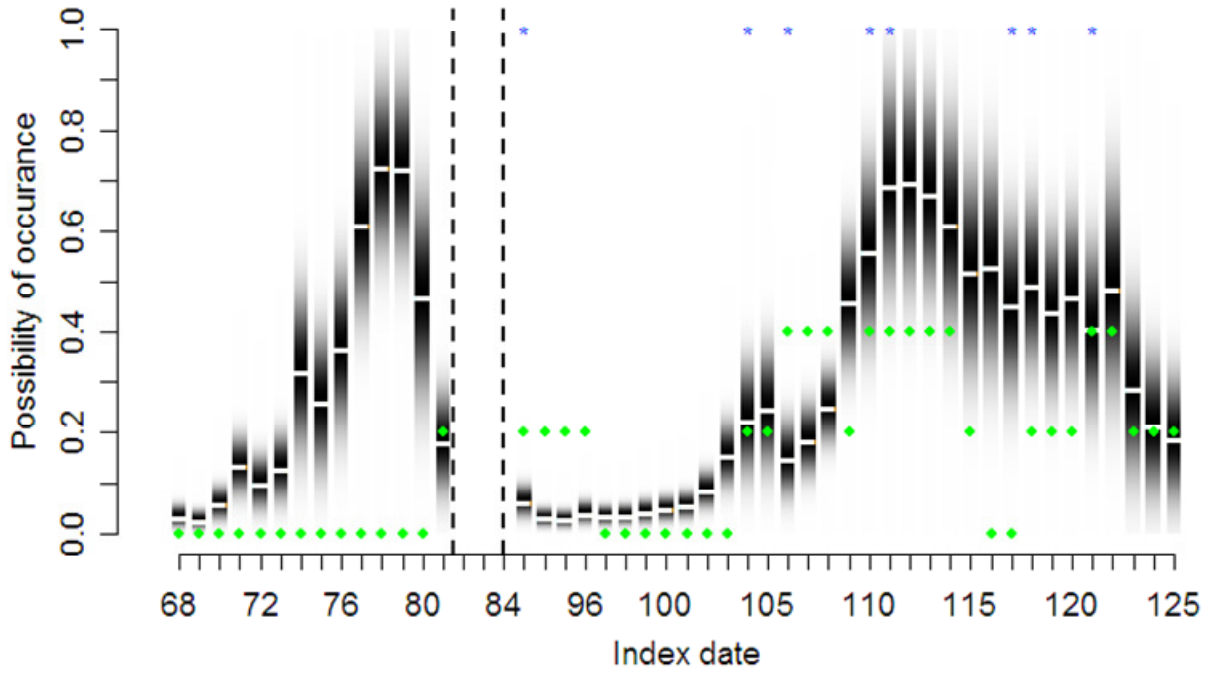


Figure 5.6: Density plot of forecast of the likelihood index of occurrence for *Physalia* in the Canterbury region during the 08/09 season. Each bar is a graphical representation of the distribution of forecasts generated from 100 neural networks. The mean forecast is represented by the horizontal white bar. The blue asterisks represent actual occurrences of *Physalia* whereas the small green diamond represents the running average of occurrence. Missing data is indicated by the gap between the dotted lines. Index date is from 1st October 2008.

5.5 Discussion

The use of a variable sliding window to pre-process the environmental data to forecast *Physalia* occurrence by ANN in two New Zealand regions was reasonably successful, especially when compared to forecasts generated from the time lagged networks. Clearly a high number of false positives were forecasted however; in this application false positives are of minor consequence and false negatives more important. When forecasting *Physalia* occurrence the forecast must target the requirements of the end user otherwise the effectiveness of the forecast become limited (Johnston *et al.* 2005). It is envisioned that that a patrol captain, in charge of a given beach, is the target end user. To a patrol captain, false

positives are not so concerning as false negatives because they are trained to prevent incidents and will use anything to minimise potential incidents. Given the patrol captains mindset it maybe a false positive is received more favourably than a false negative for which they are unprepared. However, it is key that the forecasts are simple to understand and that the patrol captains have confidence in the likelihood index of occurrence and associated thresholds (Johnston *et al.* 2005). Many authors stress the importance of minimising false predictions when creating models (Gotelli 2000; Hwang *et al.* 2005; Liu *et al.* 2005) but this requires that false predictions can be clearly identified. Given the sources of error it is impossible to determine whether inaccurate forecasts are the result of model error, few people swimming on that day, unreported stings, or that jellyfish were present but by chance nobody happened to be stung. While minimisation of false forecasts is desirable, this type of prediction is similar to weather forecasting where a level of error is accepted.

To maximise forecasts accuracy every effort is required to minimise aspects of input data and modelling technique that reduce accuracy while at the same time, maximising model representation of realistic features within the target system (Maier & Dandy 2000). For an accurate forecast of the likelihood index of occurrence, the model must be designed and trained with data that represents the desired target (Clark 2007). The time lagged networks have two inadequacies with respect to their input data that primarily accounts for their poor forecast performance. By using time lagged data in binary format, network training was optimised to give a binary output, through the use of thresholds. Given the issues with correctly assessing an actual *Physalia* occurrence, using time lagged data may create additional noise in the data and mask any patterns occurring. Whereas by using a running average to represent the occurrence of *Physalia* it allows the variable sliding window networks to differentiate between high and low periods of occurrence increasing the possibility of the networks identifying patterns in the input data (Dawson & Wilby 1998). The

other issue associated with the input data for the time lagged networks was the way a time lag was incorporated into the input data. Time lags were created by time-stepping the data (figure 5.3) which has the potential to greatly increase the number of explanatory variables, especially if the number of days is high, increasing the possibility of overtraining or the network learning the noise in the data (Nath *et al.* 1997; Maier & Dandy 2000). In comparison, the temp variable sliding window networks were applied to a moving average of the variables which minimised the number of training variables but also smoothed the data increasing the possibility of the networks identifying patterns in the input data.

Although a presence based approach has been utilised for many different ecological applications such as predicting fish presence (Mastrorillo *et al.* 1997), species distributions (Elith *et al.* 2006) and predicting bird habitat suitability (Brotons *et al.* 2004), it, like all other models, relies on the quality of the input data and explanatory variables. The input data for this study is extremely noisy reducing the adequacy of model fit and therefore the accuracy of any subsequent forecast. The use of a variable sliding window to pre-process the data improved forecast accuracy. By identifying and utilising time periods where the input variables were highly correlated with *Physalia* occurrence, more relevant data were identified, and noise was reduced such that the networks more readily learned important patterns in the data.

Variables identified using a variable sliding window indicated the oceanographic variables that are likely to influence *Physalia* presence. In the West Auckland region, winds from the Northwest lower the likelihood index of occurrence indicating that *Physalia* individuals are transported into the region from the south. The variables identified in the Canterbury region also indicate that *Physalia* seems to originate somewhere to the north of the region as Northwest winds are important in *Physalia* presence. This possibly indicates that either

Physalia are being transported down the length of New Zealand or they are passing through Cook Strait. Further modelling is required particularly of the Wellington region to ascertain if either scenario is a possibility.

The use of a variable sliding window seems to have enhanced the network's ability to identify patterns as shown by the strong patterns forecasted by the networks. Furthermore, using a variable sliding window to analyse the input data necessitated the use of a running average occurrence as the target output probably gave a more realistic target especially when the temporal variability of records of *Physalia* occurrence is considered. The use of a variable sliding window has been demonstrated by Worner *et al.* (2002) the dataset used in that study was less complex and probably was less noisy. The fact that the approach seems successful in this study indicates its potential for predicting complex highly variable environmental data.

Compared with weather forecasting the volume of data used to train and validate the models is quite small for both regions. Simply incorporating the 2008/2009 season data into the training data will increase the training examples available to the networks giving a greater representation of the problem and leading to better networks (Lek *et al.* 2000). Moreover it is possible to maximise the resolution of the input data by utilising the daily output from the Wavewatch III model further increasing the resolution of the forecasts. The need to develop an informative way of communicating the forecast to the end user is also a priority with a web based system preferred because of ease of access. The use of thresholds to forecast *Physalia* occurrence derived from the likelihood index of occurrence is possible, however, false positive rates are high which may reduce confidence in the forecasts over time (Johnston *et al.* 2005). But including the forecasted likelihood index of occurrence will allow any end user to assess the likelihood index of occurrence. That information will increase the relevance of the

information until such time more and better data reduce false positive rates to acceptable levels.

In summary, the ANN network ability to forecast *Physalia* occurrence in the West Auckland and Canterbury regions was enhanced through the use of a variable sliding window to pre-process the input data. Although the use of a variable sliding window eliminated much of the noise in the data the forecasts were limited to general trends. Despite this result there is potential to increase the resolution of the forecast so that a real time forecasting system can be implemented.

Chapter 6: Predicting the occurrence of *Physalia* at New Zealand beaches using Multi-Layer Perceptrons and Naïve Bayes Classifiers

6.1 Abstract

Observing or monitoring pelagic species population levels is very difficult. Currently there has been limited development of a theoretical framework that assists understanding how a pelagic population may aggregate or be redistributed through oceanographic processes. This study seeks to use artificial neural networks (ANN) and a Versatile Quantum-inspired Evolutionary Algorithm (VQEA) mapped to feature space with a Naïve Bayes Classifiers (NBC) to model the presence and absence of *Physalia* in New Zealand coastal waters as a model system to explore possible effects of oceanographic processes on a pelagic population. ANN accuracies achieved improved on previous studies of the *Physalia* dataset but were outperformed by the NBC. Both models achieved classification accuracies so that was possible to identify significant oceanographic influences with confidence of capturing a true representation of the system. The models indicated that New Zealand appears to have two independent systems driven by currents and oceanographic features that are responsible for the redistribution of *Physalia* from North of New Zealand and the Tasman Sea and their subsequent presence in coastal waters.

6.2 Introduction

Climate change is predicted to cause changes that will have significant as well as varied range of impacts on many species (Fischlin *et al.* 2007). There is a general expectation that the phylum Cnidaria will increase its populations, primarily through the expected increase in ocean temperatures and higher rates of eutrophication (Mills, 2001; Purcell and Arai, 2001; Parsons and Lalli, 2002). Increases in jellyfish populations have been known to block nets used in aquaculture enterprises, such as salmon farms, and high populations can affect fish stocks by severely depleting eggs, small larvae and plankton (Purcell *et al.* 2007). Additionally, jellyfish cause a great deal of discomfort or potentially death to swimmers (Bailey *et al.* 2003). Clearly it would be useful to predict when such species become troublesome, however there are several issues. First, only a limited number of datasets of appropriate population data exist (Purcell 2005), restricting the ability to establish base population levels for species in specific geographic regions. Two, Cnidarian populations naturally undergo large fluctuations, which may mask or prevent the identification of changes to their population levels (Lynam *et al.* 2004; Purcell 2005).

Because of a lack of data, researchers may be forced to use datasets collected for purposes other than research. While such data potentially lacks the rigour associated with that collected for scientific purposes such data may provide a viable alternative for study if potential shortcomings in the data are recognised and treated accordingly (Elith *et al.* 2006). Identifying population fluctuations is also difficult, especially the accurate identification of population maxima and minima in the bloom cycle so that trends can be correctly identified. Establishing when a bloom has occurred is difficult because as Graham *et al.* (2001) describes there are two types of blooms; true and apparent. True blooms are when a rapid change in a jellyfish population has occurred and apparent blooms are when a stable population has been re-

dispersed. Distinguishing between the bloom types may be impossible without detailed records of the surrounding areas. To aid in distinguishing bloom types an understanding of how the marine environment influences jellyfish movement is required.

To date, the number of studies that have attempted to determine to what extent the marine environment influences jellyfish populations either to infer population movement or actual recorded effects have been limited. Johnson *et al.* (2001) used large scale circulation models to model pelagic *Chrysaora quinquecirrha* population dynamics in the Gulf of Mexico with the aim of determining probable polyp areas. Johnson *et al*. (2001) wanted to identify possible suitable locations for an intermediate life-stage of this species using information from the Gulf of Mexico circulation model. Key variables and processes that were identified through modelling this system, were that wind forcing played a dominate role in the distribution patterns encountered and that currents were also highly relevant. The extent to which these variables are important to population movement of jellyfish species is questionable as the species that inhabit the depths would not be affected by wind forcing whereas others are capable of significant movement by themselves. For example, *Chironex fleckeri* has been recorded as having a straight line speed of 212m an hour (Seymour *et al.* 2004).

The genus is *Physalia* is considered to be one of the more primitive extant Hydrozoan genera as it lacks many of the morphological characteristics associated with later evolving species (Collins 2002; Dunn *et al.* 2005). In particular *Physalia* lacks any swimming mechanisms causing *Physalia* to be completely dependant on ocean winds and currents for movement and dispersal (Lane 1960). Two morphs have evolved with regard to their sail, one with a left hand sail and one with a right hand sail, allowing individuals to move at slightly different angles in the same wind condition (Totton 1960; Barnes 1980). This apparent reliance on wind and currents potentially means that *Physalia* population movement can be modelled

using wind, current and swell information. For this reason *Physalia* is an ideal model species to investigate and identify factors that influence pelagic cnidarian populations based on oceanographic data.

Artificial neural networks are potentially a very powerful technique for modelling species populations and identifying key factors that influence their populations, especially for Cnidarians that live in a complex and changing environment. Because oceans are highly variable a large number of features, that potentially could contain significant noise in their data, need to be considered. Standard statistical techniques have been shown to be often out-performed by ANN when applied to complex data (Lek *et al.* 1996; Brosse *et al.* 1999; Mutanga & Skidmore 2004). In particular Multi-Layer Perceptrons (MLP) have shown great promise in their application to problems with noisy and complex data (Lek *et al.* 1996; Olden & Jackson 2002b; Joy & Death 2004) however, their use in ecology is still not widely accepted.

There is a growing use of Quantum-Inspired Evolutionary Algorithms (QEA) to replace the use of Evolutionary Algorithms, as they have been shown to been to superior (Han & Kim 2004; Venayagamoorthy *et al.* 2005). A current generation QEA is the Versatile Quantum-inspired Evolutionary Algorithm (vQEA) proposed by Platel *et al.* (2007) to overcome irreversible choice and the phenomenon of hitchhiking that can occur in QEA. vQEA improves QEA by updating the attractors without considering their fitness and performing a global synchronisation so that generation t+1 corresponds to the best solution found at generation t. vQEA was found to significantly outperform both Classical Genetic Algorithm and QEA when compared using benchmark classification problems. Because it is possible to translate each generation from the algorithm to feature space it is possible to assess what features are being identified through the use of robust classifiers such as Naïve Bayes

Classifiers (NBC) (Friedman *et al.* 1997) giving a powerful technique for feature identification (Kotsiantis 2007). Moreover, NBC, which are normally less accurate than ANN (Kotsiantis *et al.* 2006), have been shown to be comparable and sometimes superior for instance-based learning when used of with state of the art algorithms (Domingos & Pazzani 1997; Kotsiantis 2007).

Initial attempts to model *Physalia* presence or absence for selected regions around the New Zealand shoreline indicated ANN were able to identify patterns within the data (Pontin *et al.* 2008). With relevant time lags added to the models, ecological realism as well as accuracy improved (Pontin *et al.* 2009) which indicated that identification of relevant features was possible. The aim of the present study was to use and compare ANN and the Versatile Quantum-inspired Evolutionary Algorithm in conjunction with an NBC to identify and clarify oceanographic features and time frames that influence *Physalia* dispersal around the New Zealand shoreline.

6.3 Methods

6.3.1 Data.

Data from the combined SLSNZ and oceanographic dataset was extracted for five regions in New Zealand (West Auckland, Bay of Plenty, Taranaki, Wellington and Canterbury) for the 2000/2001 season to the 2004/2005 season (Figure 6.1) as described in Pontin *et al.* (2009). Time lags were created by time-stepping the data from one to six days (Figure 6.2). In other words data from one to six days prior to the target event was concatenated and included as the final dataset. Time lags of up to six days were investigated because time-stepping rapidly increases the number of features in the dataset to the point that feature numbers would exceed examples, leading to overtraining.

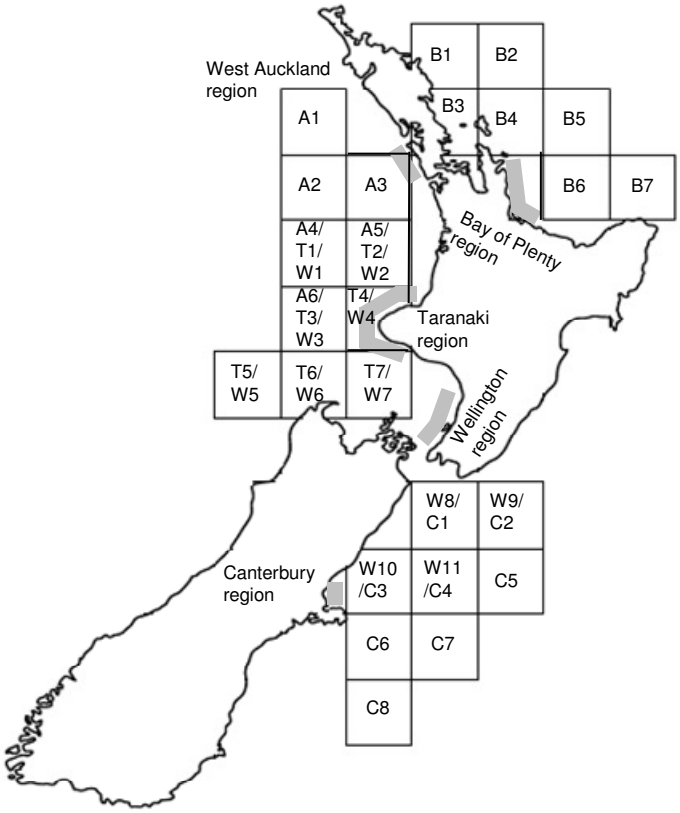


Figure 6.1: Oceanic cells associated with each of the five regions examined. Cells that are associated with a particular region are shown by ID codes in which the letter indicates the associated region, except for the West Auckland region which is represented by an A, and the number identifies individual cells within a region.

6.3.2 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 with momentum. The method of training networks was as in Chapters 3, 4 and 5. 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 the risk of overtraining training of the networks was carried out in two steps. The first step was to train the networks as above utilising all of the data and then using the Olden scores (Olden & Jackson 2002b), indentify the percentage that each input feature contributed to the network. 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 re found as above for each different total percent contribution. The accuracy of these networks was assessed as described above.

6.3.3 Naïve Bayesian Classifier

The Versatile Quantum-inspired Evolutionary Algorithm (vQEA) (Platel *et al.* 2007) was used as the wrapping optimization algorithm. 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 (Platel *et al.* 2008). The learning rate was set to $\theta = \pi/100$ and the algorithm was allowed to evolve over a total number 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 containing a much higher number of features. 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.

In every generation the chromosome of each individual in the population was translated into the corresponding feature space. Naïve Bayesian Classifier (NBC) was then trained and tested using k-fold cross-validation procedure. Parameter k was set for each dataset individually and are summarised in Table 6.1. 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 to 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

had a high correlation with the initial feature then this 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 & Dandy 2000; Guyon & Elisseeff 2003).

6.4 Results

6.4.1 MLP accuracies

The best performing network model based on a combination of time lag and network parameters for each region is shown in Table 6.2. The optimum time lag over the regions ranged between 3 and 6 days. Training kappa were between 0.97 and 0.99 which when compared to the test kappa of between 0.26 and 0.40 was an indication that overtraining has occurred because of the large difference between the training and test results.

Networks that were trained with a reduced number of features based on the percentage 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 (Table 6.2). Validation accuracy was also increased by utilising a reduced number of features with gains of between 0.01 to 0.39 recorded (Table 6.2). The percentage contribution that determined the number of features selected was 50% in all regions except for Canterbury which was 40% (Figure 6.1). In other words the network performed best when features selected from the largest contributing features represented 40% to 50% of the total contribution based on olden scores (Olden & Jackson 2002b).

6.4.2 Naive Bayesian Classifier accuracies

Because of the strong imbalance of the datasets, the percent of correctly classified samples started at a high level >80%, for all regions, at the beginning of the evolutionary run. Despite the high initial accuracy improvements to classification accuracy were still possible in later generations (Figure 6.3). The average final testing accuracy reported by each individual in the population was significantly higher than the corresponding MLPs testing accuracies for all regions ($p<0.03$, t test). Because of the reduced number of generations that the Wellington region was evolved over, although gaining comparable classification accuracy with other regions, increased generations may improve results further. As the vQEA was evolved, the accuracy steadily increased with a corresponding decrease in the number of features (Figure 6.3).

6.4.3 Feature selection

Features that had a large contribution to the MLP networks for each region are shown in Table 6.3. 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 6.1) was higher in all regions than the corresponding features selected through percent contribution in the MLP (Table 6.2). 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 are identifying the same underlying pattern with an average of 4.8 (SEM 0.86) features being the same or highly correlated (Table 6.2).

Table 6.1: Performance, parameters and number of features used to train Naïve Bayesian Classifier (NBC) associated with each region;* indicates significant increase ($p<0.05$) compared to the best testing accuracy achieved by the MLP (Table 6.4), and ** a highly significant increase ($p<0.001$) (T test).

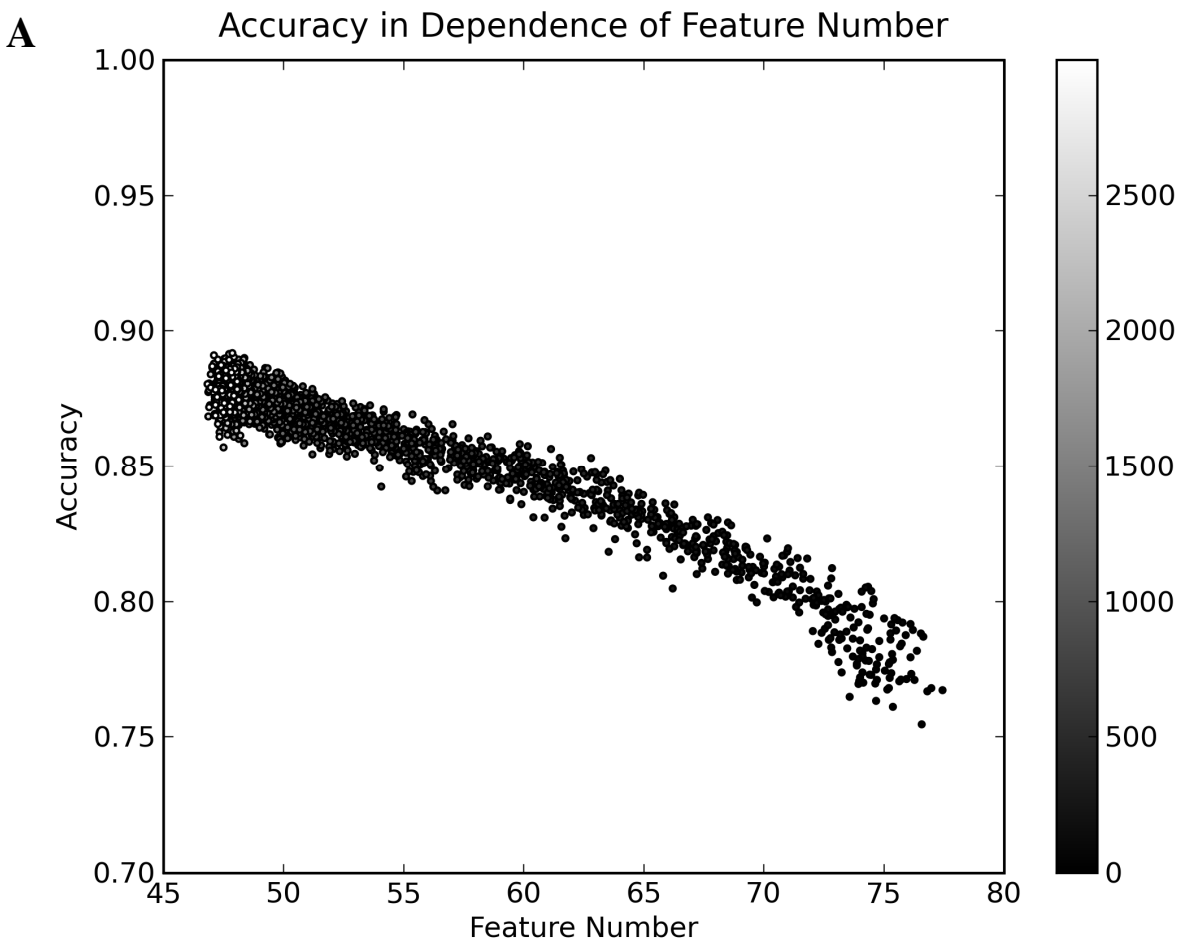
Region	Lag	Generations	Parameter *k*	Final number of features	Overall Kappa	Test Kappa
West Auckland	4	3000	12	47	0.7912	0.6276**
Bay of Plenty	3	3000	10	42	0.8159	0.6347**
Taranaki	6	3000	7	94	0.9675	0.7034*
Wellington	6	1000	7	159	0.949	0.6961**
Canterbury	6	3000	9	87	0.8844	0.7082**

Table 6.2: 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” is the number of hidden layer neurons.

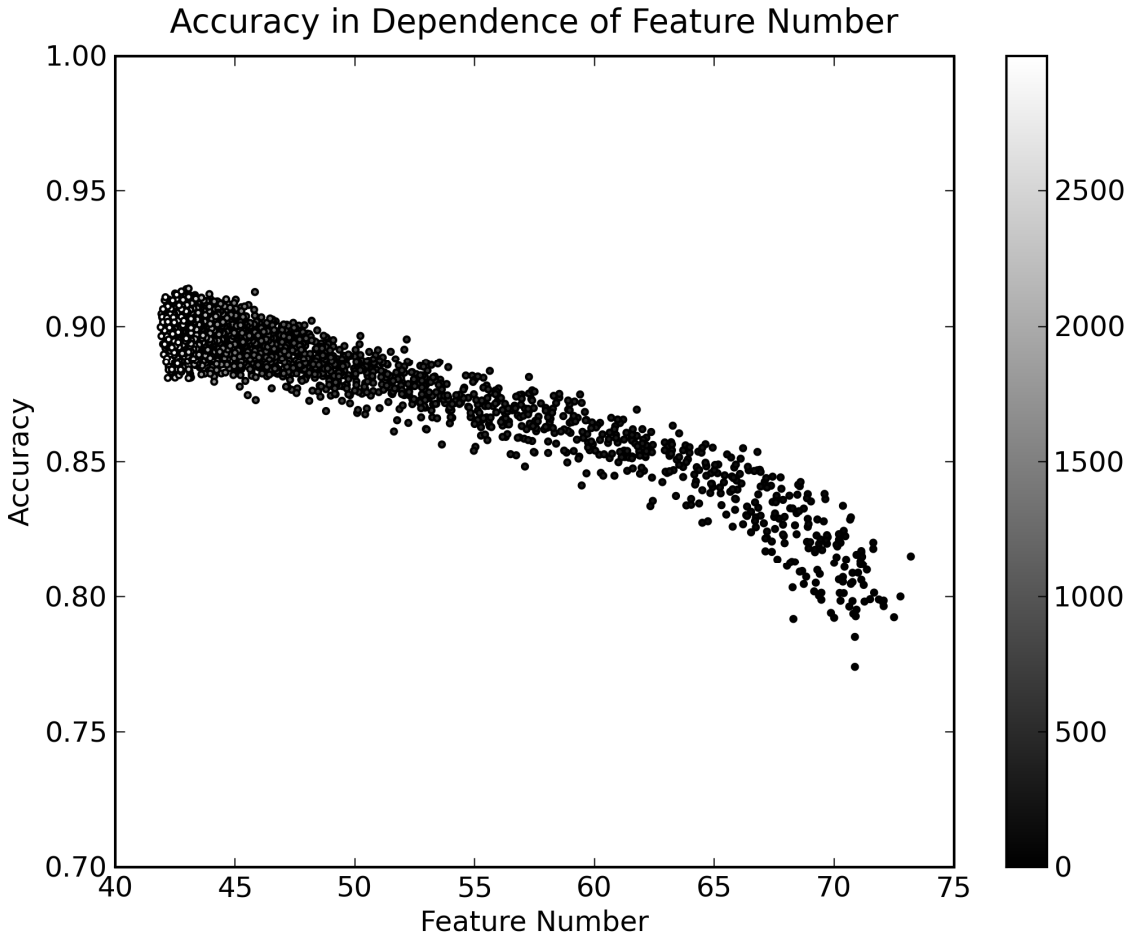
Region	Lag	Contribution	Neurons	Learning	Momentum	Epochs	Training	Test	Validation
West Auckland	4	100% (151)	7	0.1	0.05	1000	0.9749	0.4016	0.2718
Bay of Plenty	3	100% (142)	5	0.1	0.05	1000	0.9672	0.2602	0.1057
Taranaki	6	100% (246)	6	0.1	0.1	800	0.9755	0.3749	0.1917
Wellington	3	100% (386)	6	0.1	0.1	700	0.9725	0.4662	0.1216
Canterbury	6	100% (281)	3	0.1	0.1	800	0.9893	0.3941	0.2897
West Auckland	4	50% (33)	8	0.2	0.1	900	0.9004	0.5191	0.2889
Bay of Plenty	3	50% (31)	9	0.2	0.1	800	0.8824	0.4023	0.0184
Taranaki	6	50% (54)	8	0.2	0.3	950	0.9691	0.6479	0.2942
Wellington	6	50% (83)	4	0.6	0.6	400	0.9947	0.5807	0.5125
Canterbury	6	40% (45)	9	0.6	0.3	800	0.9656	0.6279	0.3021

Table 6.3: Features that were identified as the most influential contributing to the activation of the output for each of the five regions in the MLP. The letter and number after each feature indicates the oceanographic cell (Figure 6.1) in which the feature was measured and how many days prior to the data point the data were taken;* indicates that the feature, or another highly correlated feature, was selected by the NBC.

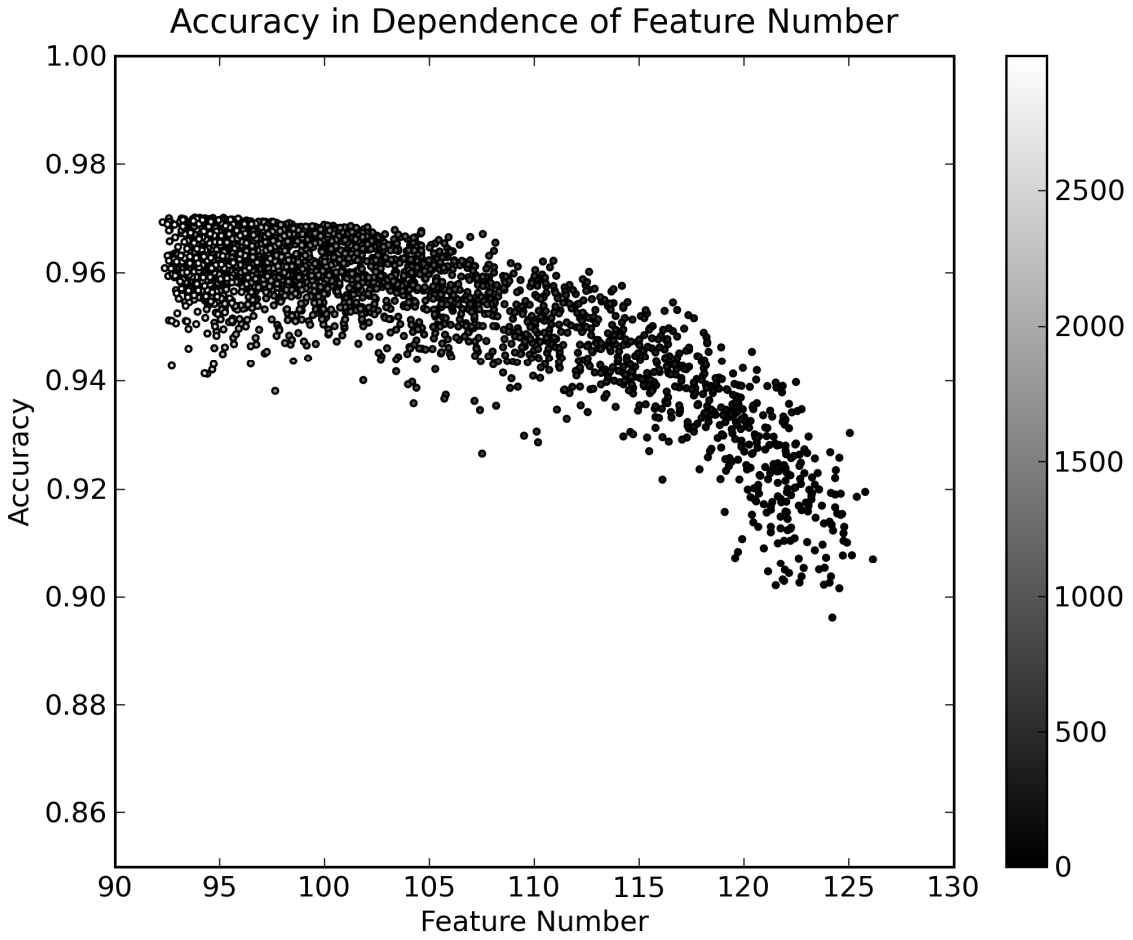
	Positive Features				Negative Features			
Region	Rank	Cell-lag	Feature	Contribution	Rank	Cell-lag	Feature	Contribution
West Auckland	3	cell 6-0	Wave direction*	58.31	1	cell 3-1	Wave period*	-84.85
	8	cell 5-0	Wave direction	43.67	2	cell 5-0	Wind speed	-84.76
	9	cell 1-3	Wave period	43.44	4	cell 1-4	Wind speed*	-57.53
	10	cell 4-2	Wind direction	41.84	5	cell 5-1	Wind speed*	-55.91
Region	Rank	Cell-lag	Feature	Contribution	Rank	Cell-lag	Feature	Contribution
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
Region	Rank	Cell-lag	Feature	Contribution	Rank	Cell-lag	Feature	Contribution
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
Region	Rank	Cell-lag	Feature	Contribution	Rank	Cell-lag	Feature	Contribution
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
Region	Rank	Cell-lag	Feature	Contribution	Rank	Cell-lag	Feature	Contribution
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


A
Accuracy in Dependence of Feature Number
Accuracy
Feature Number
1.00
0.95
0.90
0.85
0.80
0.75
0.70
45
50
55
60
65
70
75
80
2500
2000
1500
1000
500
0

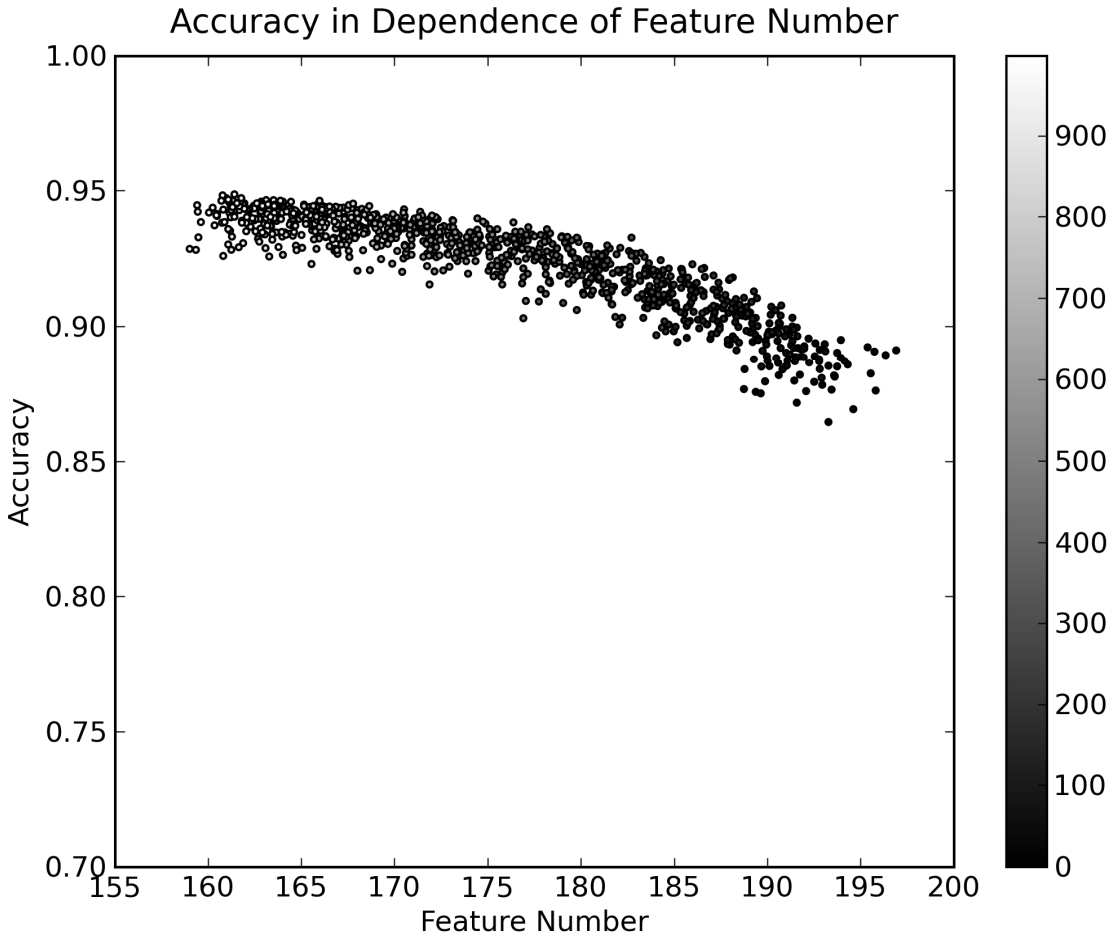
B


Accuracy in Dependence of Feature Number
Accuracy
Feature Number
1.00
0.95
0.90
0.85
0.80
0.75
0.70
40
45
50
55
60
65
70
75
2500
2000
1500
1000
500
0

C


Accuracy in Dependence of Feature Number
Accuracy
Feature Number
1.00
0.98
0.96
0.94
0.92
0.90
0.88
0.86
90
95
100
105
110
115
120
125
130
2500
2000
1500
1000
500
0

D


Accuracy in Dependence of Feature Number
Accuracy
Feature Number
1.00
0.95
0.90
0.85
0.80
0.75
0.70
155
160
165
170
175
180
185
190
195
200
900
800
700
600
500
400
300
200
100
0

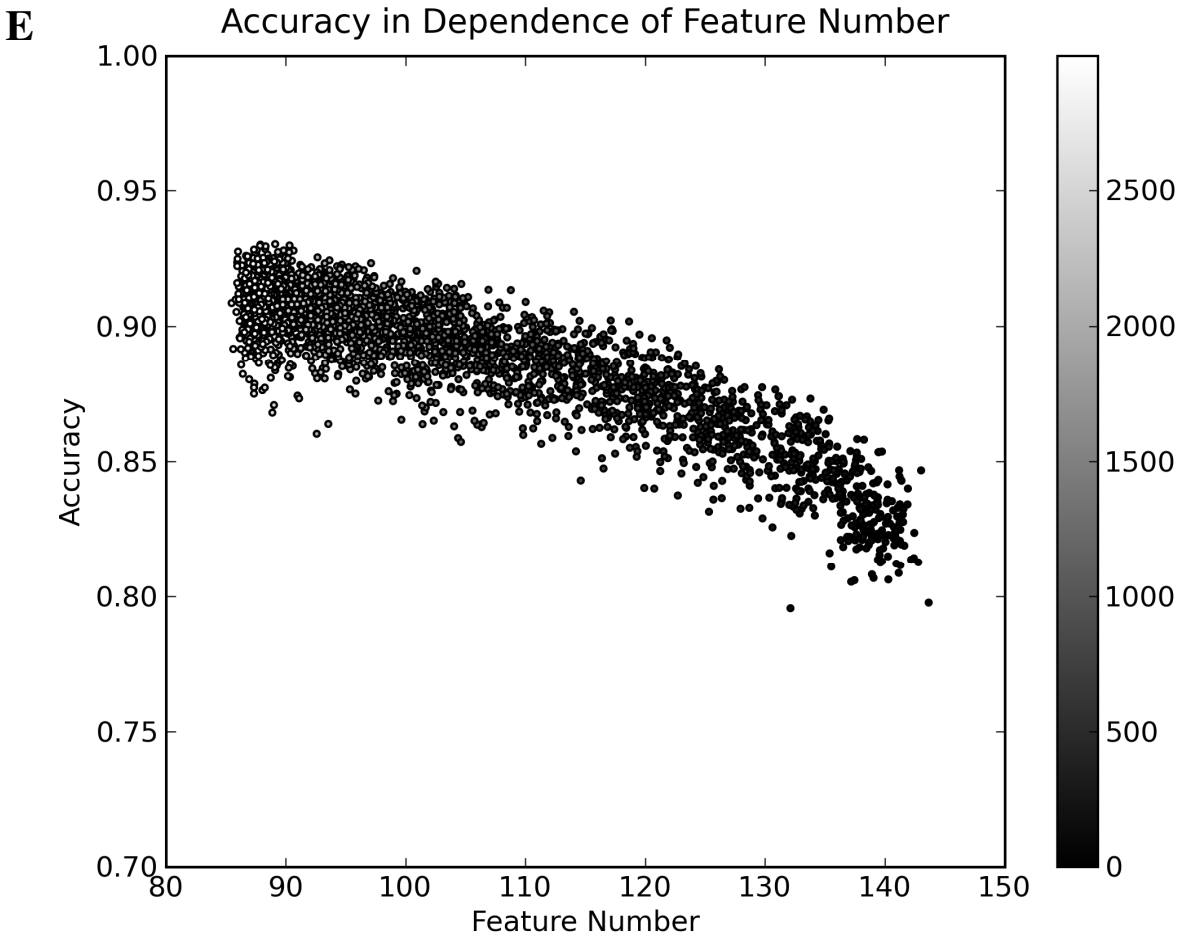


Figure 6.2: Evolution of NBC for classifying *Physalia* presence in five New Zealand regions (A: West Auckland, B: Bay of Plenty, C: Taranaki, D: Wellington and E: Canterbury) in relation to the number of features incorporated in the model and classification accuracy (percentage correctly classified). The different gray 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.

6.5 Discussion

The reduction of features included in the MLP networks, through the use of a contribution analysis, increased the network accuracy and ability to generalise over a wider area than achieved previously (Chapters 3,4 and 5) (Pontin *et al.* 2008; Pontin *et al.* 2009). The primary explanation for the improvements in accuracy is the reduction in noise within the input data and the improvement to the proportion of features to examples (Nath *et al.* 1997; Maier & Dandy 2000). As the less relevant features were eliminated the proportion of features to examples decreased subsequently reducing the noise and increasing performance. A similar process was achieved with the NBC with performance increasing as features were discarded. Although the NBC outperformed the MLP in classifying *Physalia* presence both model types identified similar noteworthy features to each other. Because these features were identified independently with similar importance they can be considered to have a greater importance and therefore more relevance to the system (Bowden *et al.* 2005; Muttil & Chau 2007). By using the variables identified by both ANN and NBC as an ensemble is possible to accurately identify variables with high contribution to the system. Moreover, a greater precision is often gained with ensembles (Araujo & New 2007; Lankin-Vega *et al.* 2008).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 how the network responds to a feature, either positively or negatively, which is not possible with a NBC.

When features with high contribution to a region's MLP networks are examined (Table 6.4) it is clear that each region has distinct patterns that can be extrapolated out to indicate possible explanations for the movement and origin of the blooms. Moreover, the patterns in all regions except Bay of Plenty appear interlinked to a degree. A similar conclusion can be drawn from the features selected by the NBC for the Bay of Plenty as it appears that oceanographic

conditions to the north of the region play an important role in determining *Physalia* occurrence. Whereas, from the features identified in the other regions, there appears to be similar linkages between regions. 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, Taranaki, Wellington and Canterbury regions (Figure 6.4).

It is unlikely that there are localised populations of *Phyaslia* inhabiting the New Zealand coastline. A recent survey of New Zealand *Physalia* species using molecular techniques (Chapter 2) indicated that only the Bay of Plenty had a genetically distinct population. As the remainder of New Zealand appears homogeneous then it is probable individuals are being sourced from the same area. The Bay of Plenty system suggests a blooming area between the North Cape eddie and East Cape eddie. There are periodic blooms of phytoplankton in that area during November (Murphy *et al.* 2001) which would provide a food source for a *Physalia* bloom. North-westerly wave and wind directions in the cells to the Northwest increase the possibility of *Physalia* occurrence (Table 6.4). But the models are indicating that similar wave directions to the north of the region one day prior decrease the possibility of occurrence. To interpret this conflicting pattern the regions current system must be taken into account. The East Auckland Current flows down the coast from North Cape and there are two permanent eddy systems in the area (Stanton *et al.* 1997). It is suspected that this current system has a greater influence on the speed and direction at which a bloom is transported. As wind and wave conditions either enhance or suppress a current strength (Stanton *et al.* 1997) it is hypothesised that if the current is enhanced to such a degree it is capable of either moving a bloom pasted the region or dispersing the bloom so that the possibility of a swimmer being stung is decreased, reflecting an issue in the source data Because the model is unable to

factor the currents and or the possibility of a number of false negatives it may account for the networks poor ability to generalise.

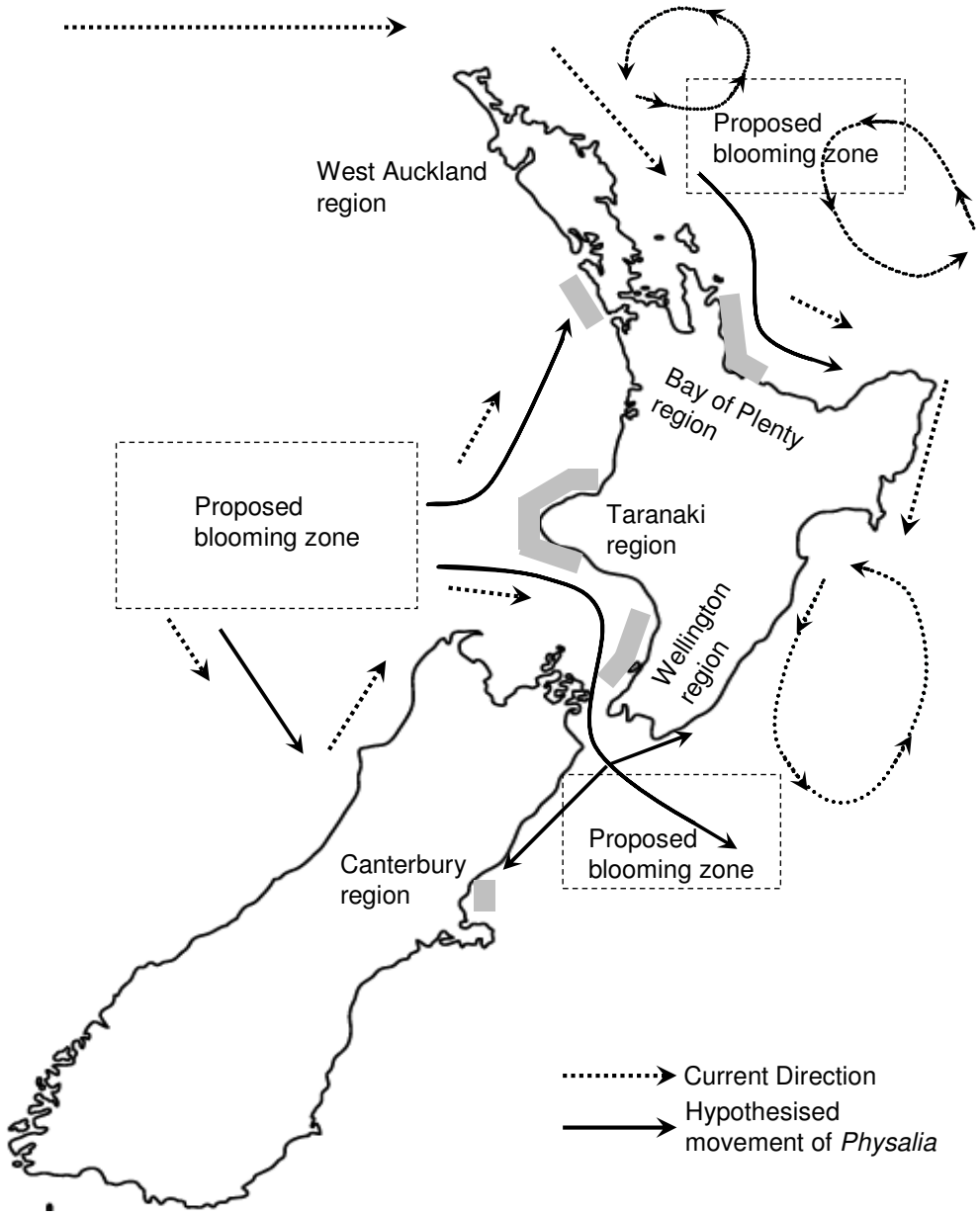


Figure 6.3: Hypothesised representation of *Physalia* movement and blooming zones around New Zealand as indicated from the ANN model.

For the other regions it is suggested that there is a blooming zone located west of the Taranaki region in the Tasman Sea (Figure 6.4). Again, each year in November there is a large increase in phytoplankton in this area (Murphy *et al.* 2001) indicating that environmental conditions

are suitable for a *Physalia* bloom. The SLSNZ data correlates with this observation as the earliest incidence records occur in late October in the West Auckland region and early November in both the Canterbury and Wellington region. The Taranaki region does not show any incidents until early December as beach patrols in the region do not start until that time. Once the bloom has formed it can be transported in multiple directions. For the West Auckland region a more westerly wind and swell in A4, A5 and A6 increases the possibility of an occurrence. These wind and wave directions also affect the Wellington region in a similar manner. It is expected that individuals will pass through Cook Strait as it is the only passage between the two main islands in over 2000km. Moreover the prevailing winds and currents will disperse any blooms formed in the blooming zone area toward the strait. Once an individual or bloom had passed through the strait it may possibly be followed by an additional bloom as phytoplankton is known to increase off the coast of Canterbury in January (Murphy *et al.* 2001). Canterbury often experiences more incidences later in the season than the other regions giving some support to this suggestion. The role that currents play in this particular system again cannot be quantified but the D'Urville Current is a significant current that passes through Cook Strait (Heath 1986) and would have a large affect on the transportation of *Physalia* in the region. However, the D'Urville current is strongly effected by wind forcing (Heath 1986) and as *Physalia* are adapted for wind dispersal the two features could be heavily confounded. North Westerly winds 3 to 6 days prior in cells C2 and C7 increase the possibility of occurrence in the Canterbury region further supporting the suggestion that *Physalia* pass through Cook Strait.

The blooming zones and dispersal patterns proposed here not only provide a hypothetical framework for further experimental and observational studies of pelagic species movement and distribution but also provide significant interpretation of other results. The geographic dispersion based on a molecular analysis of New Zealand's *Physalia* (Chapter 2) gives a

similar pattern with a similar interpretation but without any suggestion of a mechanism. By combining the two techniques it is possible to provide an independent validation of the models and provide a theoretical framework to interpretation of the molecular data. With the added effect of the identification of important geographical areas that can then be more intensively sampled to provide increased resolutions of crucial areas.

To implement management strategies to mitigate unwanted effects of Cnidarian blooms a detailed understanding of their populations is required. The models presented do not provide this but they do provide an initial point from additional which information can be incorporated and model development continued. If development can be continued with subsequent gains in accuracy then it is highly conceivable that a real time predictive model is achievable. The next step from there would be taking the techniques developed and applying them to other species or genera at the same or larger scales to assess if patterns found are generalised or species specific.

Chapter 7: General Discussion

The use of molecular techniques along with Artificial Neural Networks (ANN) provided additional interpretations about *Physalia* distributions in New Zealand waters than would not have been possible if either approach had been used in isolation. Both investigative approaches provide evidence for the existence of two separate oceanographic systems that appear to drive *Physalia* distribution around New Zealand. These systems may have contributed to possible speciation within the genus. One system is located in the Bay of Plenty extending south along the east coast of the North Island to Riversdale. The other system is thought to encompass the west coast of both main islands, Cook Strait and from Riversdale south along the South Island's East coast (Figure 6.4). By identifying key environmental variables using both ANN and NBC models it was possible to extrapolate *Physalia* movement and identify potential blooming areas, notoriously hard to establish for jellyfish. Such areas may be indicative for further study and/or validation. Despite the synergy between approaches important issues with respect to New Zealand's *Physalia* species have been identified that will require further research to clarify.

7.1 Molecular techniques

Molecular analysis of *Physalia* specimens from New Zealand, Australia and Hawaii generated a surprisingly complex taxonomic picture of the *Physalia* genus with three clans being identified. There is the potential for at least two new species to be identified from the species complex through closer morphological analysis. At present only *P. physalis* is recognised (Totton 1960; Bouillon *et al.* 2006) and given the genetic diversity found in a relatively small sample size, it is clear the entire genus will require a comprehensive global review using a fully integrated taxonomic approach that incorporates morphological and molecular analysis

(Dayrat 2005). A fully integrated approach will be vital in any review to ensure that morphological or molecular variation can be assessed accurately. From the work completed in this study it seems highly probable that a review will result in the re-description of *P. utriculus* and that the species name will be assigned to one of the clans identified.

While the molecular results do not fully clarify which species of *Physalia* are present in New Zealand they do indicate that *P. physalis* is likely to be a rare species as no specimens were sampled. A particular criticism of this conclusion is that it is solely based on molecular data but if *P. physalis* is as common as Wesrerskov & Probert (1981) and Slaughter *et al.* (2009) suggest, then *P. physalis* should have been detected even with the limited sample size in this study. Furthermore, given the associated results with a supposedly common New Zealand species such as *P. physalis* it might appropriate to question the true status of the other 133 hydrozoans (Bouillon & Barnett 1999) and 761 cnidarians (Smith & Gordon 2003) thought to be present in New Zealand. Haddock (2004) claims that cnidarians have been neglected in marine research for the past 100 years compared with other marine taxa, so it is likely that other New Zealand cnidarian genera will require a taxonomic review if their molecular information is examined.

7.2 Modelling

The development of ANN to model *Physalia* presence undertaken from the initial proof of concept to the final networks, incorporated increased ecological realism with increased modelling accuracy at each step. The initial approach used in Chapter 3 did not account for prior conditions but it did show that ANN is capable of modelling *Physalia* presence using the SLSNZ and Wavewatch III data. With species that are passive drifters, such as *Physalia*, and indeed many biological species, it is the cumulative effect of prior conditions that contribute

to the distribution observed (Ellien *et al.* 2000; Barnay *et al.* 2003; Ellien *et al.* 2004). The subsequent incorporation of a time lag that accounted for prior conditions in Chapter 4 significantly enhanced the ecological realism of the model because it allowed some form of assessment of the effect of prior conditions to be made. Moreover, the incorporation of a time lag also allowed an initial assessment of significant variables that contribute to *Physalia* presence. The time lag however, greatly increased the number of input variables on which the networks were trained, increasing the risk of overtraining.

Compounding the issue of overtraining there was a considerable imbalance in the data with absences comprising between 78 to 89% of the datasets. Imbalances of this nature have been shown to decrease model accuracy as the model will seek accurate performance over the full range of instances (Chen *et al.* 2008; Liu *et al.* 2008). The result is a model that is able to classify the majority class accurately (in this case absences) but has a poor ability to classify the minority class (Xu & Chow 2006; Chen *et al.* 2008). Because of the class imbalance the use of presence only data became questionable for incorporation into a forecast model. To provide effective forecasts a likelihood index of *Physalia* occurrence was devised for forecasting the presence only data. The likelihood index of *Physalia* occurrence mitigated the issues created by the class imbalance as the target output was transformed to a continuous index rather than a discrete occurrence. When combined with a variable sliding window as a pre-processing technique for incorporating a meaningful time component in the input data, the subsequent networks outperformed comparable networks based on time-lagged data (Chapter 5). Although, the accuracy and resolution of the variable sliding window networks were still not sufficient for detailed forecasts, useful forecasts of general trends could be made. Those variables that had a large contribution to the variable sliding window networks indicated the oceanographic variables that appear to influence *Physalia* presence but because the model and

input data were optimised for forecasting, detailed ecological conclusions about the effect of these variables on *Physalia* populations could not be made.

The main objective of Chapter 6 was to identify key variables that influence *Physalia* presence on beaches around New Zealand. Artificial neural networks trained on time stepped presence only data as in Chapter 3 and 4 showed an ability to identify variables that effected *Physalia* populations. By time stepping (Chapter 3 and 4) the data it allowed the identification and quantification of specific variable contribution to the network and hence overall system, but because of limitations, such as class imbalance and risk of overtraining, only general patterns where sought. Furthermore, as the limitations in the data were substantial, a NBC driven by a vQEA was trained on the same presence data as a comparison. By using the variables identified by both ANN and NBC as an ensemble it was possible to accurately identify variables with high contribution to the system. Greater precision is often gained with ensembles (Araujo & New 2007; Lankin-Vega *et al.* 2008).

The ensemble of the ANN and NBC was able to identify 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, Taranaki, Wellington and Canterbury regions (figure 6.4). Both models indicated the existence of both systems but when used as an ensemble gave the detail needed to hypothesise possible drivers of the systems. A strong confirmation for the existence of two systems are the results of the distribution of specimens identified through molecular techniques. The clan distribution identified in Chapter 2 closely matches the two systems identified by the models in Chapter 6, with one clan restricted to an area encompassing Bay of Plenty south to Riversdale, and another major New Zealand clan found throughout New Zealand except for the Bay of Plenty. Because of the similarities between the model predictions and molecular

patterns combined with other observations such as algal blooms (Murphy *et al.* 2001) and coastal currents (Heath 1986; Stanton *et al.* 1997), there is a degree of support for hypothesises that may direct further research to support the findings in this study.

7.3 Future directions

There are a number of potential research areas indicated by the findings in this study. Further research could clarify and confirm the findings here but there are new directions possible. Of particular importance is to clarify which species of *Physalia* occurs in New Zealand waters. To date, this research has assumed that all individuals within the complex behave in a similar matter because of the indication that as only one species was present (Wesrerskov & Probert 1981; Slaughter *et al.* 2009). Whereas there may be differences between clans that, if accounted for within the model, may be critical for increasing the resolution of the models and gaining a more detailed understanding of the system.

To clarify which species of *Physalia* actually occurs in New Zealand several approaches are possible. Given the relative ease of obtaining specimens from beaches, a wider and more intensive sampling strategy could be employed around New Zealand once funding is secured to process the additional specimens. The resultant specimens could be sequenced to increase the sample size, giving potentially greater resolution, and more information, of the relationships between clans. Also a detailed morphological analysis to determine if the molecular data correlates with the morphological characteristics of *Physalia* would be desirable. Such an integrated taxonomic approach (Dayrat 2005) has considerable benefits. Further assessment of the New Zealand specimens will not however, solve the basic ambiguity and confusion surrounding this genus, which will require a complete global review of its taxonomy.

There are a number of opportunities for further model development, both for the forecast of *Physalia* as well as understanding the oceanographic systems in which the genus inhabits. Key to future model development is improvement in the quantity and quality of all aspects of the data. Improvements to the dataset are necessary to overcome the limitations encountered with both the explanatory variables and the occurrence data. For the explanatory variables other factors should be considered such as sea surface temperature and chemical composition of the water and information on the direction and strength of the coastal currents. Both systems identified are potentially driven by the currents augmented by winds and swells requiring the incorporation of current variables. Unfortunately such data is not easily accessible and may require the implementation of a coastal monitoring program. Another possibility of obtaining such information was proposed by Bowen *et al.* (2004), they suggested the possibility of using satellite images to observe ocean features such as plankton bloom movements and infer the speed and direction of currents from such observations. This approach maybe of merit, however, the feasibility of incorporating such data into the models is limited. Despite this, such images allow, plankton blooms to be regularly identified around the New Zealand coast (Murphy *et al.* 2001). As *Physalia* are predatory and like all blooming jellyfish, require a surplus of resources to bloom (Mills 2001; Purcell 2005) the identification of a period that a bloom can occur would provide a biofix that would significantly enhance a forecast model. Moreover, if the resolution of the forecast models can be improved to reduce apparent false positives then further development and implementation of a warning system for lifeguards and beachgoers is possible. As the entire precursor information is freely available from the internet (ftp://polar.ncep.noaa.gov/pub/waves/latest_run) there is no limitation to downloading appropriate oceanographic data to automate the process and make forecasts available on the internet.

In conjunction with obtaining and improving the environmental data a concerted effort is required to improve and enlarge the occurrence data. For this research only the 2000/2001 summer season to the 2004/2005 summer season were analysed. The incorporation of data of the four seasons that have elapsed since this study began would effectively double the available data. The additional data would allow other regions to be model for example, Riversdale through to Gisborne as this region amongst others were rejected for modelling because of a lack of data. The Riversdale region is of particular interest as it was the only location sampled that had both Clan 1 and Clan 3 present (Chapter 2). If the models are accurately modelling the hypothesised system then when conditions favour *Physalia* being transported to the beach from the south, individuals should conform to the Clan 1 genotype. Whereas under more northerly conditions then it would be expected that individuals would belong to Clan 3, having been transported south from the Bay of Plenty as this is the only other location where Clan 3 is located. Because Brodie (1960) showed with float cards that there is a general ocean movement down the North Island east coast it is reasonable that individual from Clan 3 would originate from the north rather than the south.

As the quality of the data was less than ideal, in respect to recording truce absence, a small trial with Surf Lifesaving Canterbury was conducted in an attempt to improve it over the summer of 2008/2009 which removed the uncertainty of prediction for that season. Lifeguards recorded and reported to the author if they observed *Physalia* whilst on duty. An incident of someone being stung was used as a proxy for the presence of *Physalia*. Because the lifeguards train and work in other areas outside of the patrol area on the beach on any day and at a distance from shore, absence of Physalia in the general area could be confirmed removing uncertainly in absence data. In this way the effective sample area was dramatically increased improving the quality of the occurrence data for that season. If lifeguards were encouraged to record absences that could improve model forecasts overall.

The occurrence of *Physalia* on beaches reflects the presence of populations in the general locality and then local conditions determine their arrival in the swimming areas. Several lifeguards have commented that on occasions they have seen *Physalia* more than a kilometre offshore but no *Physalia* were reported by lifeguards on patrol on the beach. These observations suggest a more complex interaction between the littoral and sublittoral zones that determine *Physalia* occurrence, than what has been modeled in this study. To overcome this limitation a two step model would be appropriate with the first step comprising a prediction of the general presence of jellyfish in a region followed by the second step using local conditions determine the likelihood of their presence in the swimming areas on beaches.

With the increase in invasive marine species that are dispersed passively (Lewis *et al*. 2005; McQuaid & Phillips 2000) models that can accurately forecast potential rates and pattern of spread will assist in assessing high risk areas and the control of such species. This study has proposed several hypotheses about the important drivers of the systems identified in this research. The results may be extrapolated to identify potential regions where jellyfish blooms may occur. Plankton blooms identified from satellite imagery (potentially free on the internet) could define an initial search area to confirm the presence of jellyfish blooms. Tracking technology can be used to determine the rate of movement of model specimens as well as drift for incorporation in a model. Such an investigation can be used for any species that is dispersed by passive surface movement. Such studies could contribute to a greater understanding of passive dispersal of organisms in oceanographic systems.

Much literature has suggested that oceans provide little barrier to gene flow so there are few opportunities for allopatric divergence (Palumbi 1992; Knowlton 2000; Dawson & Jacobs 2001). This thesis has shown the distribution and composition of the New Zealand *Physalia*

represents a species complex and is not a cosmopolitan species as would be expected. Moreover, the genetic diversity found indicated that there was some form of restriction to the flow of genetic information within *Physalia* else the genetic diversity would have been more homogeneous. Furthermore the models have indicated a potential mechanism that may explain how the complex could have arisen. To test for a mechanism that may have restricted gene flow it may be necessary to develop a circulation model for coastal and near coastal New Zealand for both surface and subsurface water movement. The latter would be necessary because as *Physalia* gametes are shed into the depths (Totton 1960) and initial larval development and dispersal would be determined by subsurface currents. Such a model could be used for any marine organism that at some stage in its life cycle has a passive dispersal phases and could have important applications in fisheries as well as marine biosecurity research. Critical to model success would be the independent biological information on species distributions to provide model validation.

Acknowledgements

As I reflect back on what has gone before I consider myself extraordinarily lucky to have been at the right places at the right times and known the right people even to get the idea to make this project beyond a mere possibility. But an idea is just that an idea without people to shape and drive it forward, for that Sue, Mike and Rob you will have my undying gratitude for all your help time and passion, it has been more than inspirational. Though I don't know who has had the harder job though, Sue having to try and guide my ideas to form a PhD at the same time put up with my writing, Mike teaching me to program computers from scratch or Rob teaching me molecular techniques from scratch, but thank you all it's been fun and challenging.

Apart from the above there has been a wide and varied cast and crew at Lincoln I need to pay tribute to for everything from sanity checks to deep philosophical discussions about undergrad teaching. Hazel, Katherine, Sofia, Jagoba, Nathan, Jon, Amber and many others your support, friendship and understanding really made a difference, although you never came jellyfish hunting with me… thanks people. Then there are the people like Steve A, Steve T, Dougal, Al, Andrew, Russel, Haley, Colin, Morgan, Kate and Noel. These people and many others from the pool and the surf have given me a place to go and forget about code, DNA and first years and concentrate on more important things like life, trust, friendship, and in some cases literally just keeping my or normally someone else's head above water.

Special thanks must be given to Surf Life Saving Taranaki, Michael and Pip Taylor, the Mount Maunganui Lifeguard Service, Glenn Moore, Jared Finchley, Sam Brown and Tom Trnski for collecting the jellyfish and Brett Sullivan from SLSNZ for providing the occurrence data. Without these people this project would not have happen.

Lastly to my family, well I guess I can answer your question: it's done. Thankyou all for everything, this would not have even be a possibility with out your time, patience and counsel over the years. Most of all thanks for just being there throughout this journey I'm on.

References

Adya, M. & Collopy, F. (1998) How effective are neural networks at forecasting and prediction? A review and evaluation. *Journal of Forecasting,* **17,** 481-495.

Alam, J.M., Qasim, R., Ali, S.A., Jamal, Q. & Alam, S.M. (2002) Biochemical characterization and pathophysiological properties of two high molecular weight cytolytic proteins from the venom of a coelenterate (Jellyfish), *Physalia utriculus* (blue bottle). *Pakistan Journal of Zoology,* **34,** 9-17.

Arai, M.N. (2001) Pelagic coelentreates and eutrophication: a review. *Hydrobiologia,* **451,** 69-87.

Araujo, L.M. & New, M. (2007) Ensemble forecasting of species distributions. *Trends in Ecology & Evolution,* **22,** 42-47.

Bailey, P.M., Little, M., Jelinek, G.A. & Wilce, J.A. (2003) Jellyfish envenoming syndromes: unknown toxic mechanisms and unproven therapies. *Medical Journal of Australia,* **178,** 34-37.

Baker, R.J. & Bradley, R.D. (2006) Speciation in mammals and the genetic species concept. *Journal of Mammalogy,* **87,** 643-662.

Ballard, J.W.O. & Whitlock, M.C. (2004) The incomplete natural history of mitochondria. *Molecular Ecology,* **13,** 729-744.

Barnay, A.S., Ellien, C., Gentil, F. & Thiebaut, E. (2003) A model study on variations in larval supply: are populations of the polychaete *Owenia fusiformis* in the English Channel open or closed? *Helgoland Marine Research,* **56,** 229-237.

Barnes, R. (1980) *Invertebrate Zoology*. Saunders College, Philadelphia, US.

Batchelor, W.D. (1998) Fundamentals of neural networks. *Agricultural Systems modelling and simulation* (eds R.M. Peart & R. Bruce), pp. 597-628. Marcel Dekker, New York.

Bensasson, D., Zhang, D.X., Hartl, D.L. & Hewitt, G.M. (2001) Mitochondrial pseudogenes: evolution's misplaced witnesses. *Trends in Ecology & Evolution,* **16,** 314-321

Bouillon, J. & Barnett, T.J. (1999) *The marine fauna of New Zealand. Hydromedusae (Cnidaria : Hydrozoa)*. National Institute of Water and Atmospheric Research, NIWA, Wellington, New Zealand.

Bouillon, J., Gravili, C., Pages, F., Gili, J. & Boero, F. (2006) *An introduction to Hydrozoa*. Publications scientifiques du museum, Paris.

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.

Bowen, M., Richardson, K., Pinkerton, M.H., Korpela, A. & Uddstrom, M. (2004) Squeezing information from an elusive ocean: surface currents from satellite imagery. *Water and Atmopshere*, Vol. 12, pp. 26-27. NIWA, Wellington.

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

Brosse, S., Giraudel, J.L. & Lek, S. (2001) Utilisation of non-supervised neural networks and principal component analysis to study fish assemblages. *Ecological Modelling,* **146,** 159-166.

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.

Brotons, L., Thuiller, W., Araujo, M.B. & Hirzel, A.H. (2004) Presence-absence versus presence-only modelling methods for predicting bird habitat suitability. *Ecography,* **27,** 437-448.

Carson, R. (1965) *The sea around us*. Panther, London, United Kingdom.

Chen, M.C., Chen, L.S., Hsu, C.C. & Zeng, W.R. (2008) Information granulation based data mining approach for classifying imbalanced data. *Information Sciences,* **178,** 3214-3227.

Chiswell, S.M. & Roemmich, D. (1998) The East Cape Current and two eddies: a mechanism for larval retention. *New Zealand Journal of Marine and Freshwater Research,* **32,** 385-397.

Clark, S.J. (2007) *Models for ecological data*. Princeton University Press, Princeton.

Cocu, N., Harrington, R., Rounsevell, M.D.A., Worner, S.P. & Hulle, M. (2005) Geographical location, climate and land use influences on the phenology and numbers of the aphid, *Myzus persicae*, in Europe. *Journal of Biogeography,* **32,** 615-632.

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.

Collins, A.G., Bentlage, B., Lindner, A., Lindsay, D., Haddock, S.H.D., Jarms, G., Norenburg, J.L., Jankowski, T. & Cartwright, P. (2008) Phylogenetics of Trachylina (Cnidaria: Hydrozoa) with new insights on the evolution of some problematical taxa. *Journal of the Marine Biological Association of the United Kingdom,* **88,** 1673-1685.

Collins, A.G., Winkelmann, S., Hadrys, H. & Schierwater, B. (2005) Phylogeny of Capitata and Corynidae (Cnidaria, Hydrozoa)in light of mitochondrial 16S rDNA data. *Zoologica Scripta,* **34,** 91-99.

Collins, F.H., Kamau, L., Ranson, H.A. & Vulule, J.M. (2000) Molecular entomology and prospects for malaria control. *Bulletin of the World Health Organization,* **78,** 1412-1423.

Cotter, A.J.R., Burt, L., Paxton, C.G.M., Fernandez, C., Buckland, S.T. & Pax, J.X. (2004) Are stock assessment methods too complicated? *Fish and Fisheries,* **5,** 235-254.

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.

Crick, F. (1989) The recent excitement about neural networks. *Nature,* **337,** 129-132.

Cronin, M.A. (1993) Mitochondrial DNA in wildlife taxonomy and conservation biology: cautionary notes. . *Wildlife Society Bulletlin,* **21,** 339-348.

Daewel, U., Peck, M.A., Kuhn, W., St John, M.A., Alekseeva, I. & Schrum, C. (2008) Coupling ecosystem and individual-based models to simulate the influence of environmental variability on potential growth and survival of larval sprat (*Sprattus sprattus* L.) in the North Sea. *Fisheries Oceanography,* **17,** 333-351.

Dawson, C.W. & Wilby, R. (1998) An artificial neural network approach to rainfall runoff modelling. *Hydrological Sciences Journal-Journal des sciences hydrologiques,* **43,** 47-66.

Dawson, M.N. (2003) Macro-morphological variation among cryptic species of the moon jellyfish, *Aurelia* (Cnidaria: Scyphozoa). *Marine Biology,* **143,** 369-379.

Dawson, M.N. & Jacobs, D.K. (2001) Molecular Evidence for Cryptic Species of *Aurelia aurita* (Cnidaria, Scyphozoa). *Biological Bulletin,* **200,** 92-96.

Dayrat, B. (2005) Towards integrative taxonomy. *Biological Journal of the Linnean Society,* **85,** 407–415.

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

Dunn, C.W., Pugh, P.R. & Haddock, S.H.D. (2005) Molecular phylogenetics of the siphonophora (Cnidaria), with implications for the evolution of functional specialization. *Systematic Biology,* **54,** 916-935.

Eckman, J.E. (1996) Closing the larval loop: Linking larval ecology to the population dynamics of marine benthic invertebrates. *Journal of Experimental Marine Biology and Ecology,* **200,** 207-237.

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.

Elizondo, D.A., McClendon, R.W. & Hoogenboom, G. (1994) Neural network models for predicting flowering and physiological maturity of soybean. *Transactions of the ASAE,* **37,** 981-988.

Ellien, C., Thiebaut, E., Barnay, A.S., Dauvin, J., Gentil, F. & Salomon, J. (2000) The influence of variability in larval dispersal on the dynamics of a marine metapopulation in the eastern Channel. *Oceanologica Acta,* **23,** 423-442.

Ellien, C., Thiebaut, E., Dumas, F., Salomon, J. & Nival, P. (2004) A modelling study of the respective role of hydrodynamic processes and larval mortality on larval dispersal and recruitment of benthic invertebrates: example of Pectinaria koreni (Annelida: Polychaeta) in the Bay of Seine (English Channel). *Journal of Plankton Research,* **26,** 117-132.

Fenner, P.J. (1997) *The Global problem of Cnidarian (Jellyfish) Stinging*. University of London, London.

Fischlin, A., Midgley, G.F., Price, J.T., Leemans, R., Gopal, B., Turley, C., Rounsevell, M.D.A., Dube, O.P., Tarazona, J. & Velichko, A.A. (2007) Ecosystems, their properties, goods and services. *Climate Change 2007: Impacts, Adaptations and Vulnerability. Contribution of Working Group II to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change* (eds M.L. Parry, O.F. Canziani, P.J. Palutikof, P.J. van der Linen & C.J. Hanson), pp. 211-272. Cambridge University Press, Cambridge.

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. *Cybernetics and Systems '96, Proceedings of the 13th European Meeting on Cybernetics and Systems Research* (ed R. Trappl), pp. 1005-1008. Austrian Society for Cybernetic Studies.

Friedman, N., Geiger, D. & Goldszmidt, M. (1997) Bayesian network classifiers. *Machine Learning,* **29,** 131-163.

Funk, D.J. & Omland, K.E. (2003) Species-level paraphyly and polyphyly: frequency, causes, and consequences, with insights from animal mitochondrial DNA. *Annual Review of Ecology and Systematics,* **34,** 397-423.

Gardner, D.M. (1961) Hydrology of New Zealand Coastal Waters. *Bulletin of the New Zealand Department of Scientific and Industrial Research,* **138,** 1-84.

Gevery, M., Dimopoulos, I. & Lek, S. (2003) Review and comparison of methods to study the contribution of variables in artificial neural network models. *Ecological Modelling,* **160,** 249-264.

Gevrey, M. & Worner, S.P. (2006) Prediction of global distribution of insect pest species in relation to climate by using an ecological informatics method. *Journal of Economic Entomology,* **99,** 979-986.

Gotelli, N.J. (2000) Null model analysis of species co-occurrence patterns. *Ecology,* **81,** 2606-2621.

Goulet, T.L. & Coffroth, M.A. (2003) Genetic composition of zooxanthellae between and within colonies of the octocoral *Plexaura kuna*, based on small subunit rDNA and multilocus DNA fingerprinting. *Marine Biology,* **142**.

Govindarajan, A.F., Halanych, K.K. & Cunningham, C.W. (2005) Mitochondrial evolution and phylogeography in the hydrozoan *Obelia geniculata* (Cnidaria). *Marine Biology,* **146,** 213-222.

Graham, C.H., Ferrier, S., Huettmann, F., Moritz, C. & Peterson, A.T. (2004) New developments in museum-based informatics and applications in biodiversity analysis *Trends in Ecology & Evolution,* **19,** 497-503.

Graham, W.M., Pages, F. & Hammer, W.M. (2001) A physical context for gelatinous zooplankton aggregations: a review. *Hydrobiologia,* **451,** 199-212.

Greig, M.J., Ridgway, N.M. & Shakespeare, B.S. (1988) Sea surface temperature variations at coastal sites around New Zealand. *New Zealand Journal of Marine and Freshwater Research,* **22,** 391-400.

Guisan, A. & Thuiller, W. (2005) Predicting species distribution: offering more than simple habitat models. *Ecology Letters,* **8,** 993-1009.

Guyon, I. & Elisseeff, A. (2003) An Introduction to Variable and Feature Selection. *Journal of Machine Learning Research,* **3** 1157-1182.

Haddock, S.H.D. (2004) A golden age of gelata: past and future research on planktonic ctenophores and cnidarians. *Hydrobiologia,* **530,** 549-556.

Han, K.H. & Kim, J.H. (2004) Quantum-inspired evolutionary algorithms with a new termination criterion, h└ gate, and two phase scheme. *IEEE Transactions on Evolutionary Computation,* **8,** 156–169.

Hanley, J.A. & McNeil, B.J. (1982) The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Diagnostic Radiology,* **143,** 29-36.

Head, I.M., Saunders, J.R. & Pickup, R.W. (1998) Microbial evolution, diversity, and ecology: A decade of ribosomal RNA analysis of uncultivated microorganisms. *Microbial Ecology,* **35,** 1-21.

Heath, R.A. (1985) A review of the physical oceanography of the seas around New Zealand - 1982. *New Zealand Journal of Marine and Freshwater Research,* **19,** 79-124.

Heath, R.A. (1986) In which direction is the mean flow through Cook Strait, New Zealand — evidence of 1 to 4 week variability? *New Zealand Journal of Marine and Freshwater Research,* **20,** 119-137.

Hebert, P.D.N., Cywinska, A., Ball, S.L. & DeWaard, J.R. (2003) Biological identifications through DNA barcodes. *Proceedings of the Royal Society of London - Biological Sciences* **270,** 313-322.

Hebert, P.D.N., Ratnasingham, S. & deWaard, J.R. (2003) Barcoding animal life: cytochrome c oxidase subunit 1 divergences among closely related species. *Proceedings of the Royal Society of London Series B-Biological Sciences,* **270,** S96-S99.

Hill, B.D., McGinn, S.M., Korchinski, A. & Burnett, B. (2002) Neural network models to predict the maturity of spring wheat in western Canada. *Canadian Journal of Plant Science,* **82,** 7-13.

Hills, D.M. & Davis, S.K. (1986) Evolution of ribosomal DNA: fifty million years of recorded history in the frog genus *Rana*. *Evolution,* **40,** 1275-1288.

Hinton, G.E., McClelland, J.L. & Rumelhart, D.E. (1986) Distributed representations. *Parallel Distributed Processing* (eds D.E. Rumelhart & J.L. McClelland), pp. 77-109. MIT Press, Cambridge.

Holland, B.S., Dawson, M.N., Crow, G.L. & Hofmann, D.K. (2004) Global phylogeography of *Cassiopea* (Scyphozoa: Rhizostomeae): molecular evidence for cryptic species and multiple invasions of the Hawaiian Islands. *Marine Biology,* **146,** 1119-1128.

Huang, D.W., Meier, R., Todd, P.A. & Chou, L.M. (2008) Slow mitochondrial COI sequence evolution at the base of the metazoan tree and its implications for DNA barcoding. *Journal of Molecular Evolution,* **66,** 167-174.

Hudson, D.H. & Bryant, D. (2006) Application of Phylogenetice Networks in Evolutionary Studies. *Molecular Biology and Evolution,* **23,** 254-267

Hwang, D., Rust, A.G., Ramsey, S., Smith, J.J., Leslie, D.M., Weston, A.D., de Atauri, P., Aitchison, J.D., Hood, L., A.F., S. & Bolouri, H. (2005) A data integration methodology for systems biology. *Proceedings of the National Academy of Sciences of the United States of America,* **102,** 17296-17301.

Iosilevskii, G. & Weihs, D. (2009) Hydrodynamics of sailing of the Portuguese man-of-war *Physalia physalis*. *Jounal of the royal society interface,* **6,** 613-629.

Johnson, D.R., Harriet, P.M. & Burke, D.W. (2001) Developing jellyfish strategy hypotheses using circulation models. *Hydrobiologia,* **451,** 213-221.

Johnson, D.R. & Perry, H.M. (1999) Blue crab larval dispersion and retention in the Mississippi Bight *Bulletin of Marine Science,* **65,** 129-149.

Johnston, P.A., Archer, E.R.M., Vogel, C.H., Bezuidenhout, C.N., Tennant, W.J. & Kuschke, R. (2005) Review of seasonal forecasting in South Africa: producer to end-user. *Climate Research,* **28,** 67-82.

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. & Lester, S.E. (2005) Propagule dispersal and the scales of marine community process. *Diversity and Distributions,* **11,** 139-148.

Kerr, K.C.R., Stoeckle, M.Y., Dove, C.J., Weigt, L.A., Francis, C.M. & Hebert, P.D.N. (2007) Comprehensive DNA barcode coverage of the North American birds. *Molecular Ecology Notes,* **7,** 535-543.

Knowlton, N. (2000) Molecular genetic analyses of species boundaries in the sea. *Hydrobiologia,* **420,** 73–90.

Kohonen, T. (1990) The Self-Organizing Map. *Proceedings of the IEEE,* **78,** 1464-1479.

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.

Kotsiantis, S.B. (2007) Supervised machine learning: A review of classification techniques. *Informatica*, **31**, 249-268.

Lamark, J.B. (1801) *Systeme des animaux sans vertebres*. Museum National d'Histoire Naturelle Paris.

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.

Le Goff-Vitry, M.C., Rogers, A.D. & Baglow, D. (2004) A deep-sea slant on the molecular phylogeny of the Scleractinia. *Molecular Phylogenetics and Evolution,* **30,** 167-177.

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.

Lek, S., Giraudel, J.L. & Guegan, J.F. (2000) Neuronal Networks:Algorithms and architures for ecologist and evelutionary ecologists. *Artificial Neuronal Networks* (eds S. Lek & J.L. Giraudel), pp. 3-27. Springer, Berlin.

Lewis, P.N., Riddle, M.J. & Smith, S.D.A. (2005) Assisted passage or passive drift: a comparison of alternative transport mechanisms for non-indigenous coastal species into the Southern Ocean. *Antarctic Science,* **17,** 183-191.

Liang, C., Das, K.C. & McClendon, R.W. (2003) Prediction of microbial activity during biosolids composting using artificial neural networks. *Transactions of the ASAE,* **46,** 1713-1719.
Liu, C., Berry, P.M., Dawson, T.P. & Pearson, R.G. (2005) Selecting thresholds of occurrence in the prediction of species distributions. *Ecography,* **28,** 385-393.
Liu, J., Hu, Q. & Yu, D. (2008) A weighted rough set based method developed for class imbalance learning. *Information Sciences,* **178,** 1235-1256.
Loytynoja, A. & Goldman, N. (2005) An algorithm for progressive multiple alignment of sequences with insertions. *Proceedings of the National Academy of Sciences of the United States of America,* **102,** 10557-10562.
Lynam, C.P., Hay, S.J. & Brierley, A.S. (2004) Interannual variability in abundance of North Sea jellyfish and links to the North Atlantic Oscillation *Limnology and Oceanography,* **49,** 637-643.
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.
Mandojana, R.M. (1990) Granuloma annulare following bluebottle jellyfish (*Physalia utriculus*) sting. *Journal of Wilderness Medicine,* **1,** 220-224.
Manel, S., Dias, J.M. & Ormerod, S.J. (1999) Comparing discriminate analysis, neural networks and logistic regression for predicting species distributions: a case study with a Himalayan river bird. *Ecological Modelling,* **120,** 337-347.
Marmion, M., Hjort, J., Thuiller, W. & Luoto, M. (2008) A comparison of predictive methods in modelling the distribution of periglacial landforms in Finnish Lapland. *Earth Surface Processes and Landforms,* **33,** 2241-2254.
Marques, A.C. & Collins, A.G. (2004) Cladistic analysis of Medusozoa and cnidarian evolution. *Invertebrate Biology,* **123,** 23-42.
Mastrorillo, S., Lek, S., Dauba, F. & Belaud, A. (1997) The use of artificial neural networks to predict the presence of small-bodied fish in a river. *Freshwater Biology,* **38,** 237-246.
MathWorks (2008) MATLAB. The MathWorks Inc., Natick, Massachusetts.
McAllister, M.K. & Kirkwood, G.P. (1998) Bayesian stock assessment: a review and example application using the logistic model. *Ices Journal of Marine Science,* **55,** 1031-1060.
McFadden, C.S. (1999) Genetic and taxonomic relationships among Northeastern Atlantic and Mediterranean populations of the soft coral *Alcyonium coralloides*. *Marine Biology,* **133,** 171-184.
McFadden, C.S. & Hutchinson, M.B. (2004) Molecular evidence for the hybrid origin of species in the soft coral genus *Alcyonium* (Cnidaria: Anthozoa: Octocorallia). *Molecular Ecology,* **13,** 1495-1505.
McQuaid, C.D. & Phillips, T.E. (2000) Limited wind-driven dispersal of intertidal mussel larvae: in situ evidence from the plankton and the spread of the invasive species *Mytilus galloprovincialis* in South Africa. *Marine Ecology-Progress Series,* **201,** 211-220.
Meyer, C.P. & Paulay, G. (2005) DNA barcoding: error rates based on comprehensive sampling. *Public Library of Science, Biology,* **3,** 2229-2238.
Mills, C.E. (2001) Jellyfish blooms: are populations increasing globally in response to changing ocean conditions? *Hydrobiologia,* **451,** 55-68.
Moritz, C. (1994) Applications of mitochondrial DNA analysis in conservation: a critical review. *Molecular Ecology,* **3,** 401-411.
Moura, C.J., Harris, D.J., Cunha, M.R. & Rogers, A.D. (2008) DNA barcoding reveals cryptic diversity in marine hydroids (Cnidaria, Hydrozoa) from coastal and deep-sea environments. *Zoologica Scripta,* **37,** 93-108.

Murphy, R.J., Pinkerton, M.H., Richardson, K.M., Bradford-Grieve, J.M. & Boyd, P.W. (2001) Phytoplankton distributions around New Zealand derived from SeaWiFS remotely-sensed ocean colour data. *New Zealand Journal of Marine and Freashwater Research,* **35,** 343-362.

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. & 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. (2002a) A comparison of statistical approaches for modelling fish species distributions. *Freshwater Biology,* **47,** 1976-1995.

Olden, J.D. & Jackson, D.A. (2002b) 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,** 389-379.

Olden, J.D., Lawler, J.J. & Poff, N.L. (2008) Machine learning methods without tears: A primer for ecologists. *Quarterly Review of Biology,* **83,** 171-193.

Olson, R.R. (1985) The consequences of short-distance larval dispersal in a sessile marine invertebrate. *Ecology,* **66,** 30-39.

Ozesmi, S.L. & Ozesmi, U. (1999) An artificial neural network approach to spatial habitat modelling with interspecific interaction. *Ecological Modelling,* **166,** 15-31.

Pages, F. & Gili, J. (1992) Siphonophores (Cnidaria,Hydrozoa) of the Benguela Current (southeastern Atlantic). *Scientia Marina* **56,** 65-112.

Palumbi, S.R. (1992) Marine speciation on a small planet. *Trends in Ecology & Evolution,* **7,** 114-118.

Pelletier, D. & Mahevas, S. (2005) Spatially explicit fisheries simulation models for policy evaluation. *Fish and Fisheries,* **6,** 307-349.

Platel, M.D., Schliebs, S. & Kasabov, N. (2007) A versatile quantum-inspired evolutionary algorithm. . *IEEE Congress on Evolutionary Computation*, pp. 423–430. Singapore.

Platel, M.D., Schliebs, S. & Kasabov, N. (2008) Quantum inspired evolutionary algorithm: A multimodel eda. *IEEE Transactions on Evolutionary Computation***,** In print.

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. *International Joint Conference on Neural Networks*, pp. 1171-1176. Hong Kong.

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.

Posada, D. & Crandall, K. (1998) MODELTEST: testing the model of DNA substitution. *Bio Informatics,* **14,** 817-818.

Prechelt, L. (1996) A Quantitative Study of Experimental Evaluations of Neural Network Learning Algorithms: Current Research Practice. *Neural Networks,* **9,** 457-462.

Purcell, J.E. (2005) Climate effects on formation of jellyfish and ctenophore blooms: a review. *Journal of the Marine Biological Association of the United Kingdom,* **85,** 461-476.

Purcell, J.E., Uye, S. & Lo, W. (2007) Anthropogenic causes of jellyfish blooms and their direct consequences for humans: a review *Marine and Ecology Progress Series,* **350,** 153-174.

Robinson, B.H., Reisenbichler, K.R., Sherlock, R.E., Silguero, J.M.B. & Chavez, F.P. (1998) Seasonal abundance of the siphonophore *Nanomia bijuga* in Monterey Bay. *Deep Sea Research II,* **45,** 1741-1751.

Rumelhart, D.E., Hinton, G.E. & Williams, R.J. (1986) Learning representations by back-propagation errors. *Nature,* **323,** 533-536.

Rumelhart, D.E. & McClelland, J.L. (1986) *Parallel distributed processing: Explorations in the microstructure of cognition*. MIT Press, Cambridge.

Seymour, J.E., Carrette, T.J. & Sutherland, P.A. (2004) Do box jellyfish sleep at night? *The Medical Journal of Australia,* **181,** 707.

Shanks, A.L., Grantham, B.A. & Carr, M.H. (2003) Propagule dispersal distance and the size and spacing of marine reserves. *Ecological Applications,* **13,** S159-S169.

Shearer, T.L. & Coffroth, M.A. (2008) Barcoding corals: by interspecific divergence, not intraspecific variation. *Molecular Ecology Resources,* **8,** 247-255.

Shearer, T.L., Van Oppen, M.J.H., Romano, S.L. & Wörheide, G. (2002) Slow mitochondrial DNA sequence evolution in the Anthozoa (Cnidaria). *Molecular Ecology,* **11,** 2475-2487.

Sheiner, L.B. & Beal, S.L. (1981) Some Suggestions for Measuring Predictive Performance. *Journal of Pharmacokinetics and Biopharmaceutics,* **9,** 503-512.

Siegel, D.A., Kinlan, B.P., Gaylord, B. & Gaines, S.D. (2003) Lagrangian descriptions of marine larval dispersion. *Marine Ecology-Progress Series,* **260,** 83-96.

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.

Smith, F. & Gordon, D. (2003) Sessile Invertebrates. *The Living Reef: The Ecology of New Zealand's Rocky Reefs* (eds N. Andrew & M. Francis), pp. 80-91. Craig Potton Publishing, Nelson, New Zealand.

Stanski, H.R., Wilson, L.J. & Burrows, W.R. (1989) Survey of common verification methods in meteorology. p. 18. Atmospheric Environment Service, Ontario.

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.

Swofford, D. (2002) PAUP*: Phylogenetic Analysis Using Parsimony (and Other Methods) 4.0 Beta. Sinauer Associates, Sunderland.

Tang, J.M., Toe, L., Back, C. & Unnasch, T.R. (1996) Intra-specific heterogeneity of the rDNA internal transcribed spacer in the *Simulium damnosum* (Diptera: Simuliidae) complex. *Molecular Biology and Evolution,* **13,** 244-252.

Thomas, G.G., Goldwin, G.K. & Tatchell, G.M. (1983) Association between weather factors and the spring migration of the damson-hop aphid, *Phorodon humuli*. *Annals of Applied Biology,* **102**.

Toba, T., Lida, N., Kawamura, H., Educhi, N. & Jones, I.S.F. (1990) Wave dependence of sea-surface wind stress *Journal of Physical Oceanography,* **20,** 705-721.

Tolman, H.L. (1998) Validation of a new global wave forecast system at NCEP. *Ocean Wave Measurements and Analysis* (eds B.L. Edge & J.M. Helmsley), pp. 777-786. ASCE.

Totton, A.K. (1960) Studies on *Physalia physalis*: Natural history and morphology. *Discovery Reports,* **30,** 301-368.

van Oppen, M.J.H., Willis, H.W., van Vugt, W.J.A. & Miller, D.J. (2000) Examination of species boundaries in the Acropora cervicornis group (Scleractinia, Cnidaria) using nuclear DNA sequence analyses. *Molecular Ecology,* **9,** 1363-1373.

Venayagamoorthy, G.K., Singhal, G.J.o.C.a.T. & Nanoscience, December 2005. (2005) Quantum-inspired evolutionary algorithms and binary particle swarm optimization for training MLP and SRN neural networks. . *Journal of Computational and Theoretical Nanoscience,* **2,** 561–568.

Visen, N.S., Paliwal, J., Jayas, D.S. & White, N.D.G. (2002) Specialist Neural Networks for Cereal Grain Classification. *Biosystems Engineering,* **82,** 151-159.

Ward, R.D., Zemlak, T.S., Innes, B.H., Last, P.R. & Hebert, P.D.N. (2005) DNA barcoding Australia's fish species. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences,* **360,** 1847-1857.

Wesrerskov, K. & Probert, K. (1981) *The Seas Around New Zealand*. Reed Ltd., Wellington, New Zealand.

Wilkinson, M., McInerney, J.O., Hirt, R.P., Foster, P.G. & Embley, T.M. (2007) Of clades and clans: terms for phylogenetic relationships in unrooted trees. *Trends in Ecology and Evolution,* **22,** 114-115

Willis, B.L., Van Oppen, M.J.H., Miller, D.J., Vollmer, S.V. & Ayre, D.J. (2006) The role of hybridization in the evolution of reef corals. *Annual Review of Ecology and Systematics,* **37,** 489-517.

Worner, S.P., Lankin, G.O., Samarasinghe, S., Teulon, D.A.J. & Zydenbos, S.M. (2002) Improving prediction of aphid flights by temporal analysis of input data for an artificial neural network. *New Zealand Plant Protection,* **55,** 312-316.

Xu, L. & Chow, M.-Y. (2006) A classification approach for power distribution systems fault cause identification. *IEEE Transactions on Power Systems,* **21,** 53-60.

Yanagihara, A.A., Kuroiwa, M.Y., Louise, M. & Kunkel, D.D. (2002) The ultrastructure of nematocysts from the fishing tentacle of the Hawaiian bluebottle, *Physalia utriculus* (Cnidaria, Hydrozoa, Siphonophora). *Hydrobiologia,* **489,** 139-150.

Zaniewski, A.E., Lehmann, A. & Overton, J.M.C. (2002) Predicting species spatial distributions using presence-only data: a case study of native New Zealand ferns. *Ecological Modelling,* **157,** 261-280.

Zhang, G.Q., Patuwo, B.E. & Hu, M.Y. (1998) Forecasting with artificial neural networks: The state of the art. *International Journal of Forecasting,* **14,** 35-62.

Zhang, L.Q., Wang, G.T., Yao, W.J., Li, W.X. & Gao, Q. (2009) Molecular systematics of medusae in the genus Craspedacusta (Cnidaria: Hydrozoa: Limnomedusae) in China with the reference to the identity of species. *Journal of Plankton Research,* **31,** 563-570.

Appendix A

Maximum parsimony trees

A 1.0 Cytochrome c oxidase I

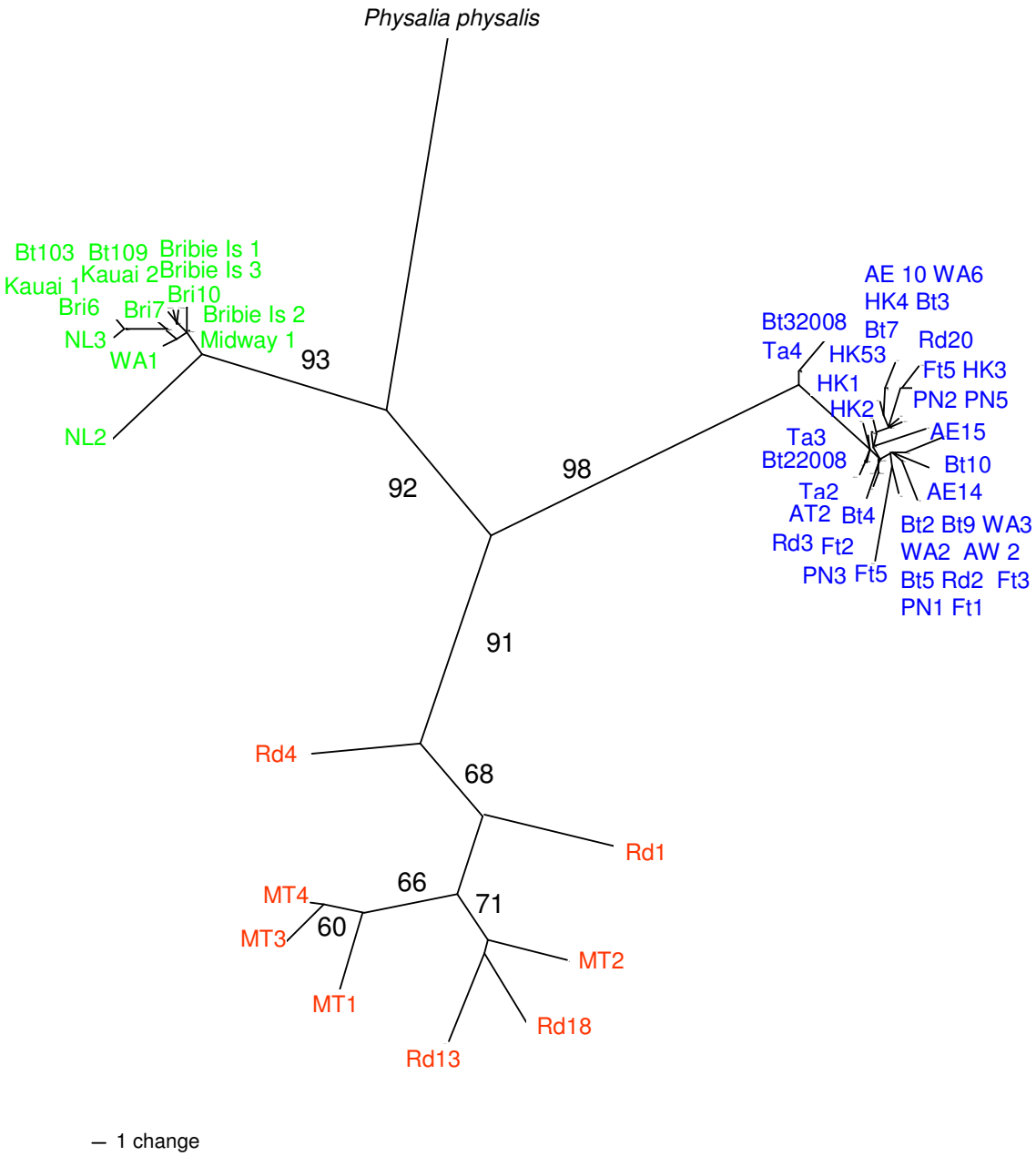


Figure A1.1: Unrooted maximum parsimony tree for COI. Numbers on branches indicate bootstrap support (1000 replicates).

A 2.0 Internal transcribed spacer

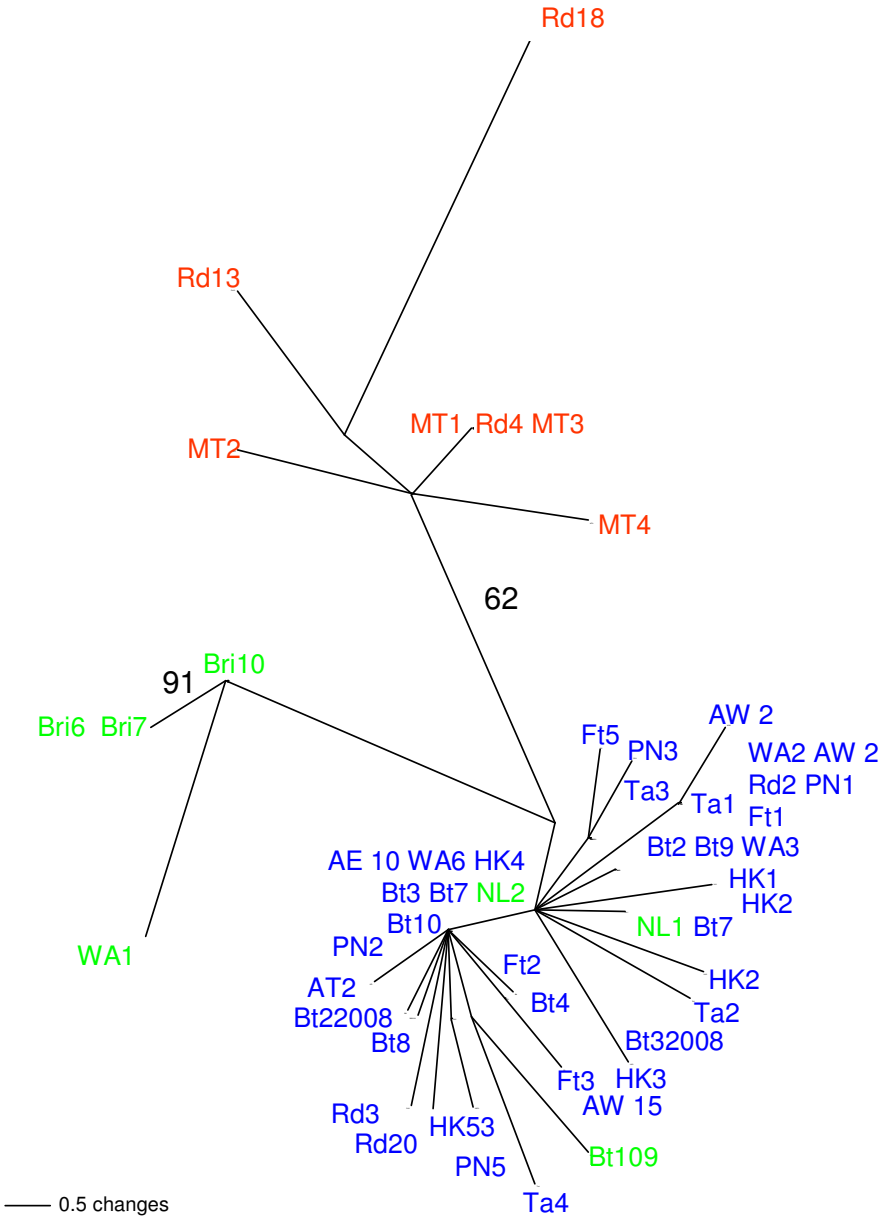


Figure A1.2: Unrooted maximum parsimony tree for ITS. Numbers on branches indicate bootstrap support (1000 replicates).

Appendix B

Associated publications

Using Multi-Layer Perceptrons to Predict the Presence of Jellyfish of the Genus *Physalia* at New Zealand Beaches

David R. Pontin[1], Michael J. Watts[1], and S. P. Worner[1]

[1] *Bio-Protection Research Centre, Lincoln University, Canterbury, New Zealand;*

International Joint Conference on Neural Networks, Hong Kong, pp. 1171-1176.

Abstract:

The apparent increase in number and magnitude of jellyfish blooms in the worlds oceans has lead to concerns over potential disruption and harm to global fishery stocks. Because of the potential harm that jellyfish populations can cause and to avoid impact it would be helpful to model jellyfish populations so that species presence or absence can be predicted. Data on the presence or absence of jellyfish of the genus *Physalia* was modelled using Multi-Layer Perceptrons (MLP) based on oceanographic data. Results indicated that MLP are capable of predicting the presence or absence of *Physalia* in two regions in New Zealand and of identifying significant biological variables.

Using Time Lagged Input Data to Improve Prediction of Stinging Jellyfish Occurrence at New Zealand Beaches by Multi-Layer Perceptrons

David R. Pontin[1], Sue P. Worner[1] and Michael J. Watts[2]

[1] Bio-Protection Research Centre, Lincoln University, Canterbury, New Zealand; [2] School of Biological Sciences, University of Sydney, NSW, Australia.

Lecture Notes in Computer Science, **5506,** 907-914.

Abstract:

Environmental changes in oceanic conditions have the potential to cause jellyfish populations to rapidly expand leading to ecosystem level repercussions. To predict potential changes it is necessary to understand how such populations are influenced by oceanographic conditions. Data recording the presence or absence of jellyfish of the genus *Physalia* at beaches in the West Auckland region of New Zealand were modelled using Multi-Layer Perceptrons (MLP) with time lagged oceanographic data as input data. Results showed that MLP models were able to generalise well based on Kappa statistics and gave good predictions of the presence or absence of *Physalia*. Moreover, an analysis of the network contributions indicated an interaction between wave and wind variables at different time intervals can promote or inhibit the occurrence of *Physalia*.

Appendix C

Matlab functions associated with genetic analysis

The matlab functions created for implementing a sliding window analysis for genetic data are presented below. Functions created for pre-processing oceanographic data are detailed on the accompanying compact disc to this thesis.

C 1.0 Obtaining the sliding window

```
function slidwindN (inputfile,sizofwindow,numbbasepairs,numbtaxa)

%function takes and input file containing sequence data in a nexus format
%as generated from Seeq (MAC) and passes a sliding window over the data
%writing a new nexus file in the target output directory for each window
%selected

%Inputs:  inputfile = the input nexus file
%         sizofwindow = size of sliding window desired
%         numbbasepairs = Length of the sequences
%         numbtaxa = number of taxa contained in the nexus file

%NOTE:   This code only works with windows that are multiples of 10
and    %       will generate a window size of 1+ the size specified.

%         Total sequence length must be the same

%         You must manually change the output directory in the code below


%Writen by David Pontin, Lincoln University 5/9/09 v1.0

%Functions called opennexus.m, changeDi.m

Windnumber=numbbasepairs-sizofwindow;
%calc number of windows to be used
Data=opennexus(inputfile,numbtaxa);
%reads the nexusfile into a cell array
Data{4,1}=changeDi(Data{4,1},numbbasepairs,sizofwindow);
%changes the nexus parameters to be correct for the new number of base
pairs
spacer=Data{7,1};
tf=isspace(spacer);
counter=0;
%ender for while loop
start=3;
%started at 3 to avoid [charter at start
    while counter==0,
        %loop identifies where the sequences beings
        if tf(1,start)==0
            counter=1;
        else
            start=start+1;
        end;
    end;
```

```
frount=spacer(1,1:start-1);
%changes the two alignment lines to match
endnumb=num2str(sizofwindow+1);
seq= [spacer(1,start:start+(sizofwindow-1)) endnumb ' ]'];
endseq=[frount seq];
Data{7,1}=endseq;
spacer=Data{8,1};
seq=spacer(1,start:start+sizofwindow);
frount=spacer(1,1:start-1);
endseq=[frount seq ' ]'];
Data{8,1}=endseq;
 starting=start;
for Windnum=1:Windnumber,
    %for the number of windows need
    Out=fopen(['Window\Window',num2str(Windnum),'.nex'],'w');
    %opens file
    %NOTE CHANGE OUTPUT DIRECTORY FILENAME HERE
    Dataseq=Data;
    for taxa=1:numbtaxa,
        %loop selects the target basepairs 1 taxa at a time
        working=Dataseq{(8+taxa),1};
        seq=working(1,starting:starting+sizofwindow);
        frount=working(1,1:start-1);
        endseq=[frount seq];
        Dataseq{8+taxa,1}=endseq;
    end;
    Nrows=numbtaxa+11;
    for BBB=1:Nrows,
        % loop writes new file out
    Line=char(Dataseq(BBB,1));
    Stransformed= num2str(Line);
    Outstring = [Stransformed, '\n'];
    fprintf(Out, Outstring);
    end;
    fclose(Out);
    starting=starting+1;
end;


function [output]=opennexus(inputfile,numbtaxa)
%reads the input nexusfile into a cell array
%called by slidwindN function
%Written by David Pontin, Lincoln University 5/9/09 v1.0

%Called by slidwindN.m
output=[];
infile=fopen(inputfile,'r');
CurrLine1=fgetl(infile);
A={CurrLine1};
output=[output;A];
CurrLine1=fgetl(infile);
A={CurrLine1};
output=[output;A];
CurrLine1=fgetl(infile);
A={CurrLine1};
output=[output;A];
CurrLine1=fgetl(infile);
A={CurrLine1};
output=[output;A];
CurrLine1=fgetl(infile);
A={CurrLine1};
output=[output;A];
CurrLine1=fgetl(infile);
A={CurrLine1};
```

```
output=[output;A];
CurrLine1=fgetl(infile);
A={CurrLine1};
output=[output;A];
CurrLine1=fgetl(infile);
A={CurrLine1};
output=[output;A];
for rowbumb=1:(numbtaxa),
    CurrLine1=fgetl(infile);
    A={CurrLine1};
    output=[output;A];
end;
CurrLine1=fgetl(infile);
A={CurrLine1};
output=[output;A];
CurrLine1=fgetl(infile);
A={CurrLine1};
output=[output;A];
CurrLine1=fgetl(infile);
A={CurrLine1};
output=[output;A];
fclose(infile);


function [dime]=changeDi(input,numbbasepairs,sizofwindow)
%changes the nexus  parameters to correct for the new number of base pairs

%Called by slidwindN.m
%Written by David Pontin, Lincoln University 5/9/09 v1.0

target=['nchar=',num2str(numbbasepairs)];
out=['nchar=',num2str(sizofwindow+1)];
dime=strrep(input,target,out);
```

C 2.0 Analysing the sliding window

```
function [output]=raidlog(number_of_files,folder)
%this function reads paup log files of bootstrap output of a sliding window
%analysis, isolates the bootstrap output that conforms to the nodes in
%each file and then concatenates them all together. The resultant matrix is
%then sorted by the mean support for each node in descending order and then
%plotted out in a series of 3X3 subplots

%Inputs:  number_of_files = the number of input (bootstrap output) files
%         folder = folder where the input files are located

%NOTE:   the other lesser supported nodes are not plotted but
can          %     be viewed in the output file generated.


%M files called seqarray.m and sortcol.m

%Written by David Pontin Lincoln University 19/8/09 v1.0

for filenumb=1:number_of_files,

data=importdata([folder,'\logfile',num2str(filenumb),'.txt'],'\t',1000);
       startrow=strmatch('12345678901234567890',data);
       startrow=startrow+1;
       cont=1;
```

```
counter=1;
while cont==1,
    targetrow=data(startrow+counter,1);
    targetrow=char(targetrow);
    logica=isempty(targetrow);
    if logica(1,1)==1,
        cont=0;
    else
        counter=counter+1;
    end;%ifelse
end;%while
data=data(startrow+1:startrow+counter,1);
sz=size(data);
Nrows=sz(1)-1;
Strings={};
for rownumb=1:Nrows,
    row=char(data(rownumb,1));
    sx=size(row);
    ch=sx(2);
    tf=isspace(row);
    colnum=1;
    Start=1;
    End=2;
    cont=1;
    while cont==1,
        if colnum==3,
            Strings(rownumb,3)=cellstr(row(1,ch-5:ch-1));
            cont=0;
        else
            if tf(1,End+1)==1,
                if tf(1,End)==0,
                  Strings(rownumb,colnum)=cellstr(row(1,Start:End));
                    colnum=colnum+1;
                    Start=End;
                    End=End+1;
                else
                    End=End+1;
                end;
            elseif tf(1,End)==1
                if tf(1,End+1)==0,
                    Start=End+1;
                    End=End+2;
                end;
            else
                End=End+1;
            end;
        end;
    end;
end;
clear End Start ch colnum cont counter data logica startrow
clear targetrow tf sx
if filenumb==1,
    output=[Strings(:,1) Strings(:,3)];
else
     ss=size(output);
     outrows=ss(1);
     output=[output cellstr(num2str(zeros(outrows,1)))];
     counter=1;
     for rownumb=1:Nrows,
         target=char(Strings(rownumb,1));
         ss=size(output);
         outrows=ss(1);
         outcol=ss(2);
         Tx=[];
```

```
                    for Orow=1:outrows,
                        alt=char(output(Orow,1));
                        T=strcmp(target, alt);
                            if T==1,
                               output(Orow,outcol)=Strings(rownumb,3);
                            end;%if
                        Tx=[Tx T];
                    end;%for
                    Tx=sum(Tx);
                    if Tx==0,
                       sy=size(output);
                       NCOL=sy(2);
                       temp=cellstr(num2str(zeros(NCOL,1)));
                       temp=temp';
                       output=[output; temp];
                       output(Orow+counter,1)=Strings(rownumb,1);
                       output(Orow+counter,outcol)=Strings(rownumb,3);
                    end;%if
            end;%for
        end;
end;
 clear NCOL Nrows Orow Strings T Tx alt counter filenumb number_of_files
 clear outrow outcol outrows row rownumb ss sy sz target temp
 ss=size(output);
 Nrow=ss(1);
 Ncol=ss(2);
 headers=output(:,1);
 temp=output(:,2:Ncol);
 Data=[];
 for colnum=1:Ncol-1,
 tempe=char(temp(:,colnum));
 tempe=str2num(tempe);
 Data=[Data tempe];
 end;
 m=mean(Data,2);
 markers=seqarray(Nrow);
 Data=[markers Data m];
 Data=sortcol(Data,Ncol+1);
 figs=floor(Nrow/9);
 count=1;
 counter=1;
 Data(:,Ncol+1)=[];
 markers=Data(:,1);
 Data(:,1)=[];
 for fignum=1:figs,
    figure1 = figure('PaperSize',[29.68
20.98],'PaperOrientation','landscape');
    subplot1 = subplot(3,3,1,'Parent',figure1);
    xlim(subplot1,[0 Ncol-1]);
    ylim(subplot1,[0 100]);
    box(subplot1,'on');
    hold(subplot1,'all');
    plot(Data(count,:),'Parent',subplot1);
    title(headers(markers(counter,1),1));
    count=count+1;
    counter=counter+1;

    subplot2 = subplot(3,3,2,'Parent',figure1);
    xlim(subplot2,[0 Ncol-1]);
    ylim(subplot2,[0 100]);
    box(subplot2,'on');
    hold(subplot2,'all');
    plot(Data(count,:),'Parent',subplot2);
    title(headers(markers(counter,1),1));
```

```
count=count+1;
counter=counter+1;

subplot3 = subplot(3,3,3,'Parent',figure1);
xlim(subplot3,[0 Ncol-1]);
ylim(subplot3,[0 100]);
box(subplot3,'on');
hold(subplot3,'all');
plot(Data(count,:),'Parent',subplot3);
title(headers(markers(counter,1),1));
count=count+1;
counter=counter+1;

subplot4 = subplot(3,3,4,'Parent',figure1);
xlim(subplot4,[0 Ncol-1]);
ylim(subplot4,[0 100]);
box(subplot4,'on');
hold(subplot4,'all');
plot(Data(count,:),'Parent',subplot4);
title(headers(markers(counter,1),1));
count=count+1;
counter=counter+1;

subplot5 = subplot(3,3,5,'Parent',figure1);
xlim(subplot5,[0 Ncol-1]);
ylim(subplot5,[0 100]);
box(subplot5,'on');
hold(subplot5,'all');
plot(Data(count,:),'Parent',subplot5);
title(headers(markers(counter,1),1));
count=count+1;
counter=counter+1;

subplot6 = subplot(3,3,6,'Parent',figure1);
xlim(subplot6,[0 Ncol-1]);
ylim(subplot6,[0 100]);
box(subplot6,'on');
hold(subplot6,'all');
plot(Data(count,:),'Parent',subplot6);
title(headers(markers(counter,1),1));
count=count+1;
counter=counter+1;

subplot7 = subplot(3,3,7,'Parent',figure1);
xlim(subplot7,[0 Ncol-1]);
ylim(subplot7,[0 100]);
box(subplot7,'on');
hold(subplot7,'all');
plot(Data(count,:),'Parent',subplot7);
title(headers(markers(counter,1),1));
count=count+1;
counter=counter+1;

subplot8 = subplot(3,3,8,'Parent',figure1);
xlim(subplot8,[0 Ncol-1]);
ylim(subplot8,[0 100]);
box(subplot8,'on');
hold(subplot8,'all');
plot(Data(count,:),'Parent',subplot8);
title(headers(markers(counter,1),1));
count=count+1;
counter=counter+1;
```

```
    subplot9 = subplot(3,3,9,'Parent',figure1);
    xlim(subplot9,[0 Ncol-1]);
    ylim(subplot9,[0 100]);
    box(subplot9,'on');
    hold(subplot9,'all');
    plot(Data(count,:),'Parent',subplot9);
    title(headers(markers(counter,1),1));
    count=count+1;
    counter=counter+1;
    saveas(figure1,[folder,'\',folder,num2str(fignum),'.fig']);
 end;


function [output]=sortcol(input, col_to_be_sorted_by)
%sorts all the data by a column determined by the user
%called by raidlog function
%Written by David Pontin, Lincoln University 5/9/09 v1.0

output=[];
sz=size(input);
Nrow=sz(1);
[T O]=sort(input(:,col_to_be_sorted_by),'descend');
for rownumb=1:Nrow,
    row=input(O(rownumb,1),:);
    output=[output; row];
end;


function [output]=seqarray(size)
%creates a marker column with each marker being one greater than
the       %previous
%called by raidlog function
%Written by David Pontin, Lincoln University 5/9/09 v1.0

output=[];
start=0;
for rownumb=1:size,
    start=start+1;
    output=[output;start];
end;
```