University of South Wales
2059207

Bound by
Abbey
Bookbinding Co.
116 Cathays Terrace, Cardiff CF24 4HY
South Wales, U.K. Tel: (029) 2039 5882
www.bookbindersuk.com

University of Glamorgan
Prifysgol Morgannwg

Neural Network Techniques for the Identification and Classification of Marine Phytoplankton from Flow Cytometric Data

Luan Marie Al-Haddad

School of Computing
University of Glamorgan
Pontypridd

A thesis submitted in partial fulfilment of the requirements of the University of Glamorgan for the degree of Doctor of Philosophy.

This research programme was carried out in collaboration with Plymouth Marine Laboratory

2001

Table of Contents

Table of Contents i

Table of Figures vi

Table of Tables xiii

Acknowledgements xv

Declaration xvi

Abstract xvii

1 Introduction **1**
1.1 Phytoplankton and Their Importance 1
1.2 Major Phytoplankton Classes 1
1.3 Phtyoplankton Analysis 3
1.4 Flow Cytometry 4
1.4.1 Marine Analytical Flow Cytometry 4
1.4.2 Becton-Dickinson Flow Cytometer 5
1.4.3 Characteristics of Flow Cytometry Data 7
1.5 Plankton Reactivity in the Marine Environment-PRiME 9
1.6 Pattern Recognition 10
1.6.1 Statistical Pattern Recognition Methods 10
1.6.2 Neural Computing 13
1.6.2.1 The Biological Pattern Recognition System 13
1.6.2.2 Artificial Neural Networks 14
1.6.2.3 Neural Network Characteristics 15
1.7 Flow Cytometry and Neural Networks 18
1.8 Aims and Objectives 20

2 Supervised Neural Networks – MLP or RBF? **21**
2.1 Introduction 21
2.2 Multi-Layer Perceptron 21
2.2.1 Architecture 21
2.2.2 Algorithm 21
2.2.2.1 Forward Pass 22
2.2.2.2 Backward Pass – Update 25
2.2.3 Network Initialisation 26
2.2.3.1 Hidden Nodes 26
2.2.3.2 Weights 27
2.2.3.3 The Learning Parameter : η 27
2.2.3.4 Momentum : α 27
2.3 Radial Basis Functions and Interpolation 28
2.4 Radial Basis Functions as Networks 29
2.4.1 Architecture 29
2.4.2 Learning Algorithm 31

2.4.2.1 Location of the Basis Functions 31
2.4.2.2 Form of the Basis function 32
2.4.2.3 Output Layer Weights 37
2.5 Paradigm Summary 38
2.6 Data Collection, Preparation and Pre-processing 39
2.7 Experimental Procedure 42
2.7.1 Training and Testing Files 42
2.7.2 MLP training 42
2.7.3 RBF training 45
2.7.4 Testing Procedure 45
2.7.4.1 Probability Matrices 45
2.7.4.2 Rejection of Unknowns 46
2.8 Results 46
2.9 Discussion 51
2.10 Conclusion 54

3 Multi-Class RBF Network for Phytoplankton Analysis 55
3.1 Introduction 55
3.2 Training Set Size 55
3.3 Imbalanced Training Sets 56
3.4 Experimental Procedure 57
3.4.1 Balanced Event Numbers 57
3.4.2 Imbalanced Event Numbers 58
3.4.3 Compensation for Imbalanced Event Numbers 61
3.5 Results 61
3.5.1 Balanced Event Numbers 61
3.5.2 Imbalanced Event Numbers 64
3.5.3 Compensation for Imbalanced Event Numbers 64
3.6 Discussion 70
3.6.1 Balanced Event Numbers 70
3.6.2 Imbalanced Event Numbers 70
3.6.3 Compensation for Imbalanced Event Numbers 72
3.7 Summary 72
3.8 Analysis of 62 Phytoplankton Species 73
3.8.1 Experimental Procedure 74
3.8.1.1 Identification to Taxonomic Group Level 74
3.8.1.2 Comparison of Distance Metrics with Large Data Sets 74
3.8.1.3 Number of Hidden Layer Nodes 74
3.8.1.4 Principal Component Analysis 74
3.8.1.5 Species Combinations 75
3.8.2 Results 77
3.8.2.1 Identification to Taxonomic Group Level 77
3.8.2.2 Comparison of Distance Metrics with large data sets 77
3.8.2.3 Number of Hidden Layer Nodes 80
3.8.2.4 Principal Component Analysis 80
3.8.2.5 Species Combinations 80
3.8.3 Discussion 80
3.8.3.1 Identification to Taxonomic Group Level 80

3.8.3.2 Comparison of Distance Metrics with large data sets 82
3.8.3.3 Number of Hidden Layer Nodes 84
3.8.3.4 Principal Component Analysis 84
3.8.3.5 Species Combinations 84
3.8.4 Rejection of Species as Unknown 87
3.8.4.1 Experimental Procedure 87
3.8.4.2 Results 87
3.8.4.3 Discussion 90
3.9 Conclusion 92

4 Alternative Multiple Neural Network Architecture 93
4.1 Introduction 93
4.2 Restraints of the Original Multi-Class Network Architecture 93
4.3 Combinatorial Neural Networks 94
4.4 Multiple Neural Network Architecture 96
4.4.1 Maximum Valued Output 99
4.4.2 RBF Network Decision 99
4.4.3 Bayesian A Posteriori Probabilities 100
4.5 Background Class – Content and Quantity 100
4.6 Experimental Procedure 101
4.6.1 Single Species Training and Testing Files 101
4.6.2 Single Species Network Training 103
4.6.3 Maximum Valued Output 103
4.6.4 RBF Network Decision 103
4.6.4.1 Training and Test Files 103
4.6.4.2 Network Training 104
4.7 Results 104
4.7.1 Single Species Networks 104
4.7.2 Maximum Valued Output 104
4.7.3 RBF Network Decision 109
4.8 Discussion 109
4.8.1 Single Species Networks 109
4.8.2 Output Combinations 111
4.9 Conclusion 112
4.10 Multiple Network Architecture - Evaluation 113
4.10.1 Species Combinations 113
4.10.2 Rejection of Unknowns 113
4.10.3 Dynamic Selection of Species 116
4.10.4 Addition of Novel Species 117
4.11 Results 117
4.11.1 Species Combinations 117
4.11.2 Rejection of Unknowns 119
4.11.3 Dynamic Selection of Species 119
4.11.4 Addition of Novel Species 122
4.12 Discussion 122
4.12.1 Species Combinations 122
4.12.2 Rejection of Unknowns 125
4.12.3 Dynamic Selection of Species 125

4.12.4 Addition of Novel Species 126
4.13 Conclusion 127

5 Classification and Unsupervised Neural Networks 130
5.1 Introduction 130
5.2 Data Classification 130
5.3 Flow Cytometric Classification and Neural Networks 133
5.4 Cluster Definition Problem 133
5.5 Kohonen's Self Organising Map 134
5.5.1 Architecture 134
5.5.2 Algorithm 134
5.5.3 Initialisation 138
5.5.3.1 Update region 138
5.5.3.2 Map Size and Dimension 138
5.6 Boundary and Cluster Detection on the Kohonen SOM 140
5.7 Boundary Detection Methods 141
5.7.1 Visualisation of Hyper-Dimensional Euclidean Distances 141
5.7.1.1 Borders between Nodes 141
5.7.1.2 Grey Scale Representation of Euclideans 141
5.7.1.3 Edge Detection 142
5.7.2 Redundant Nodes 142
5.7.3 Proportional Node Responses 143
5.7.4 Visual Population Density 143
5.7.5 Agglomerative Clustering 144
5.7.6 Decomposition 144
5.7.7 Notes on Boundary Detection Methods 145
5.8 Experimental Procedure 146
5.8.1 Training Files 146
5.8.2 Kohonen Network Training 146
5.9 Results 153
5.9.1 Visualisation of Hyper-Dimensional Distances 153
5.9.1.1 Borders between Nodes 153
5.9.1.2 Grey Scale Representation of Euclideans 167
5.9.1.3 Edge Detection 167
5.9.2 Redundant Nodes 173
5.9.3 Proportional Node Responses 173
5.9.4 Visual Population Density 180
5.9.5 Agglomerative Clustering 187
5.9.6 Decomposition 187
5.10 Discussion 198
5.10.1 Visualisation of Hyper-Dimensional Distances 198
5.10.1.1 Borders between Nodes 198
5.10.1.2 Grey Scale Representation of Euclideans 199
5.10.2 Redundant Nodes 199
5.10.3 Proportional Node Responses 200
5.10.4 Visual Population Density 201
5.10.5 Agglomerative Clustering 202
5.10.6 Decomposition 203

5.11 Summary 203
5.12 Conclusion 206

6 Biological Variation 207
6.1 Introduction 207
6.2 Illumination Variance 207
6.2.1 Experimental Procedure 207
6.2.1.1 Multi-Class Network Architecture 207
6.2.1.2 Multiple Network Architecture 208
6.2.2 Results 210
6.2.2.1 Multi-Class Network Architecture 210
6.2.2.2 Multiple Network Architecture 212
6.2.3 Discussion 212
6.3 Laboratory Cultured Mix Data 215
6.3.1 Experimental Procedure 215
6.3.2 Results 218
6.3.3 Discussion 218
6.4 Field Tests 223
6.4.1 PRiME Cruise 223
6.4.1.1 Procedures and Results 223
6.4.2 Plymouths Coastal Waters 226
6.4.2.1 Experimental Procedure 227
6.4.2.2 Results 229
6.4.2.3 Discussion 229
6.5 Conclusion 235

7 Synthesis 236
7.1 Introduction 236
7.2 Paradigm Selection 236
7.3 Training Set Size 236
7.4 Multi-class RBF Network 237
7.5 Multiple Network Architecture 238
7.6 Morphology versus Flow Cytometry 238
7.7 Boundary Recognition on the SOM 239
7.8 Culturing Variation and Field Trials 240
7.9 Conclusions and Future Work 240

References 242

Appendix 1 - Phytoplankton Characteristics 252

Appendix 2 – Biological Glossary 255

Appendix 3 - Kohonen's Self Organising Map 256

Appendix 4 – Publications and Posters 259

Table of Figures

		Page
Figure 1.1	Morphological illustration of a selection of phytoplankton species.	2
Figure 1.2	Exploded schematic diagram of the Becton-Dickinson FACSort flow cytometer used for this research.	6
Figure 1.3	Scatter plots indicating the lack of consistency between the flow cytometric signatures of some species and their morphometric groupings	8
Figure 1.4	Architecture of the feed-forward Multi-Layer Perceptron network.	16
Figure 2.1	Transfer functions of the MLP network (a) Sigmoid (b) Hyperbolic tangent.	24
Figure 2.2	Architecture of the Radial Basis Function network.	30
Figure 2.3	Gaussian basis function for the RBF network.	33
Figure 2.4	Spatial extent and shape of the basis functions with varying forms of the variance-covariance matrix.	36
Figure 2.5	Two-dimensional scatter plots of the two data sets constructed to compare the performance of the MLP and RBF networks.	43
Figure 2.6	Percentage of species rejected from each data set when a threshold is imposed upon the winning node for both network architectures.	50
Figure 2.7	Percentage of species rejected from each data set, when a threshold is imposed upon the difference between the winning node and the second highest node, for both network architectures.	50
Figure 2.8	Hypothetical illustration of an imaginary data point in relation to boundaries constructed by the RBF and MLP network architectures.	53
Figure 3.1	Overall identification and individual identification of some species as node and event numbers are increased for three balanced data sets.	63
Figure 3.2	Overall correct identification for three imbalanced data sets as event numbers are increased.	65
Figure 3.3	Individual identification of extreme species in both halves of each of the three imbalanced data sets.	66
Figure 3.4	Performance identification, after adjustment, for each half of the three imbalanced data sets.	69
Figure 3.5	Confusion dendrogram of the misidentification matrix for the 62 species analysed by the RBF network.	76
Figure 3.6	Scatter plots depicting two species exhibiting non-isotropic and isotropic spatial orientation respectively.	83

Figure 3.7	Overall percentage of correct identification, and percentage of species rejected by imposing three different thresholds upon the parameters of the optimum RBF network, trained on all 62 species.	89
Figure 3.8	Overall percentage of correct identification, confidence of identification and percentage of species rejected by imposing a threshold upon the maximum hidden layer node output of the optimum RBF network.	91
Figure 4.1	Schematic of the alternative multiple network architecture presented in this thesis.	97
Figure 4.2	Confusion dendrogram of the misidentification matrix for the 62 species analysed by the multiple network architecture.	114
Figure 4.3	Overall percentage of correct identification, confidence of identification and percentage of species rejected by imposing thresholds upon the parameters of the optimum multiple network.	120
Figure 4.4	Training times for the original multi-class network architecture and the alternative multiple network architecture employing both Mahalanobis and Euclidean distance metrics.	121
Figure 4.5	Overall percentage correct identification by both network architectures, employing each distance metric for a separable and overlapping data set.	121
Figure 5.1	Two dimensional plot illustrating data sets exhibiting elliptical clusters and the problems with a statistical analysis technique that uses a distance metric for clustering.	132
Figure 5.2	Schematic of Kohonen's Self Organising Map in two dimensions.	135
Figure 5.3	Mexican Hat Function. Update region of the Kohonen layer of the SOM showing excitatory and inhibitory regions.	137
Figure 5.4	Progression of a Kohonen map learning to map 2 dimensional data arranged in a circular area (a) 6 by 6 map (b) one dimensional map.	139
Figure 5.5	Scatter plots depicting 2-d views of two sets of optical parameters for the eleven species data set, constructed to assess performance of the boundary detection methods for the SOM.	147
Figure 5.6	Scatter plots depicting 2-d views of two sets of parameters for the two class data set.	148
Figure 5.7	Scatter plots depicting 2-d views of two sets of parameters for the thirty species data set.	150
Figure 5.8	Scatter plots depicting 2-d views of two sets of parameters for the sixty species data set.	152
Figure 5.9	Two dimensional plot depicting 8 by 8 SOM produced for the generated two group data set after training.	155

Figure 5.10	Two dimensional plot depicting 24 by 24 SOM produced for the generated two group data set after training.	155
Figure 5.11	Two dimensional 8 by 8 grid produced for the generated two group data set showing borders between allocated nodes.	156
Figure 5.12	Two dimensional 24 by 24 grid produced for the generated two group data set, showing borders between allocated nodes.	156
Figure 5.13	Two dimensional 8 by 8 grid produced for the generated two group data set, showing thresholds imposed on borders between allocated nodes.	158
Figure 5.14	Two dimensional 24 by 24 grid produced for the generated two group data set showing thresholds imposed on borders between allocated nodes.	158
Figure 5.15	Two dimensional plot depicting 16 by 16 SOM produced for the eleven species data set after training.	159
Figure 5.16	Two dimensional 22 by 22 grid produced for the eleven species data set showing borders between allocated nodes.	159
Figure 5.17	Two dimensional 22 by 22 grid produced for the eleven species data set, showing thresholds imposed on borders between allocated nodes.	160
Figure 5.18	Two dimensional plot depicting 24 by 24 SOM produced for the thirty species data set after training.	160
Figure 5.19	Two dimensional plot depicting 30 by 30 SOM produced for the thirty species data set after training.	161
Figure 5.20	Two dimensional 24 by 24 grid produced for the thirty species data set, showing borders between allocated nodes.	161
Figure 5.21	Two dimensional 24 by 24 grid produced for the thirty species data set, showing thresholds imposed on borders between allocated nodes.	162
Figure 5.22	Two dimensional 30 by 30 grid produced for the thirty species data set, showing borders between allocated nodes.	162
Figure 5.23	Two dimensional 30 by 30 grid produced for the thirty species data set, showing thresholds imposed on borders between allocated nodes.	163
Figure 5.24	Two dimensional plot depicting 22 by 22 SOM produced for the sixty species data set after training.	163
Figure 5.25	Two dimensional plot depicting 30 by 30 SOM produced for the sixty species data set after training.	164
Figure 5.26	Two dimensional 22 by 22 grid produced for the sixty species data set showing borders between allocated nodes.	164

Figure 5.27	Two dimensional 22 by 22 grid produced for the sixty species data set showing thresholds imposed on borders between allocated nodes.	165
Figure 5.28	Two dimensional 30 by 30 grid produced for the sixty species data set showing borders between allocated nodes.	165
Figure 5.29	Two dimensional 30 by 30 grid produced for the sixty species data set showing thresholds imposed on borders between allocated nodes.	166
Figure 5.30	22 by22 map for the eleven specie data set after grey scale imaging of average Euclidean distances and Sobel application.	168
Figure 5.31	22 by22 map for the eleven specie data set after grey scale imaging of four corner Euclidean distances and Sobel application.	168
Figure 5.32	24 by24 map for the thirty species data set after grey scale imaging of average Euclidean distances and Sobel application.	169
Figure 5.33	24 by24 map for the thirty specie data set after grey scale imaging of four corner Euclidean distances and Sobel application.	169
Figure 5.34	30 by30 map for the thirty species data set after grey scale imaging of average Euclidean distances and Sobel application.	170
Figure 5.35	30 by 30 map for the thirty specie data set after grey scale imaging of four corner Euclidean distances and Sobel application.	170
Figure 5.36	22 by22 map for the sixty species data set after grey scale imaging of average Euclidean distances and Sobel application.	171
Figure 5.37	22 by 22 map for the sixty specie data set after grey scale imaging of four corner Euclidean distances and Sobel application.	171
Figure 5.38	30 by 30 map for the sixty species data set after grey scale imaging of average Euclidean distances and Sobel application.	172
Figure 5.39	30 by 30 map for the sixty specie data set after grey scale imaging of four corner Euclidean distances and Sobel application.	172
Figure 5.40	Groupings produced by the proportional node response method on the 8 by 8 SOM for the two class data set.	175
Figure 5.41	Groupings produced by the proportional node response method on the 24 by 24 SOM for the two class data set.	175
Figure 5.42	Groupings produced by the proportional node response method on the 22 by 22 SOM for the eleven species data set.	176
Figure 5.43	Groupings produced by the proportional node response method on the 24 by 24 SOM for the thirty species data set.	176
Figure 5.44	Groupings produced by the proportional node response method on the 24 by 24 SOM for the thirty species data set. Alternative threshold.	177

Figure 5.45	Groupings produced by the proportional node response method on the 30 by 30 SOM for the thirty species data set.	177
Figure 5.46	Groupings produced by the proportional node response method on the 30 by 30 SOM for the thirty species data set. Alternative threshold.	178
Figure 5.47	Groupings produced by the proportional node response method on the 22 by 22 SOM for the sixty species data set.	178
Figure 5.48	Groupings produced by the proportional node response method on the 22 by 22 SOM for the sixty species data set. Alternative threshold.	179
Figure 5.49	Groupings produced by the visual population density method on the 8 by 8 SOM for the two class data set.	181
Figure 5.50	Groupings produced by the visual population density method on the 22 by 22 SOM for the eleven species data set.	181
Figure 5.51	Groupings produced by the visual population density method on the 24 by 24 SOM for the two class data set, with varying centres.	182
Figure 5.52	Groupings produced by the visual population density method on the 24 by 24 SOM for the thirty species data set.	183
Figure 5.53	Groupings produced by the visual population density method on the 30 by 30 SOM for the thirty species data set.	183
Figure 5.54	Groupings produced by the visual population density method on the 22 by 22 SOM for the sixty species data set.	184
Figure 5.55	Groupings produced by the visual population density method on the 22 by 22 SOM for the sixty species data set. Alternate centre .	184
Figure 5.56	Groupings produced by the visual population density method on the 30 by 30 SOM for the sixty species data set.	185
Figure 5.57	Groupings produced by the visual population density method on the 30 by 30 SOM for the sixty species data set. Alternative centres.	185
Figure 5.58	Groupings produced by the visual population density method on the 30 by 30 SOM for the sixty species data set. Alternative centres.	186
Figure 5.59	Groupings produced by the agglomerative method on the 8 by 8 SOM for the two class data set. 20 groups	188
Figure 5.60	Groupings produced by the agglomerative method on the 24 by 24 SOM for the two class data set. 20 groups	188
Figure 5.61	Groupings produced by the agglomerative method on the 22 by 22 SOM for the eleven species data set. 100 groups	189
Figure 5.62	Groupings produced by the agglomerative method on the 22 by 22 SOM for the eleven species data set. 150 groups	189

Figure 5.63	Groupings produced by the agglomerative method on the 22 by 22 SOM for the eleven species data set. 200 groups	190
Figure 5.64	Groupings produced by the agglomerative method on the 24 by 24 SOM for the thirty species data set. 100 groups	190
Figure 5.65	Groupings produced by the agglomerative method on the 24 by 24 SOM for the thirty species data set. 150 groups	191
Figure 5.66	Groupings produced by the agglomerative method on the 24 by 24 SOM for the thirty species data set. 200 groups	191
Figure 5.67	Groupings produced by the agglomerative method on the 30 by 30 SOM for the thirty species data set. 100 groups	192
Figure 5.68	Groupings produced by the agglomerative method on the 30 by 30 SOM for the thirty species data set. 150 groups	192
Figure 5.69	Groupings produced by the agglomerative method on the 30 by 30 SOM for the thirty species data set. 200 groups	193
Figure 5.70	Groupings produced by the agglomerative method on the 22 by 22 SOM for the sixty species data set. 100 groups	193
Figure 5.71	Groupings produced by the agglomerative method on the 22 by 22 SOM for the sixty species data set. 150 groups	194
Figure 5.72	Groupings produced by the agglomerative method on the 22 by 22 SOM for the sixty species data set 200 groups	194
Figure 5.73	Groupings produced by the agglomerative method on the 30 by 30 SOM for the sixty species data set. 100 groups	195
Figure 5.74	Groupings produced by the agglomerative method on the 30 by 30 SOM for the sixty species data set. 150 groups	195
Figure 5.75	Groupings produced by the agglomerative method on the 30 by 30 SOM for the sixty species data set. 200 groups	196
Figure 5.76	Chart illustrating the percentage of the five taxonomic groups allocated, after training, to the five nodes of a 1 by 5 Kohonen SOM.	196
Figure 5.77	Primary node allocation to eleven nodes of a 1 by 11 Kohonen SOM, trained on the eleven species data set.	197
Figure 5.78	Chart illustrating the percentage of the eleven species allocated, after training, to the eleven nodes of a 1 by 11 Kohonen SOM.	197
Figure 6.1	Orange fluorescence histograms for two species cultured under different illumination intensities.	214
Figure 6.2	Kohonen grid produced from the 27 group data set showing borders remaining after threshold imposition.	217

Figure 6.3	Charts depicting analysis results of the seven laboratory cultured mixtures.	219
Figure 6.4	Transect of the 1996 PRiME cruise.	224
Figure 6.5	Region gated for analysis assumed to be *Coccolithus pelagicus*.	224
Figure 6.6	Two dimensional scatter plots of a sample collected during the PRiME cruise.	225
Figure 6.7	Two dimensional scatter plots depicting red fluorescence against side scatter for the second sample collected during the PRiME cruise.	225
Figure 6.8	Two dimensional scatter plots of a sample collected during field studies at waters off the coast of Plymouth.	228
Figure 6.9	Percentage of species determined to be present in the field samples taken at the five stations on Day 1.	230
Figure 6.10	Percentage of species determined to be present in the field samples taken at the five stations on Day 2.	231
Figure 6.11	Percentage of species determined to be present in the Day 1 field samples after 2 days incubation in the refrigerator	232
Figure 6.12	Orange fluorescence distributions for *Micromonas pusilla* under different culturing conditions	233
Figure 6.13	Two dimensional scatter plots depicting red fluorescence against side scatter for *Micromonas pusilla* under different culturing conditions	233

Table of Tables

		Page
Table 2.1	Database of 62 species cultured under PRiME 1 conditions from five taxonomic groups	40
Table 2.2	Two sets of species used to construct training and testing files for analysis of the MLP and RBF network architectures.	43
Table 2.3	Species used as unknown to test rejection performance of the MLP and RBF networks.	44
Table 2.4	Overall probability of correct identification of two data sets by each paradigm, as hidden layer nodes are increased.	48
Table 2.5	Individual identification of species in the Mixed data set by the optimum RBF and MLP networks.	49
Table 2.6	Individual identification of species in the Dinoflagellate data set by the optimum RBF and MLP networks.	49
Table 3.1	Species membership of the balanced and imbalanced data sets	59
Table 3.2	Event numbers for the three constructed data sets, as well as respective training data class probabilities used to assess performance of the RBF network after Bayesian adjustment.	62
Table 3.3	Misidentification matrix for an RBF network trained on flow cytometric data to identify to taxonomic group level.	78
Table 3.4	Mean identification and standard error of mean for individual species across 5 RBF networks employing Mahalanobis and Euclidean distance metrics trained on 62 phytoplankton species.	79
Table 3.5	Performance of networks employing both Mahalanobis and Euclidean distance metrics as hidden layer nodes are increased.	81
Table 3.6	Identification performance produced from combining species at various levels	81
Table 3.7	Correct identification and confidence of identification for individual species from the optimum RBF network.	85
Table 3.8	Eleven additional species selected to assess the RBF networks ability to reject novel classes when trained on the original 62 species.	88
Table 4.1	Event numbers for the 8 constructed sets of 62 single species network training files. Overall identification success for each set of single species networks as node numbers are increased in the class of interest.	102
Table 4.2	Percentage correct identification, mean and standard deviation of three sets of 62 single species networks.	105
Table 4.3	Results of decision processes for combining the trained single species network outputs.	107

Table 4.4	Correct identification and confidence of identification for individual species by the multiple network architecture, by both methods for combining outputs.	110
Table 4.5	Twelve species of phytoplankton used as unknowns, to assess the ability of the alternative multiple network to reject novel species.	115
Table 4.6	Overall identification and confidence of identification by the RBF decision network produced from combining species in various groups.	118
Table 4.7	Correct identification, mean and standard deviation for 12 single species networks trained to recognise an additional 12 novel species.	123
Table 4.8	Individual correct identification and confidence of identification for the original 62 species and the additional 12 novel species.	124
Table 4.9	Comparison of individual identification results by the optimum multi-class network and the optimum multiple network approach.	128
Table 5.1	Eleven species data set constructed to assess the performance of the boundary detection methods on the SOM.	147
Table 5.2	Thirty species data set constructed to assess the performance of the boundary detection methods on the SOM.	149
Table 5.3	Sixty species data set constructed to assess the performance of the boundary detection methods on the SOM.	151
Table 5.4	Euclidean threshold ranges and values used for border depiction between nodes, to illustrate the detection method for the four data sets.	157
Table 5.5	Threshold values selected for the four data sets constructed to assess the performance of the Proportional Node Response Method and the Visual Population Density Method.	174
Table 6.1	Event numbers for each of the 9 sets of 60 files constructed to train single species networks to assess variance in performance for different illumination intensities.	209
Table 6.2	Comparison of individual identification of species by the optimum original multi-class network and the optimum multiple network for the PRiME 2 data set (i.e. different illumination intensity).	211
Table 6.3	27 separately cultured species used to train networks of both architectures for identification of 7 laboratory cultured mixtures.	216
Table 6.4	Stations locations and depths for field sample analysis around Plymouths coastal areas.	228
Table A1.1	Some of the primary physical features used to identify phytoplankton cells by morphology.	252
Table A1.2	Characteristics of some of the phytoplankton classes used in this research	253

Acknowledgements

I would like to express my thanks to my supervisors at Glamorgan University, Colin W. Morris and Dr. Andrew Ware, and at Cardiff University, Prof. Lynne Boddy; to Dr. Glen Tarran and Dr. Peter Burkill of Plymouth Marine Laboratory, who provided the PRiME data used in this thesis; to Dr. Malcolm Wilkins of Cardiff University.

My personal thanks go to my family for their continual love, support and endless cups of tea, and to Tony for his love and patience.

This work was supported by the NERC grant number GT22/95/PRiME/2 (project number 3101A259). The additional 12 species used in chapters 3 and 4 to assess rejection of unknowns, were supplied courtesy of the MAST III project AIMS (contract number MAS3-CT97-0080).

Certificate of Research

This is to certify that, except where specific reference is made, the work described in this thesis is the result of the candidate. Neither this thesis, nor any part of it, has been presented, or is currently submitted, in candidature for any degree at any other University.

Signed  (Candidate)

Signed C W Morris (Director of Studies)

Date April 2001

Abstract

This thesis documents the research that has led to advances in the Artificial Neural Network (ANN) approach to analysing flow cytometric data from phytoplankton cells. The superiority of radial basis function networks (RBF) over multi-layer perceptron networks (MLP), for data of this nature, has been established, and analysis of 62 marine species of phytoplankton represents an advancement in the number of classes investigated.

The complexity and abundance of heterogeneous phytoplankton populations, renders an original multi-class network redundant each time a novel species is encountered. To encompass the additional species, the original multi-class network requires complete retraining, involving long optimisation procedures to be carried out by ANN scientists. An alternative multiple network approach presented (and compared to the multi-class network), allows identification of the expanse of real world data sets and the easy addition of new species. The structure comprises a number of pre-trained single species networks as the front end to a combinatorial decision process for determining species identification. The simplicity of the architecture, and of the subsequent data produced by the technique, allows scientists unfamiliar with ANNs to dynamically alter the species of interest as required, without the need for complete re-training.

Kohonens Self Organising Map (SOM), capable of discovering its own classification scheme, indicated areas of discrepancy between flow cytometric signatures of some species and their respective morphological groupings. In an attempt to improve identification to taxonomic group or genus level by supervised networks, class labels more reflective of flow cytometric signatures must be introduced. Methods for boundary recognition and cluster distinction in the output space of the SOM have been investigated, directed towards the possibility of an alternative flow cytometric structuring system.

Performance of the alternative multiple network approach was comparable to that of the original multi-class network when identifying data from various environmental and laboratory culturing conditions. Improved generalisation can be achieved through employment of optical characteristics more representative of those found in nature.

1. Introduction

1.1 Phytoplankton and Their Importance

Plankton is the collective name given to the microscopic animals and plants found at all depths in the Earth's oceans and fresh water regions. Phytoplankton, the plant constituent of plankton, exist either as independent cells or in a colonial state. The main components of fresh water areas are the Diatoms, Dinoflagellates, Desmids and Cyanobacteria (also a marine inhabitant), whilst primarily Diatoms and Dinoflagellates populate the world's seas (Harris, 1986) (Fig. 1.1).

These organisms are photoautrophic and fuel marine food chains throughout the planet. They supply protein, carbohydrates, fats, vitamins and mineral salts to the primary consumers, and the oxygen liberated by photosynthesis is a vital part of the life support system of the planet (Boney, 1989). They are also known to have a more subtle effect on the earth's climate, impacting on the exchange of CO_2 between atmosphere and sea. To a lesser extent they contribute to atmospheric cooling through the production of dimethyl sulphide, which once oxidised forms sulphate aerosols (Wigley, 1994). Some species form toxic blooms, which cause death to many aquatic vertebrates and invertebrates in the affected area, not only by oxygen depletion but by toxin production. Cyanobacteria, being one of the major culprits of toxin emission, are capable of causing severe illness in humans as well as the aquatic life. It is therefore imperative that in order to understand the functioning of the world's aquatic systems, the community structure and function of these primary producers be fully understood. The work reported in this thesis forms part of a major study programme in this area (Section 1.5).

1.2 Major Phytoplankton Classes

Characterising phytoplankton is an exceptionally difficult process. Variation within a single species can be considerably complex due to a number of factors (Section 1.4.3). Morphological characteristics (Appendix 1, Table A1.1) have historically been used for distinguishing between species and taxonomic groups, producing a number of major phytoplankton classes, some of which are listed here:

- *Bacillariophyceae* (Diatoms)
- *Dinophyceae* (Dinoflagellates)

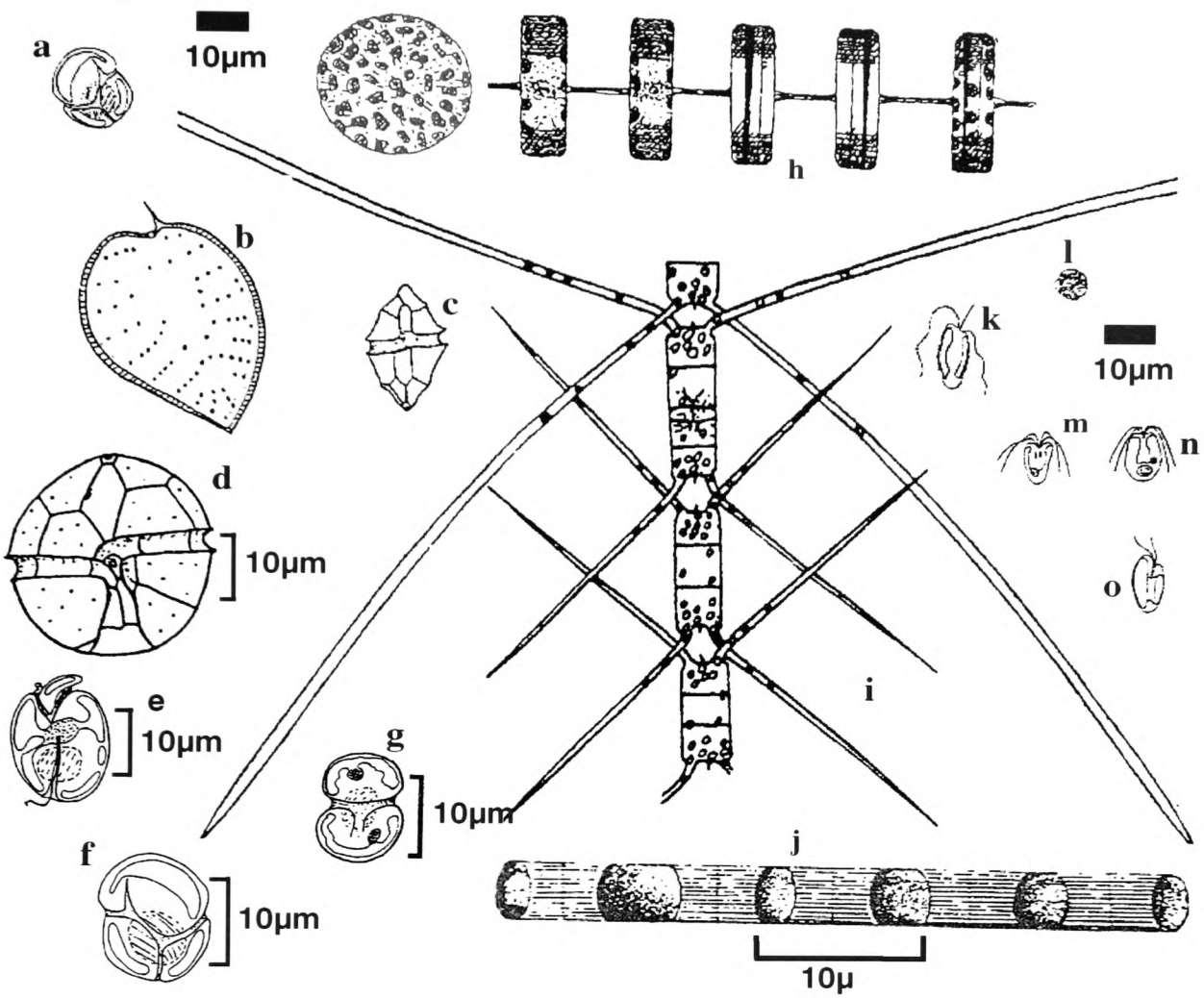


Figure 1.1 Morphological illustration of a selection of phytoplankton species. Dinoflagellates (a) *Gymnodinium micrum* (b) *Prorocentrum micans* (c) *Heterocapsa triquetra* (d) *Alexandrium tamarense* (e) *Amphidinium carterae* (f) *Gymnodinium micrum* (g) *Aureodinium pigmentosum.* Diatoms (h) *Thalassiosira* sp. (i) *Chaetoceros* sp. (j) *Skeletonema costatum.* Prymnesiomonads (k) *Prymnesium parvum* (l) *Emiliania huxleyi.* Flagellates (m) *Pyramimonas* sp. (n) *Tetraselmis* sp. Cryptomonads (o) *Chroomonas* sp. (Supplied by Dr. G.A. Tarran of Plymouth Marine Laboratories).

- *Chlorophyceae* (Grass-green algae)
- *Prasinophyceae* (Green algae)
- *Euglenophyceae* (Euglenoid flagellates)
- *Chrysophyceae* (Golden-brown phytoflagellates)
- *Prymnesiophyceae* (Brown phytoflagellates)
- *Cryptophyceae* (Flagellate)
- *Cyanophyceae* (Cyanobacteria; blue-green algae)
- *Rhodophyceae* (Red algae)

Details of these classes are given in Table A1.2 (Appendix 1). It should be noted however, that the descriptions are generic, and genera and species within the classes vary considerably.

1.3 Phytoplankton Analysis

The most commonly used tool for quantitative and qualitative analysis of phytoplankton is microscopy. Once samples are harvested, staining techniques and fluorescence can be used to highlight certain cellular components under microscope analysis, allowing subsequent identification to be made by a highly trained taxonomist. Both light and electron microscopy are used in the analysis, providing information on external morphology (*e.g.* special structures) and sub-cellular structure respectively. Light microscopy is also used for living organisms, where a species can be identified by its swimming motion.

Other techniques include satellite remote sensing, which supplies a spatial and temporal monitoring of phytoplankton blooms, employing colour scanners to determine biomass through optical effects on pigment (Harris, R. 1987; Fukushima, 1993). Fluorometry is used to evaluate biomass on a smaller scale, where chlorophyll fluorescence is measured and converted using chlorophyll to carbon conversion factors to produce an evaluation. A more accurate procedure than fluorometry is chromatography (Drucker, 1987). This technique separates chemical substances by taking advantage of the rates at which they are absorbed by a stationary material from a moving stream of gas or liquid. Gas chromatography, which can detect the various pigments in a sample and quantify them, is commonly used. This method allows identification of certain species

or groups whose pigments are unique, producing a number of chemo-taxonomic markers.

From the techniques mentioned, microscopy provides the most definitive and complete analysis for identifying and quantifying a phytoplankton sample, however, it does have its limitations. The procedure is painstakingly long and tedious and cannot normally be achieved at sea. It requires exceptional expert knowledge and a clear set of distinguishable features. The study of individual cells covers areas such as intra-population variability, and is an important aspect of phytoplankton analysis (Burkill, 1987). When this is required of a cell from a less common species, microscopy is unsuitable, as it is unable to identify a single cell that may be obscured by a mass of organic and inorganic materials. Thus, automated cell analysers were introduced to marine biology (Trask *et al.*, 1982; Olson *et al.*, 1983; Yentsch *et al.*, 1983; see below).

1.4 Flow Cytometry

Flow cytometers are particle analysers that simultaneously record a set of optical measurements, providing information on the physical characteristics of individual cells. Purpose designed for the study of mammalian cells and cell immunology, there use in oceanography required a number of alterations to account for the many differences between mammalian and phytoplankton cells (Peeters *et al.*, 1989). In addition to variance in size and shape, mammalian cells are generally homogenous, whereas marine organisms are exceptionally heterogeneous; internal fluorescence is one of the primary areas of investigation in aquatic cells, a characteristic rarely studied in mammalian cells; medical flow cytometers are designed for controlled, stable laboratory use and are not robust or flexible enough to transport to sea (Burkill, 1987; Steen, 1991). In order to advance this obvious benefit to aquatic science, marine flow cytometers needed to be built to specification, to increase their operation in allowing for these requirements.

1.4.1 Marine Analytical Flow Cytometer

A number of the limitations of the first marine flow cytometers were addressed by the design and experimental implementation of the Optical Plankton Analyser (OPA) (Dubelaar *et al.*, 1989; Peeters *et al.*, 1989). Although the OPA allowed the diversity of phytoplankton size and shape to be considered, it was determined unsuitable for field use,

only achieving experimental status; advancing this was the European Optical Plankton Analyser (EurOPA) (Dubelaar *et al.*, 1994a; 1994b). This flow cytometer was designed specifically for on board use, with the ability to measure particles in the range of 1μm to 1000μm, at a rate of about 1 ml of sample per minute. It is capable of detecting species that exist as colonies and chains, without damage to the structure. The instrument has the facility for image capture and, through improved electronics, allows a greater number of optical measurements to be taken (Dubelaar *et al.*, 1994a).

1.4.2 Becton-Dickinson Flow Cytometer

The flow cytometer used in this research is the Becton-Dickinson FACSort (Fig. 1.2). The instrument is triggered on chlorophyll fluorescence for data acquisition, indicating the presence of photosynthetic cells. Laboratory cultured *Micromonas pusilla* (1-3μm) is used as the lower analysis detector. An electrolyte containing the sample travels through a flow cell, crossing the focus of a very intense 15mW argon ion laser beam. The laser projects monochromatic light at a wavelength of 488nm, which irradiates 10^3 cells sec^{-1}. As cells pass individually through the beam, light is scattered in various directions. A number of detectors and filters are placed at different positions to measure pulse peak height or integrated signals. These are stored as a multi-parameter set of data, providing the optical characteristics of the particular cell. The Becton-Dickinson records 7 optical parameters for each event (one pass of a cell through the laser beam):

- Forward light scatter (FSC) - gauges light scattered between 1 and 10 degrees and is used as a sizing parameter.

90^o light scatter is focussed by a light collection lens and then spectrally filtered to detect the remaining 6 parameters:

- Side scatter (SSC) - used to quantify cell granularity and complexity.
- Depolarised light scatter (FL1) - This is used to indicate the presence of a group of phytoplankton, the Coccolithophores, which have a covering of calcite plates that depolarises the vertically polarised beam into the horizontal plate.

Red fluorescence is an indicator of cellular chlorophyll, the photosynthetic pigment of phytoplankton, recorded as three measurements:

- FL3-Height - measures the intensity of chlorophyll fluorescence.

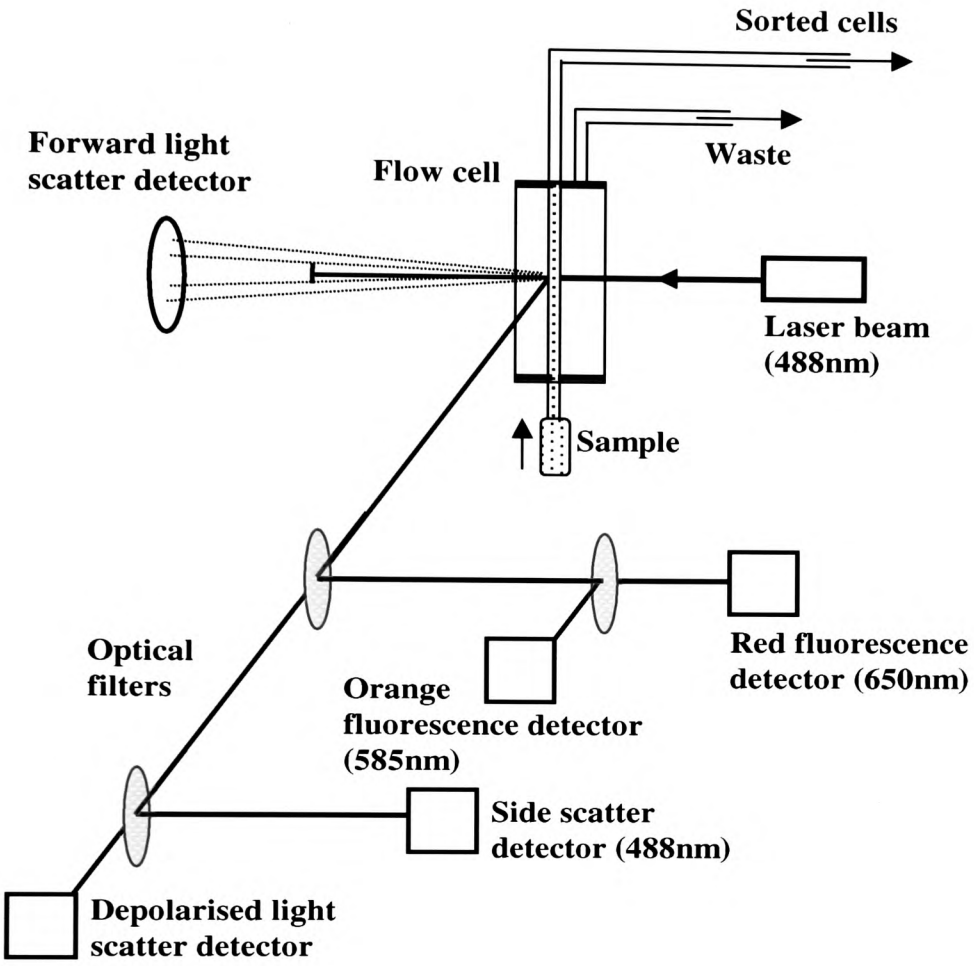


Figure 1.2 Exploded schematic diagram of a flow cytometer with cell sorting facility (the Becton-Dickinson FACSort)

- FL3-Area - measures total chlorophyll fluorescence for a single cell.
- FL3-Width - measures time of flight, the length of time a particle fluoresces as it passes through the beam.
- Orange fluorescence (FL2) - measures cellular accessory pigments, primarily phycoerythrin. This pigment is especially dominant in a particular taxonomic group of phytoplankton, the Cryptomonads.

The signals are measured for each event, amplified, digitised and stored in a *listmode* (binary) file, generally as logarithmic values. The flow cytometric signature for each cell is represented in the multi-parameter data. In addition to single cell analysis, this particular flow cytometer is capable of sorting, by electrostatically charging those droplets containing cells exhibiting the required parameters and deflecting them into sorting vessels. These can then be used for subsequent analysis or purification of cultures (Burkill, 1987; Burkill and Mantoura, 1990).

1.4.3 Characteristics of Flow Cytometry Data

With the exception of size and the detection of calcite plates, flow cytometry provides limited information about a cell's external physical features. Some of the defining characteristics that have placed species into their respective taxonomic groups are not detected by flow cytometry. Thus, not all species that are distinguishable through their morphometric characteristics, will be as easily separated by their flow cytometric signatures. For example, the placement of *Hemiselmis rufescens* (Fig. 1.3a) and *Chlorella salina* (Fig. 1.3b) into different taxonomic groups, is supported by the variance in their flow cytometric signatures. However, this is not the case with *Chlorella salina* and *Hemiselmis virescens* (Fig. 1.3c), where the two-dimensional plot of *Hemiselmis virescens* appears more characteristic of the Flagellate distribution than its own genus. This overlap and complex nature of flow cytometric signatures is a consequence of the inherent variability of phytoplankton data. Whether a sample is cultured or natural, individual cells from the same species will exhibit considerable biological variation, as well as the possibility of some multi-modal data distributions. In a field environment, additional factors contribute to the heterogeneous nature of phytoplankton (Morris *et al.*, 1994). A sample may contain cells at varying stages of growth or decay; differences may be

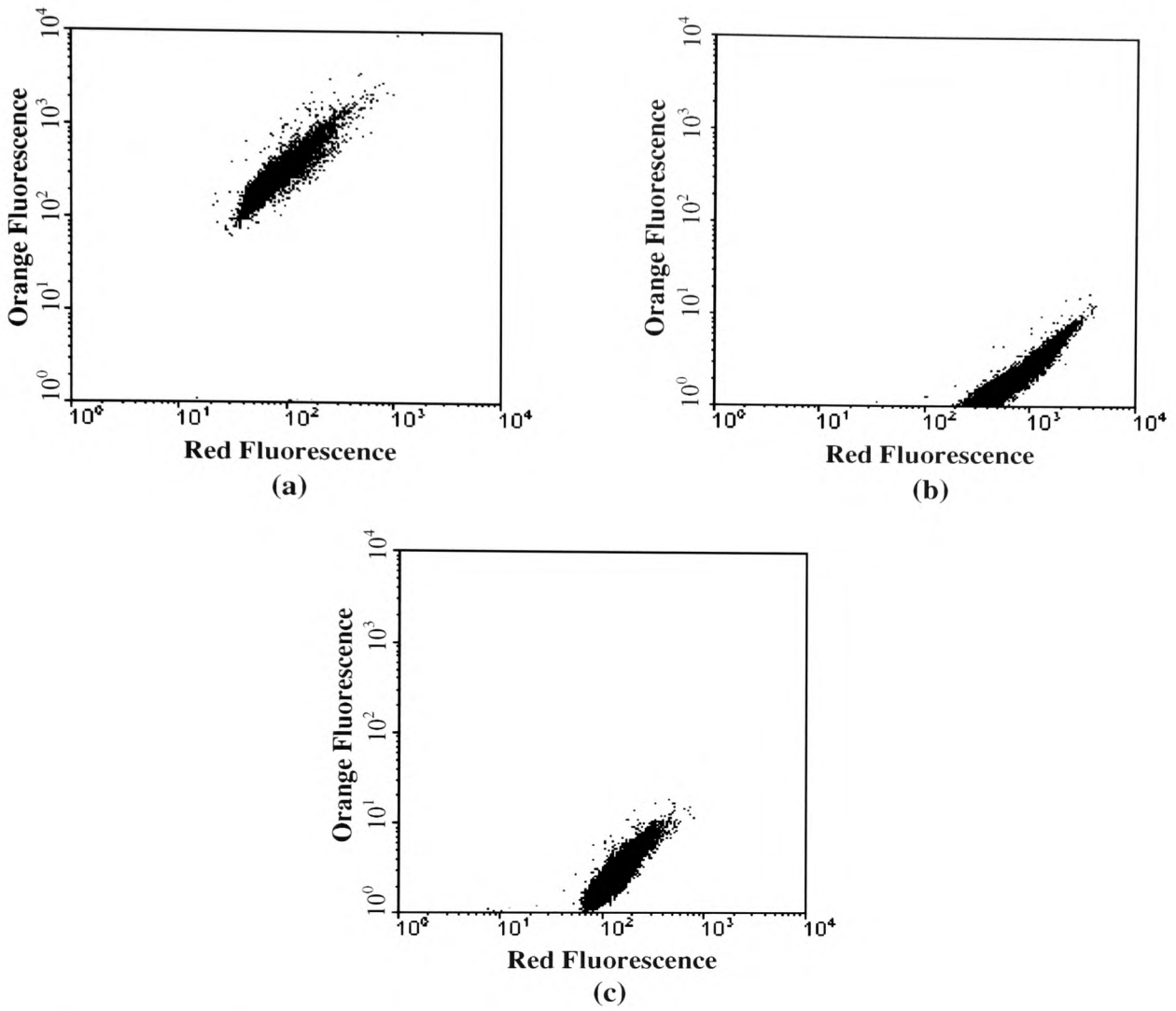


Figure 1.3 Two-dimensional scatter plots of Red Fluorescence (H) against Orange Fluorescence for (a) *Hemiselmis rufescens* (Cryptomonad), (b) *Chlorella salina* (Flagellate) and (c) *Hemiselmis virescens* (Cryptomonad). The morphometric placement of *Hemiselmis rufescens* and *Chlorella salina* into different groups is supported by their distinct flow cytometric signatures. This is not the case with *Chlorella salina* and *Hemiselmis virescens* where, in this instance, the latter appears more representative of the Flagellates distribution than the Cryptomonads.

attributed to environmental conditions (nutrient availability, light intensity, temperature, depth), and/or temporal and evolutionary factors; cell debris, zooplankton, bacteria and inorganic particles will add to the complexity of a sample (although chlorophyll content allows relatively easy exclusion of non-photosynthetic extraneous particles). These factors illustrate two primary requirements when considering identification through flow cytometry. Firstly, an adequate number of optical parameters must be considered in the analysis in order to achieve maximum separation of classes. However, the variables that make up the multi-parameter patterns are not independent of each other and therefore cannot be analysed individually. This dictates the requirement for a method of analysis which examines simultaneously all the variables that contribute to a particular cell (Demers *et al.*, 1992). Secondly, in order to cover as much of the range of biological variation as possible, a sufficient number of flow cytometric signatures, representative of each species, must be analysed. The need for such an abundance of variable data intensifies the extent of overlap between different species, as well as limiting the choice of algorithm to those that can cope with large data sets.

1.5 Plankton Reactivity in the Marine Environment – PRiME

The Biogeochemical Ocean Flux Study programme introduced the PRiME project in the early 1990s. At the start of 1995, a number of Oceanographic Research groups were brought together to lay the basis for comprehensive, mathematical models of the nature, distribution and interaction of ocean plankton, with a particular understanding of the contribution of these organisms to biogeochemical processes in the world's aquatic regions. Modelling employs both historical and empirical data generated during the course of the project. Varying techniques, old and new, including microscopy and acoustical engineering, have been employed in the programme to analyse areas such as grazing, normally containing a predator (*e.g.* Zooplankton), and competition experiments, addressing the survival of individual phytoplankton species in a fight for nutrients. The requirements of the PRiME project for the research reported here, was to further the coupling of Automated Flow Cytometry with Artificial Neural Networks, producing a rapid, portable and flexible multi-variate analysis technique, for identification and classification of large numbers of phytoplankton (Section 1.7 & 1.8).

1.6 Pattern Recognition

Pattern Recognition encompasses many areas of data interpretation. Its implementation is found in numerous fields such as, computer vision, seismic analysis, face and speech recognition, character recognition and medical diagnosis. Any application which, broadly speaking, involves the description, identification or classification of measurements, may be termed a pattern recognition problem. It may be categorised as an information reduction, information mapping or information labelling process (Schalkoff, 1992). It is based on recognising a set of states which provide information about a certain event, such as the pattern set of optical parameters which represent the flow cytometric signature of a phytoplankton cell. The primary approaches to pattern recognition fall into three main categories:

- Statistical pattern recognition considers class-conditioned probabilities or probability density functions (Section 1.6.1).
- Syntactic pattern recognition looks at the internal relationship between features in a pattern, rather than the features themselves (not documented here).
- Artificial Neural Networks are the most recent form of pattern recognition and are the main focus of this thesis (Section 1.6.2.2).

Although varying in internal algorithm, the structure of each system is relatively similar, employing either a classification, identification or descriptive procedure to produce a decision.

1.6.1 Statistical Pattern Recognition Methods

Multivariate statistical analysis is commonly used for large data sets, where P measurements are recorded simultaneously on each of N individuals, allowing expression of the data set as a data matrix:

$$\mathrm{X} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1P} \\ x_{21} & x_{22} & \cdots & x_{2P} \\ \vdots & \vdots & & \vdots \\ x_{N1} & x_{N2} & \cdots & x_{NP} \end{pmatrix}$$

where x_{21} denotes the 1st parameter recorded for the 2nd individual, and each row of the matrix can be considered as a vector representation of the N^{th} individual in P-dimensional

space. A number of classical methods of multivariate statistical analysis exist, all of which are well-explored and documented (Dillon & Goldstein, 1984; McLachlan, 1992; Krzanowski, 1993), some are briefly discussed here.

Cluster Analysis

Cluster analysis involves partitioning the data space into particular regions (sub-sets or clusters) and then assigning each pattern in the data set to one of the clusters. It is a data reduction technique, where residence in a particular sub-set is dependent upon similarities more common to the other members than to the remaining patterns being considered. This similarity is commonly assessed via a distance metric and requires an inference to '*What is similarity?*' and subsequently '*What is dissimilarity?*'. Once a similarity threshold has been established, the data must be partitioned based on the concluded measure, a process for which numerous methods exist (Dillon & Goldstein, 1984).

Discriminant Analysis – Bayesian Statistics

Discriminant analysis uses a set of independent variables to classify events into mutually exclusive and exhaustive groups. Bayesian statistics is concerned with the separation of two points in hyper-dimensional space, generally concluded by deciding how to partition the sample space. Considering a set of k classes of which pattern $\mathbf{x}$ belongs to class j, the analysis expresses the *a posteriori* probability $Pr(class\ j \mid \mathbf{x})$, *i.e.* the probability that the class is j given the event $\mathbf{x}$, in terms of the *a priori* probability of class j, $Pr(class\ j)$, and the class conditional probability or likelihood, $Pr(\mathbf{x}|class\ j)$:

$$Pr(class\ j \mid x) = \frac{Pr(x \mid class\ j)\ Pr(class\ j)}{Pr(x)}$$

The unconditional probability is denoted by the denominator and ensures that the posterior probabilities sum to one. It can be expressed in terms of the class conditional probability and the prior probabilities:

$$Pr(class\ j \mid x) = \frac{Pr(x \mid class\ j)\ Pr(class\ j)}{\sum_{j=1}^{k} Pr(x|class\ j)\,Pr(class\ j)}$$

The classifier then assigns x to the class with the highest *a posteriori* probability. Calculated *a posteriori* probabilities can subsequently be used for analysis of new data (Bishop, 1995).

Feature Selection

Feature selection is a technique used for reducing dimensionality by selecting only a subset of the inputs, disregarding any that have little or no influence on classification. The procedure requires defining a threshold criterion to select influential parameters (Tou and Gonzalez, 1974)

Principal Component Analysis

A commonly used approach, Principal Component Analysis (Jolliffe, 1986; Preisendorfer, 1988), transforms a set of vectors, **x**, in a p-dimensional space, onto a set of vectors, **z**, in a m-dimensional space (where m<p), with minimum loss to variance of the original set. Generally, 2 or 3 principal components are extracted, where each is a linear combination of the input parameters and weights w_{II}:

$$PC_C = w_{c1}x_1 + w_{c2}x_2 + \ldots\ldots + w_{cp}x_p$$

where $c = 1 \ldots m$ principal components and $i = 1 \ldots p$ variables.

Subject to the constraint $\sum_{i=1}^{p} w_{ci}^2 = 1$, the weights are chosen so that the variance of PC_c is a maximum (Krzanowski, 1993). The first principal component accounts for the majority of variance in the data and subsequent ones represent variation that is unrelated to any preceding principal component.

Multivariate statistical techniques have been employed in the analysis of flow cytometric data (*e.g.* Demers *et al.*, 1992; Carr *et al.*, 1996), however, some of these procedures have limitations, requiring decisions to be made prior to analysis. Dimensionality reduction techniques may cause loss of discriminatory information and possibly increase the overlap between classes, thereby making it more difficult to determine class membership. Bayesian statistics requires an *a priori* measure, normally

the probability of a class' existence within the data set, to produce an accurate identification or allocation of a pattern to 1 of k classes. Prior knowledge is not generally available for the data considered in this research. Conversely, cluster analysis does not allocate to a pre-determined class. The threshold requirements of this method can be very much user-dependent *i.e.* similarity/dissimilarity measurements, and as the process of grouping or partitioning can continue until there exists either one group containing all patterns, or as many groups as there are patterns, knowledge of the number of groups actually present in the data set may be required. Whilst an appropriate statistical procedure may work well, the choice of technique is not always easy with difficulties arising if the assumptions made about the algorithm or data are incorrect.

Bayesian statistics (Chapter 3 & 4) and cluster analysis (Chapter 5) are further investigated and used in conjunction with ANNs.

1.6.2 Neural Computing

1.6.2.1 The Biological Pattern Recognition System

The basic element of information processing in the human brain is the neuron, an electrochemically respondent cell. Projecting from the cell are numerous dendrites and a single axon. The dendrites receive information from other neurons as electric impulses. These are transmitted across synaptic junctions, between the dendrites of one cell and the axons of others, through a chemical transfer using neurotransmitters (Heimer, 1995). The dendritic spines and the branch-like nature of the dendrites, ranging from one micrometer to one millimetre in length, designates a considerably large surface area upon which to receive information from other neurons. The human brain contains approximately 10^{11} neurons, each one making connections with thousands of other neurons, indicating the presence of over 10^{14} synapses.

As a nerve impulse reaches the synaptic junction of a neuron, neurotransmitters are released to a receiving neuron. This interaction produces an action potential (electrical impulse), which will either depolarise or hyperpolarise the receiving synapse (Heimer, 1995). The nature of the synapse may be excitatory or inhibitory and will influence whether the receiving neuron 'fires'. Each neuron receives impulses from thousands of other neurons. These impulses are scaled through adaptive strengths associated with each

synapse. If the summation of the incoming signals passes the excitation threshold, the neuron fires a sequence of action potential spikes down the axon to the dendrites of receiving neurons and the process continues.

The processing time of a biological neuron is relatively slow when compared to electronic devices. However the overall processing, performed at an immensely parallel level, delivers complete computation in the order of 10^2 ms, far exceeding any automated processing (Bishop, 1994). The real-time capability of the brain is exceptional, as is its ability to adapt to various states, to be able to extrapolate and interpolate data and still function given a 'noise' situation. Additionally, it has a vast memory capacity and continues to function in the event of damage or death of individual neurons. It was these characteristics that inspired the onset of neural computing.

1.6.2.2 Artificial Neural Networks

Typical von Neumann computers (von Neumann, 1945; Goldstine, 1972; Randell, 1982) have the ability to perform a great many tasks at tremendous speeds. However, these computers are capable only of serial processing, requiring an external, pre-determined, logical set of instructions to solve a particular problem. Thus, in an attempt to model the pattern recognition ability of the human brain, neural computation was developed, encompassing the capacity to simulate parallel performance with a distributed control that does not require the specific dictation of logical algorithms.

An Artificial Neural Network, or connectionist system, consists of a number of inter-connected nodes (neurons), linked by weighted junctions (analogous to synapses). Each of the nodes performs a relatively simple processing function, simulating an overall parallel computation. The basic features observed in the biological brain translate to ANNs in various ways, the majority of which are listed below:

- The neuron or processing unit in an ANN has a synaptic strength or weight modifying the signals.
- Neurons receive and pass on signals to multiple neurons.
- Some function is applied to the incoming signals.
- Memory storage can be long term and adaptive through the weighted connections.

- Weights can be modified during *learning* (Chapter 2) to adapt to changes in environment.
- No programming is involved.
- They have the ability to recognise patterns immersed in noisy or uncertain data.
- ANNs are relatively robust to loss of information.
- They operate in real-time.
- They are capable of generalising from specific data.

These characteristics make ANNs advantageous to a diverse number of fields, especially where the problem supplies a vast quantity of multi-variate data. Their generalisation ability minimises the need for *a priori* knowledge and they do not require assumptions to be made about the distribution of the data, or the selection of an appropriate algorithm.

1.6.2.3 Neural Network Characteristics

There are many types of ANNs, distinguished primarily by their learning/training algorithm, the node characteristics and the network topology. A typical structure of a basic Multi Layer Perceptron (MLP) (Hush & Horne, 1993; Chapter 2, Section 2.2) is shown in Figure 1.4. The processing units are arranged in layers. Typically, one hidden layer (Chapter 2, Section 2.2.2.1) is used but more can be introduced, depending on the structure and complexity of the data. Each node in a source layer is linked, through a weighted connection, to every node in its destination layer. The number of nodes in the input layer are representative of the number of input parameters for the specific problem, in this case, seven optical flow cytometry parameters, one node per parameter. The transformed output from the hidden layer node is received at the output layer, where the class membership is determined. There are a number of distinct categories for distinguishing ANNs.

Dynamic or Static

Dynamic networks, *e.g.* The Hopfield Network and Continuous Time Recurrent Network (Hush and Horne, 1993), are capable of memory storage represented by node equations, either differential or difference. Conversely, static networks have memory-less

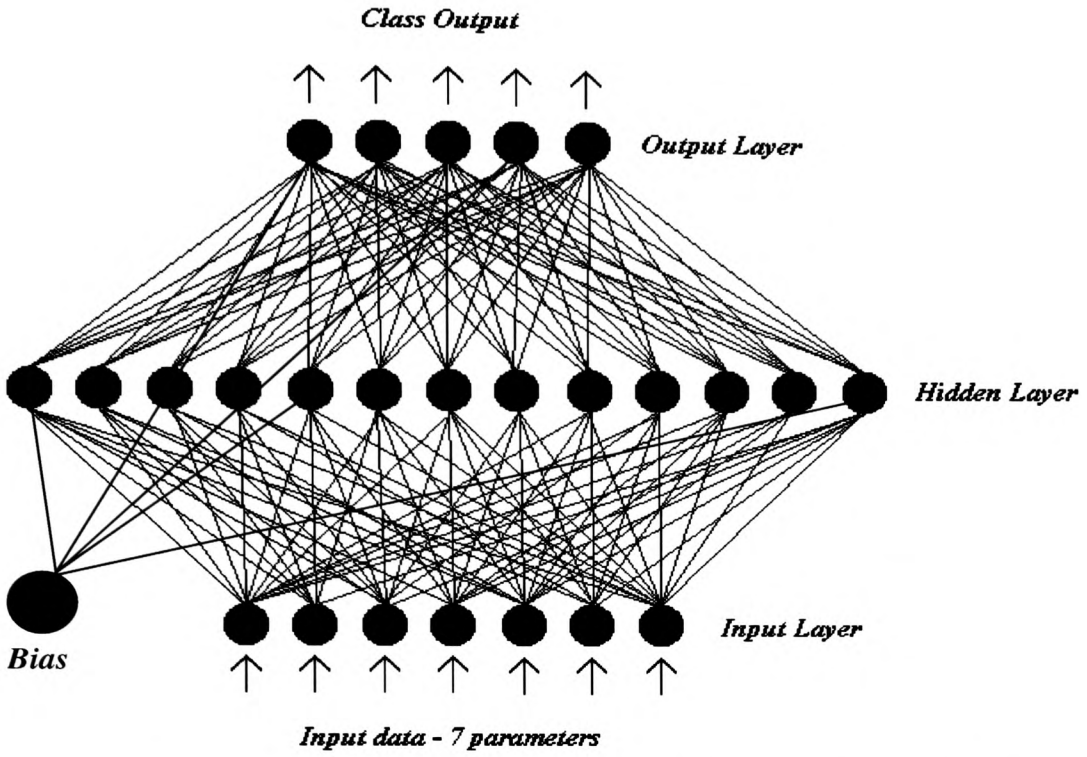


Figure 1.4 A feed forward Multi Layer Perceptron with one layer of hidden nodes, seven input nodes, representing the 7 parameters of flow cytometric data, and an output layer with as many nodes as there are possible classes. The bias node acts as an extra hidden unit, with a constant value of 1.0, connected to all nodes in the hidden and output layer (not all connections are shown) (Chapter 2, Section 2.2).

node equations, where each output is a function of the present input only, *e.g.* Radial Basis Function Network (RBF) (Moody and Darken, 1989; Broomhead and Lowe, 1988; Chapter 2, Section 2.4), and the Multi Layer Perceptron.

Binary or Continuous

Although many networks may be capable of processing both binary and continuous valued inputs, certain networks were designed for one type of data, for example, Adaptive Resonance Theory 1 (ART1) (Carpenter & Grossberg, 1987a) analyses binary data only. The ART2 network, by the same authors, accepts both binary and continuous data (1987b).

Supervised or Unsupervised

The learning mode is a primary differentiation between neural network algorithms. *Supervised networks*, like many classical pattern recognition systems, require the presence of some external knowledge. These paradigms, including the MLP, RBF and Learning Vector Quantisation network (LVQ) (Kohonen 1988, 1990), are used for identification rather than classification and learn through pattern association. The network requires *a priori* information in order to assign identified patterns correctly. In supervised networks, learning is via a set of n patterns, each of which belongs to 1 of N classes, where N may or may not be equal to n. Each of these n patterns has an associated label (target) indicating to which of the N classes it belongs. As training proceeds the network parameters are adapted, according to the learning algorithm, until the relationship between inputs and targets has been sufficiently modelled. For example, those patterns belonging to class 2 will have a target output 0,1,0,...,0, and the network output should reflect this association. The values produced by the network will generally be continuous and the node producing the highest valued output indicates class membership.

Unsupervised networks, such as Kohonens Self Organising Map (SOM) (Lippman, 1987; Kohonen 1990, 1997; Chapter 5), and ART networks, are independent of any external knowledge. These networks are not given a patterns associated identity and are therefore not forced to model any pre-existing relationship between input data and class membership. As no label is presented, training involves the network's own natural

determination of pattern relationship and clusters accordingly. This is performed through some measure of mathematical similarity, generally a hyper-dimensional distance (*e.g.* Euclidean).

Data Transfer

Networks are also identified by the order and direction of data transfer. All of the architectures considered in this research, *i.e.* MLP, RBF and SOM, are *feedforward* networks. Used in most applications, these networks map a set of input values directly to a set of output values through some transformation. The input layer of the basic architecture, acts only as the receptor for the input parameters, performing no function of any kind. The data are transformed by some function in the hidden layer nodes and passed through weighted connections to the output layer, where interpretation of the signals produces a decision. This process can continue for as many epochs (the number of presentations of the training set) as required, but only in one direction. There is no lateral or reverse propagation of signals in the networks and no time-varying behaviour.

Alternatively, *recurrent* networks, such as ART, exhibit dynamic behaviour capabilities. These networks are considered as an interconnection of units, rather than layers, introducing lateral and reverse direction connections. They are applied to complex computations which require the additional ability of a network to map temporal events.

1.7 Flow Cytometry and Neural Networks

Identification and classification of phytoplankton by neural networks has been investigated in a number of fields, including image analysis (Culverhouse *et al.*, 1996; Ellis *et al.*, 1997) and remote sensing data (Scardi, 1996 & 1998). Much of the earlier research documented for ANN analysis of flow cytometric data, employed only MLP networks as the supervised approach to these pattern recognition problems. One of the first works which demonstrated the potential of the ANN/AFC approach, Frankel *et al.* (1989), used a back-propagation network and a Kohonen Self Organising Map to distinguish between large, naturally occurring phytoplankton species (Prochlorophytes and *Synechococcus*) from noise and calibration beads. Later, Frankel *et al.* (1996) used back-propagation networks to identify a number of laboratory grown cultures of phytoplankton,

as well as natural populations, demonstrating further the robustness and flexibility of ANNs. Previous to this, Smits *et al.* (1992) and Balfoort *et al.* (1992) used MLPs to distinguish between cyanobacteria and non-cyanobacteria data, analysed on the Optical Plankton Analyser (OPA). The results for separating the poisonous from the non-poisonous algae were in the order of 99% correct. However, these two species were visually discernible through two- dimensional scatter plots, showing the nature of the identification to be relatively simple. Subsequent investigations include Wilkins *et al.* (1994a), which used LVQ and SOM networks for analysis of 7 species of freshwater phytoplankton, and Wilkins *et al.* (1994b) where MLP and RBF networks were employed to identify a maximum of 12 marine phytoplankton classes.

All of the above works established the enormous possibilities in using ANNs to analyse flow cytometric data. However, the research mentioned so far considered only small data sets, with a limited number of classes (*i.e.* species), where the probability of correct identification/classification is expected to be high. In the natural world, the taxonomic categories would by far exceed those considered in many of the above studies. Exceptions to this are Boddy *et al.* (1994a & 2000), of which the initial work used a back-propagation network to discriminate between 42 species of phytoplankton, while the latter employed an RBF to identify 72 species, and Wilkins *et al.* (1999), which employed RBF networks to discriminate between 34 species. These works advanced the application of ANNs to AFC data by increasing the number of classes under analysis.

The research documented to date, illustrates that analysis of multi-variate flow cytometry data by ANNs is an extremely powerful approach to this area of pattern recognition. However, the natural abundance of phytoplankton populations requires the non-trivial task of analysis of a large number of both laboratory and field cultured species (classes), not always available in equal quantities. The discovery of a new species is inevitable, and its inclusion in subsequent analysis is a necessity. When this is encountered, the addition of the species to the database can cause problems with existing network architectures. Additionally, identification to taxonomic or genus level by supervised networks may be improved, if the labelled classes were more reflective of flow cytometric signatures than their original morphological groupings. This involves analysis by unsupervised networks to provide an indication of flow cytometric similarities. Thus,

to fully realise the potential of the approach, further development is required to overcome these problems.

1.8 Aims and Objectives

The primary aim of this research was to investigate the application of neural networks to analytical flow cytometry data, for the identification and classification of phytoplankton. The rationale for the work was prompted by the requirements of the PRiME project (Section 1.5). A number of areas of investigation have been considered, an outline of which is presented.

1. Analysis of different supervised network paradigms to determine the optimal algorithm for the data considered (Chapter 2).
2. Analysis of network performance as the number of classes and events per class are varied, for both balanced and imbalanced training sets (Chapter 3 & 4).
3. Analysis of the selected paradigm with a high number of classes (Chapter 3).
4. Development of an alternative multiple network approach, simulating a parallel dynamic decision process, to overcome some of the limitations of the original multi-class network architecture (Chapter 4).
5. The subsequent improvement of supervised network performance, by using training data for which class membership is determined through flow cytometric similarities, rather than forcing morphometric groupings. (Chapter 3 & 4).
6. The development of various techniques of grouping data through determination of cluster centres and cluster boundaries on a Kohonen SOM (Chapter 5).
7. Illustration of the variance between morphology and flow cytometric signatures for some phytoplankton species (Chapter 5).
8. Analysis of the generalisation ability of a network, trained on a particular set of laboratory grown cultures (Chapter 2, Section 2.6), to be able to identify species from both differing synthetic cultures and actual field samples (Chapter 6).

2 Supervised Neural Networks - MLP or RBF?

2.1 Introduction

With the exception of a few papers (Wilkins *et al.*, 1994b, Wilkins *et al.*, 1999; Al-Haddad *et al.*, 2000; Boddy *et al.*,2000), the majority of research aimed at identification of phytoplankton has utilised probably the most common supervised network architecture, the Multi Layer Perceptron (Frankel *et al.*, 1989, 1996; Balfoort *et al.*, 1992; Smits *et. al.*, 1992). Although this network has a documented performance success in pattern recognition areas, an alternative paradigm, the Radial Basis Function (Moody & Darken, 1989; Broomhead & Lowe, 1988), has a number of advantages, including relatively rapid training times and the ability to reject unknowns, which appear to make it more suitable for the flow cytometric patterns being analysed.

This chapter discusses each of the architectures and applies both networks to a selection of 12 species from various taxonomic groups. The networks are assessed on suitability for the data, identification, and rejection of unknowns. An optimum paradigm is chosen and the reasons discussed.

2.2 Multi Layer Perceptron Artificial Neural Network (MLP ANN)

2.2.1 Architecture

The basic MLP consists of a number of processing elements arranged in layers (Chapter 1, Fig. 1.4). The N_{in} nodes in the input layer represent the N_{in} parameters of the training data. These input nodes serve only to distribute the input pattern to the one or more hidden layers of nodes, where a transfer function is implemented. The number of nodes in the hidden layer(s) is determined either heuristically or empirically. Computation also takes place within the output layer nodes, of which there exist as many as there are classes in the training data. Every node in a single layer is connected, via weights, to every node in a preceding or succeeding layer, and the computations that take place create an arbitrarily close approximation to any non-linear mapping.

2.2.2 Algorithm

The most common MLP algorithm is the back-propagation network, which is based on the error-correction learning rule (Rumelhart & McClelland, 1986; Haykin, 1994). This algorithm is performed in three main stages. Initially, training patterns are presented to the

input layer and propagated forward (forward pass) via inhibitory or excitatory weighted connections. The calculated output from the network is compared to the desired output for the particular pattern and an error value determined. Next, the associated error is propagated backwards (backward pass) through the layers, and the final stage involves small adjustments to the weights, to reduce the calculated error. This process continues for a number of epochs until a defined limit is reached, or the error converges to a minimum. At this point the network is considered trained and any subsequent testing or usage involves only data flow in the forward direction.

2.2.2.1 Forward Pass

Hidden Layer Nodes

The input signal of the training vector **x**, where $\mathbf{x} = x_1, x_2 \ldots . x_{Nin}$, is passed on to the hidden layer nodes for initial computation. These nodes act as feature extractors, forming an arbitrary mapping from N_{in}-dimensional input space to N_{out}-dimensional sample space. When considering a network with one hidden layer, the input of hidden node j is a weighed summation of the outputs from the N_{in} nodes in the input layer (for a network with a second hidden layer, this is the summation of the outputs from the preceding hidden layer nodes):

$$a_j = \sum_{i=1}^{N_{in}} w_{ji} x_i + w_{j0}$$

where w_{ji} is a weight vector from input node i to hidden node j and, w_{j0} denotes the bias for hidden node j, which acts as a weight vector with a constant value of 1. This can be represented within the summation by setting x_0 to 1:

$$a_j = \sum_{i=0}^{N_{in}} w_{ji} x_i$$

The activation, a_j, is transformed by a non-linear transfer function g() and the subsequent output, o_j, is passed on to the next layer:

$$o_j = g\left(\sum_{i=0}^{N_{in}} w_{ji} x_i \right)$$

in the case of one hidden layer node this is passed to the output layer.

The hidden layers of an MLP form infinite linear hyper-plane boundaries bisecting

the data space. Identification of a pattern is dependent upon an input vector's proximity to one of these boundaries and upon which side of the boundary it falls, both of which are depicted by its activation level, a_j. The orientation of the decision plane is defined by the node's weight vector, while the bias governs its perpendicular distance to the origin, ensuring the hyperplanes are not constrained to pass through the origin.

Transfer Function

A two layer perceptron, containing just input and output layers, is not capable of mapping a complex function that may bound non-linearly separable regions. A multi-layer linear system also lacks this ability, performing only as a single layer linear network. The activation function in the hidden layers makes the MLP more powerful than the basic perceptron, by introducing non-linearity into the network.

The elementary requirement for the transfer activation function is that it is differentiable and preferably bounded. Two of the most common non-linearly increasing activation functions are the sigmoidal, such as the monotonic increasing logistic function:

$$g(x) = \frac{1}{1 + e^{-x}}$$

bounded between 0 and 1 (Fig. 2.1a), and the hyperbolic tangent function (Fig. 2.1b), bounded between –1 and +1:

$$g(x) = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}}$$

Output Layer

Each node in the output layer receives a weighted summation of the nodes from the preceding hidden layer. After application of a transfer function the value produced at the output layer nodes, y, indicates the network's response to a particular input pattern:

$$y_n = g\left(\sum_{j=0}^{N_{hid}} w_{nj} o_j\right)$$

where N_{hid} is the number of hidden layer nodes, w_{nj} is the weight from hidden node j to output node n, and the bias term has been again been absorbed in the summation. Although g() is depicted as the transfer function, there is no constraint on the form of the function within the output layer being the same as that of the hidden layer.

When the network uses a squared error cost function (Section 2.2.2.2), and the

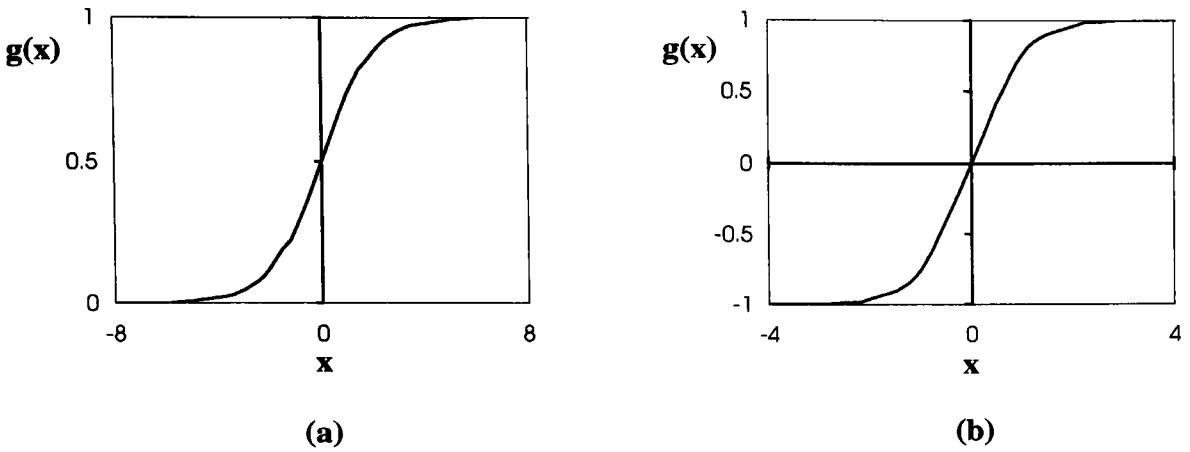


Figure 2.1 Non-linear transfer functions applied to the activation of the nodes in the MLP. (a) Sigmoid Function bounded between 0 and 1 (b) Hyperbolic Tangent bounded between –1 and 1. Other possibilities include trigonometric, log or Gaussian functions.

outputs are termed 1 of N_{out}, it has be shown that with sufficient training data, the network outputs estimate Bayesian *a posteriori* probabilities, providing the class conditional probabilities and *a priori* probabilities are accurately reflected in the composition of the data (Richard & Lippman, 1991); this is discussed further in Chapters 3 & 4.

2.2.2.2 Backward Pass - Update

As the final calculation reaches the output layer, the backward pass begins with the determination of the associated error for a particular pattern, **x**:

$$e_n(x) = \left(t_n(x) - y_n(x)\right)$$

where $y_n(x)$ is the network output for node n with pattern **x**, and $t_n(x)$ is the target output for node n with pattern **x**, which takes a value of 1 if the pattern belongs to the class represented by node n, and 0 otherwise. The error or *cost function* is evaluated over all nodes for a particular pattern as the squared error:

$$E(x) = \sum_{n=1}^{N_{out}} e_n^2(x) \qquad \text{where } n = 1,2\ldots N_{out}$$

In order to achieve an optimum trained level, the free parameters within the network, *i.e.* the weights, must be adjusted so as to minimise the error. This optimisation is done via the error back-propagation, or *Generalised Delta Rule* (GDR). This is a process of gradient descent, where the errors computed at the output layer are propagated backwards through the network, and the local errors, *i.e.* local gradient (δ), for each node are calculated (Lippman, 1987; Schalkoff, 1992; Hush & Horne, 1993; Bishop, 1995). Weights are then successively updated layer by layer in an attempt to reduce the error.

The local gradient (delta value), δ_n, is computed at the output layer as a product of the error for the particular node, n, and the derivative of the associated activation function:

$$\delta_n = e_n(x)\, g'(a_n)$$

The weight and bias correction terms are then defined respectively:

$$\Delta w_{nj} = \eta \delta_n o_j \qquad j = 1,2\ldots N_{hid}$$

$$\Delta w_{n0} = \eta \delta_n$$

where η is a learning-rate parameter.

The hidden nodes have no associated target and therefore the error term, δ_j, for a node j in the hidden layer, is calculated in terms of the delta inputs from the nodes in the

output layer and the derivatives of the activation functions:

$$\delta_j = g'(a_j) \sum_{n=1}^{N_{out}} \delta_n w_{nj}$$

with weight and bias update as before:

$$\Delta w_{ji} = \eta \delta_j x_i$$

$$\Delta w_{j0} = \eta \delta_j$$

where, for a one hidden layer network, x_i is the output from node i in the input layer.

Once all delta values have been calculated, the weights are updated simultaneously throughout the network for each particular connection:

$$w_{new} = w_{old} + \Delta w$$

This iterative training process proceeds with continuous pattern presentations until the error reaches a stable minimum, or the algorithm is halted.

2.2.3 Network Initialisation

2.2.3.1 Hidden Nodes

Kolmogrov's Mapping Neural Network Existence Theorem (Kolmogrov, 1957) generalised that any multi-variable continuous function, for a closed and bounded domain, could be represented by the superposition of a small number of single variable functions (ref. by Bishop, 1995). This suggested that a 3 layer network (*i.e.* one hidden layer) can perform any continuous mapping *g(x)*, from an input space of *d* dimensions exactly to an output space of *m* dimensions, where the single hidden layer has (2d+1) units. This theorem incorporates a monotonic increasing function within the hidden layer nodes, but is limited in its usage as no indication of the form of the function is known.

Generally the number of layers is dependent upon the complexity of the data and the decision boundaries required. However, a network with two hidden layers should be sufficient to map any complex function, and it has been suggested that there should be at least 3 times as many nodes in the second hidden layer as in the first (Lippman, 1987). The performance of one and two hidden layers was researched by de Villiers and Barnard (1992), where a comparison yielded no significant difference and in fact found that a one hidden layer network had a higher identification success.

An increased number of hidden nodes does not always improve network performance and in some cases is counterproductive, for two main reasons. Firstly, too many nodes can have a significant influence on the time taken to achieve an optimum training level and secondly, surplus nodes may start to memorise the data causing the network to reach a state of overtraining, becoming unable to generalise on unseen data. The choice of how many hidden nodes and the structure of the layers is still very much a trial and error process, and can sometimes become a trade-off between training time and performance level.

2.2.3.2 Weights

The initial selection of weights has a significant influence on the successfulness of the network to learn and converge. An incorrect initialisation can cause slow convergence due to *premature saturation* (Lee *et. al.*,1991). This effect is apparent when the cost function appears to have reached a stable state, remaining almost constant for a period, but subsequently decreases further indicating a local minimum was reached instead of a global one. Lee suggests premature saturation can be avoided if the initial values of the weights are set to small, uniformly distributed, random numbers of the same dimension as the input data.

2.2.3.3 The Learning Rate Parameter : η

The learning rate parameter controls the degree of change to the synaptic weights. For too small a value of η the weight changes are minimal, resulting in prolonged training periods and a risk of the error being trapped in local minima, where the value of η is not large enough to allow escape. However, too large a value of η causes instability in the network and convergence will be slow or unobtainable.

Setting η to an initial value and sequentially decreasing it as training progresses, avoids early risk of falling into local minima and increases the chance of convergence (Bishop, 1994).

2.2.3.4 Momentum : α

The introduction of a momentum term, α, to the GDR helps the system to avoid weight oscillation and aid convergence. This is achieved by adding a fraction, α (where

$0<\alpha<1$), of the previous weight adjustments at iteration t, to the present weight adjustment at iteration $t+1$. The momentum term, like the learning rate parameter, is also decreased during training. For further discussion of MLP and back-propagation see Rumelhart & McClelland 1986; Hush & Horne, 1993; Bishop, 1995.

2.3 Radial Basis Functions and Interpolation

The RBF network is an alternative supervised paradigm, which has been used more recently in pattern recognition of phytoplankton flow cytometric signatures (Wilkins *et al.*, 1994b, Wilkins *et al.*, 1999; Al-Haddad *et al.*, 2000; Boddy *et al.*, 2000). The network uses basis functions, or kernels, to determine the node activation via an arbitrary distance measure between the input vector and the kernels. Their initial use was for interpolation in hyper-dimensional space (Powell, 1985), and they were implemented into an ANN by Broomhead and Lowe (1988).

Given a set of N_{in} input vectors, x_i, each with a target vector t_i, the interpolation problem requires finding a function when mapping from a d-dimensional input space, $\mathbf{x}$, to a one-dimensional output space, t, such that:

$$f(x_i) = t_i \qquad i = 1,2,.....N_{in}$$

This is implemented via a set of N_{in} basis functions of the form $\vartheta(\|x - x_i\|)$, where x_i are the centres of the basis functions, and ||.|| indicates the norm, usually Euclidean (Broomhead & Lowe, 1988; Haykin, 1994). A weighted, linear combination of the basis functions forms the output of the mapping:

$$f(x) = \sum_{i=1}^{N_{in}} w_i \vartheta(\|x - x_i\|)$$

This translates to multiple output variables as each input vector, x_i, is associated with a target vector, t_i, generalising to:

$$f_j(x_i) = t_{ij} \qquad i = 1,2,.....N_{in} \quad j = 1,2.....N_{out}$$

which leads to:

$$f_j(x) = \sum_{i=1}^{N_{in}} w_{ij} \vartheta(\|x - x_i\|)$$

2.4 Radial Basis Functions as Neural Networks

A direct implementation of the interpolation function as a neural network, requires a number of modifications to overcome certain restrictions (Broomhead and Lowe 1988; Moody and Darken 1989). Primarily, the requirement of as many basis functions as there are input vectors makes the procedure cumbersome and mathematically demanding. In cases where noise may be present, forcing the function to pass through every data point creates a continuous differentiable surface, that has a high oscillatory character (Bishop, 1995). When modelling the underlying nature of the data this fitting of misleading variations is undesirable and results in poor generalisation. A function is required that is capable of averaging over any noisy data or anomalies, and is able to generalise enough to recognise unseen data. Allowing basis function number (N_{hid}), position and width to be variable and data dependent, makes the RBF a powerful interpolation tool, which is capable of mapping high dimensional data to a linearly separable space:

$$y_j(x) = \sum_{k=0}^{N_{hid}} w_{jk} \vartheta_k(x)$$

where w_{jk} is the weight vector between kernel centre k (hidden node) and output node j and the bias term has been absorbed in the summation by including an extra basis function ϑ_0, whose output is 1.

2.4.1 Architecture

The three layer architecture of the RBF network (Fig. 2.2) is similar to that of the MLP. The N_{in} nodes in the input layer, one for each parameter of the data, transfer the N_{in}-dimensional training data to the hidden layer (of which there is normally one). The hidden layer nodes implement the basis functions and propagate their output, via weighted connections, through a linear summation that takes place in the output layer. A network decision on class membership is formed in the 1 of N_{out} output nodes.

Unlike the MLP, training in an RBF takes place in two separate stages. The first stage involves locating the positions of the basis functions and can be determined through supervised or unsupervised methods. The second stage involves calculating the connection weights between the hidden layer nodes and the output layer nodes.

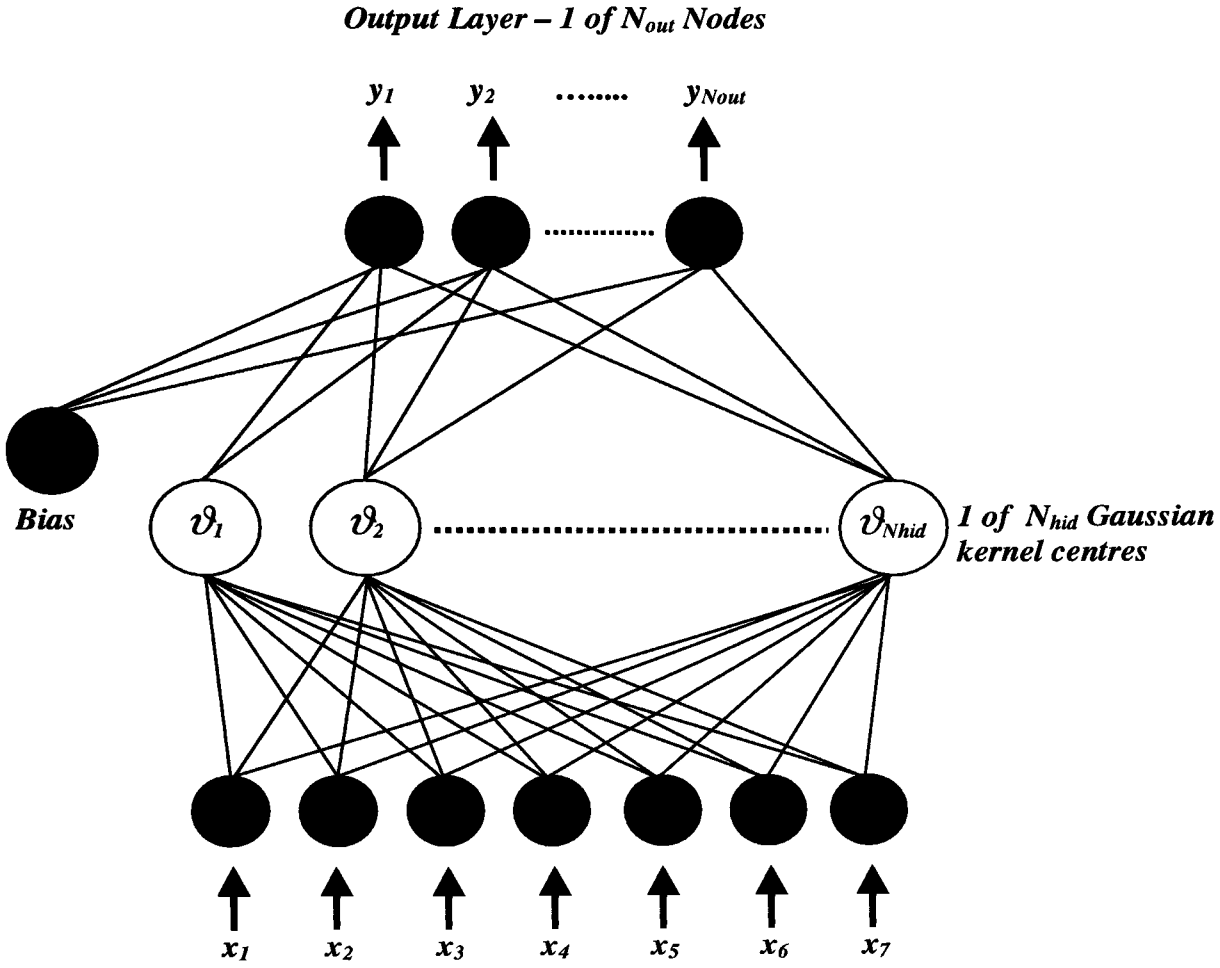


Figure 2.2 A Radial Basis Function Network with one layer of hidden nodes, seven input nodes (N_{in}), representing the 7 parameters of flow cytometric data, and an output layer with as many nodes as there are possible classes. The bias node acts as an extra hidden unit, with a constant value of 1.0, connected to all nodes in the output layer (not all connections are shown).

2.4.2 Learning Algorithm

2.4.2.1 Location of the Basis Functions

The basis functions in an RBF network each represent a finite area of the input space. Each responds only to data points that fall within a small localised region of the centre of the particular kernel. A suitable number of kernels must therefore be chosen to span the entire input space and suitably model the data. This will allow the network to generalise without causing overfitting or memorisation, where only data found in the training set is recognised.

A number of methods exist to select and optimise the position of the basis function centres, including subsets of input data, orthogonal least squares, K-means and supervised clustering, all of which are summarised below.

Subsets of input data

k random pattern vectors are selected from the training data and implemented as kernel centre locations. Although the simplest method, it requires a large number of data points for good performance and is a poor approximator of density estimation, leaving many highly populated areas over-represented, while sparser regions are empty. It is used mainly as a starting point for other more optimal methods, where placement is generally an iterative process.

Orthogonal Least Squares

This method defines the network as a linear regression model and involves the selection of basis functions centred at different data points (Chen *et al.*, 1991). Like the random sub-selection, the process defines which data points will be chosen to represent basis functions. Using an orthogonal procedure, such as Gram-Schmidt (Nering, 1970), a set of orthogonal vectors are constructed from the regressor vectors (basis functions), and the contribution of each to the output is established. The data point chosen as a basis function, is that which produces the greatest reduction in the sum of squares error.

K-means (Unsupervised Clustering)

This algorithm determines the number of k kernel centres in advance, via unsupervised clustering. It is based on the minimisation of the sum-of-squared distances

between a kernel centre and the data points in its cluster domain (Tou & Gonzalez, 1974; Moody & Darken 1989). Initially, the centres are chosen at random from the data set. Hyper-dimensional distances between every input vector and each kernel centre are calculated and assigned to the centre for whom their distance is a minimum. This partitions the data points into k disjoint subsets, S_j, where j = 1,2....k. After primary separation, the centre positions are adjusted by moving the kernel to the mean of each cluster. The minimum distance between the new cluster centres and all data points is again calculated and, if required, the data points are assigned to new kernels. The means are re-computed and the iterative process continues until there is little, or no change in update of the centres.

Supervised Clustering

Unlike the previous methods a supervised placement of basis functions such as Learning Vector Quantisation (LVQ) (Kohonen 1988, 1990), will allow information about class membership to be encoded directly into the network. LVQ utilises a finite number of discrete *codebook vectors* or *reference vectors,* to model the distribution of individual classes, instead of the entire data set as one. M vectors are allocated to each of the identified classes present and the distribution is approximated defining class regions. In each of the subsets, S_j, the initial placement of the reference vector is random, after which competitive learning is employed to optimise placement of the winning node. The node deemed the winner is that one having the minimum Euclidean distance to the presented input pattern. If this node is representative of the class the data point is a member of, it is moved closer to that point. Conversely, if the winning node is representative of a different class, it is moved away from the point.

2.4.2.2 Form of the Basis Function

The exact form of the basis function appears to have little effect on the performance of the RBF networks (Haykin, 1994). As this is not an issue here only the most common form, the Gaussian kernel function (Fig. 2.3) is implemented throughout:

$$\vartheta_k(x) = \exp(-\frac{\phi_k{}^2}{2\sigma_k{}^2})$$

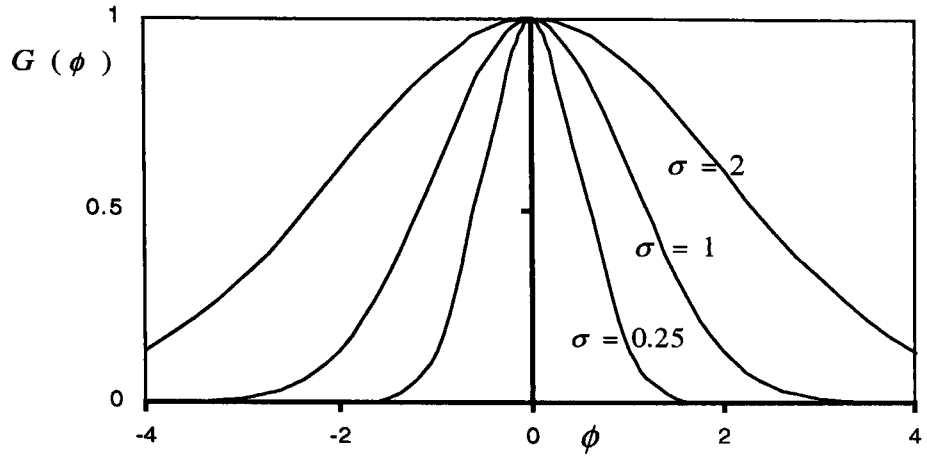


Figure 2.3 The Gaussian kernel function with varying values of σ, the width parameter. This is the most common form of the basis function employed by RBF networks.

Translated to the output of the hidden layer node, the function is defined in terms of a Euclidean distance metric between the kernel centres and the data points:

$$\vartheta_k = \exp\left[-\frac{(x-m_k)^T(x-m_k)}{2\sigma_k^2}\right]$$

where **x** is the N_{in}-dimensional input vector, $\mathbf{m_k}$, is the position vector of the k^{th} kernel centre and σ_k^2 is the normalisation parameter, which controls the width of the k^{th} kernel.

The basis functions form radially symmetric concentric boundaries around their centre locations, at which point the node output is one, reducing to zero as the distance from the centre increases, thereby producing a localised response to input patterns. The normalisation parameter can be calculated as the average Euclidean distance between a kernel centre and the N_k corresponding data points it represents (Hush and Horne, 1993):

$$\sigma_k^2 = \frac{1}{N_k}\sum_{x\in S_k}(x-m_k)^T(x-m_k)$$

where S_k is the set of input vectors represented by kernel k. This can also be represented by the trace of the variance-covariance matrix, C_k, which summarises the distribution of data assigned to the particular kernel, k (Wilkins *et al.*, 1994b):

$$C_k = \begin{bmatrix} \sigma_{11}^2 & \sigma_{12}^2 & \cdots & \sigma_{1N}^2 \\ \sigma_{21}^2 & \sigma_{22}^2 & \cdots & \vdots \\ \vdots & \vdots & & \vdots \\ \sigma_{N1}^2 & \sigma_{N2}^2 & \cdots & \sigma_{NN}^2 \end{bmatrix}$$

and

$$trace(C_k) = \sum_{i=1}^{N_{in}} \sigma_{ii}^2$$

where σ_{ii} is the variance of the ith component of the input vector **x**:

$$\sigma_{ii}^2 = \frac{\sum_{x\in S_k} x_i^2}{N_k} - \left[\frac{\sum_{x\in S_k} x_i}{N_k}\right]^2$$

and σ_{ij} is the covariance between the ith and jth components of **x**:

$$\sigma_{ij}^2 = \frac{\sum_{x\in S_k} x_i x_j}{N_k} - \frac{\sum_{x\in S_k} x_i \sum_{x\in S_k} x_j}{N_k^2}$$

The functionality of the basis functions can be increased by replacing the Euclidean distance metric with the Mahalanobis distance metric, thus:

$$\vartheta_k = \exp(-\frac{1}{2}(x-m_k)^T \sum_k^{-1} (x-m_k)) \qquad \text{where} \quad \frac{1}{2}\sum_k^{-1} = C_k^T C$$

This more versatile distance metric alters the spatial extent and shape of the basis functions, making them hyper-ellipsoids with varying principle axes (Hush & Horne, 1993). The actual distribution and dimensions of the function are defined by the *normalisation matrix*, $\sum_k^{-1}$ which takes one of three forms depending on C_k.

1. The principal axes of the hyper-ellipsoids are indicated by the eigenvectors of C_k, whose eigenvalues give the variances along the respective principal axes directions (Bishop, 1995). If C_k is non-diagonal, the axes are not restricted to the co-ordinate axes of the data, but can be oriented along the axes of possible clusters (Fig. 2.4a). Since the shape of the basis functions is defined by the eigenvalues and eigenvectors of the matrix C_k, it is this that defines basis overlap. Although using large eigenvalues produces good generalisation, it can increase the loss of local properties due to overlap of different class clusters (Musavi *et al.*, 1992). This can be avoided if the covariance matrix for a basis function is constrained by the location of the nearest training point that does not belong to it. Musavi *et al.*, (1992) suggests a method for this using the Gram-Schmidt orthogonalisation procedure to incorporate the information into the matrix.
2. If C_k is diagonal and the diagonal elements are not equal, the basis functions are still hyper-ellipsoids but with principal axes parallel to the co-ordinate axes (Fig. 2.4b). A generalised version of the normalisation parameter can be applied to determine kernel width by defining individual scaling factors such that:

$$\sigma_{ik}^2 = \frac{1}{N_k} \sum_{x \in S_k} (x_i - m_{ik})^2$$

   giving the variance of the i^{th} component of the patterns allocated to kernel k.
3. If C_k is diagonal and the diagonal elements are equal, the basis functions are radially symmetric hyper-spheres. This is in fact the Euclidean distance, where the scaling factor, σ, controls the width (Fig. 2.4c).

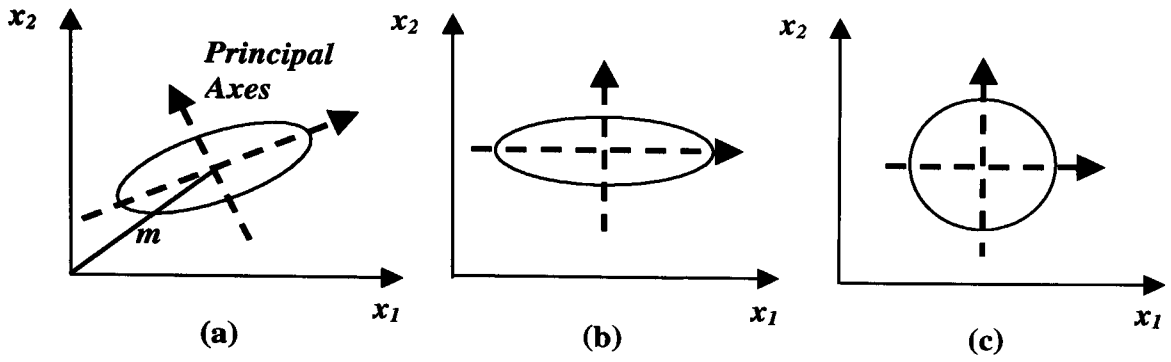


Figure 2.4 Spatial extent and shape of the basis functions with varying forms of the variance-covariance matrix where m is the vector defining the kernel centre. The principal axes are indicated by the eigenvectors of C_k, and the variances along the respective principal axes directions are given by the eigenvalues of the matrix. (a) C_k is a non-diagonal matrix forming hyper-ellipsoid basis functions whose principal axes are not restricted to the co-ordinate axes of the input data. (b) C_k is a diagonal matrix, where the diagonal elements are not equal, forming hyper-ellipsoid basis functions whose axes are confined to the co-ordinate axes of the input data. (c) C_k is a diagonal matrix, where the diagonal elements are equal, forming radially symmetric hyper-spheres.

2.4.2.3 *Output Layer Weights*

Once the hidden layer node parameters have been determined, the output from the basis function is propagated forward where a weighted linear summation occurs in each of the output layer nodes. Translated to matrix notation to give:

$$\Theta W = Y$$

where W = (w_{jk}), and w_{jk} is the weight connecting the k^{th} hidden layer node to output node j, Θ = (ϑ_{ik}), and ϑ_{ik} is the output of the k^{th} hidden node when the i^{th} input pattern is presented and Y= (y_{ij}) where y_{ij} is the output of node j when presented with pattern i.

The optimum weights are found by minimising the total squared error calculated over all patterns present in the training data:

$$E = \sum_{x} (y-t)^T (y-t)$$

where $\mathbf{y} = (y_1, y_2, y_3 \ldots y_{Nout})^T$ represents the actual network outputs and $\mathbf{t} = (t_1, t_2, t_3 \ldots t_{Nout})^T$ represents the actual target values, where t_1 is 1 if the pattern comes from class 1 and 0 otherwise. Although the solution can be solved via an iterative least mean squares procedure, placing the initial interpolation condition directly into the RBF matrix equation resolves the weights exactly, producing a set of linear equations (Bishop, 1995; Haykin, 1994):

$$\begin{bmatrix} \vartheta_{11} & \vartheta_{12} & \cdots & \vartheta_{1Nhid} \\ \vartheta_{21} & \vartheta_{22} & \cdots & \vartheta_{2Nhid} \\ \vdots & \vdots & & \vdots \\ & & & \end{bmatrix} \begin{bmatrix} w_{11} & w_{12} & \cdots & w_{1Nout} \\ w_{21} & w_{22} & \cdots & w_{2Nout} \\ \vdots & \vdots & & \vdots \\ w_{Nhid1} & w_{Nhid2} & \cdots & w_{NhidNout} \end{bmatrix} = \begin{bmatrix} t_{11} & t_{12} & \cdots & t_{1Nout} \\ t_{21} & t_{22} & \cdots & t_{2Nout} \\ \vdots & \vdots & & \vdots \\ & & & \end{bmatrix}$$

where t_{ij} is the target value of output node j when presented with pattern i.

This is the interpolation matrix $\Theta W = T$, which produces the formal solution for the weights that best minimises E as:

$$W^T = \Theta^{\dagger} T$$

where $\Theta^{\dagger}$ is the pseudoinverse termed $\Theta^{\dagger} \equiv (\Theta^T \Theta)^{-1} \Theta^T$.

Once the optimum weights are determined the matrix calculation, $\Theta W = Y$, can be solved, producing the network outputs which, like the MLP, represent the Bayesian *a posteriori* probability that an input belongs to 1 of N_{out} classes (Richard & Lippman, 1991; Chapter 3).

When all network parameters are defined, *i.e.* basis functions and weights, they can be further optimised through a procedure of gradient descent. This requires defining an error surface and gradient with respect to a particular parameter. The parameter adjustment is then set proportional to the negative gradient, in order to move them towards an optimum solution of minimum error.

2.5 Paradigm Summary

The MLP and RBF are both capable of forming an arbitrarily close approximation to any non-linear mapping between multi-dimensional spaces. However, there are significant differences between the two paradigms making each appropriate for different tasks. These are summarised below:

- Both architectures consist of an input and output layer. However, where the RBF normally has only one hidden layer, the MLP may have one or more.
- Training in an RBF takes place in two completely autonomous procedures whereas the MLP parameters are generally determined simultaneously, as part of a supervised global training strategy.
- The activation functions of the hidden layer of an MLP are non-zero over an infinitely large region of the input space. This will produce activation from a number of hidden layer nodes when presented with an input pattern. Conversely the RBF uses localised basis functions (*e.g.* Gaussian) which cover only limited hyperspherical or hyperellipsoidal regions of the input space, producing notable activation from only a small number of hidden nodes within the vicinity of the input pattern (Haykin, 1994).
- In an RBF the non-linear operations of the hidden layer have a different purpose and are completely separate to the linear activations of the output layer. The hidden and output layers of an MLP normally share a common non-linear model, whose activation functions are not necessarily the same.
- In the hidden layer of a RBF, a distance metric between the input pattern and kernel location, forms the argument of the activation function. In a MLP the activation function transforms the weighted summation of the input signals.

Whilst the differences between the algorithms of the two architectures are obvious, their empirical performance for the particular data set has to be established. The following

sections provide a comparison of both architectures when applied to the phytoplankton data.

2.6 Data Collection, Preparation and Pre-processing

A data set of laboratory cultured phytoplankton (PRiME 1) was received for analysis, supplied by the Plymouth Culture Collection. PRiME 1 consists of 62 species of phytoplankton from 5 taxonomic groups. The species spanned a range of sizes and morphologies, representative of natural phytoplankton populations (Table 2.1). The phytoplankton cultures were maintained in a Gallenkamp INF-781 incubator at approximately 15°C and were illuminated on a 12 hr light, 12 hr dark cycle at 50μmol quanta $m^{-2}s^{-1}$. Batch cultures were grown for several weeks before analysis and were sub-cultured every 3 – 4 days to maintain cultures in exponential growth. With the exception of *Emiliania huxleyi B11*, which was grown in F/10 medium, all species were cultured in F/2 medium (Guillard, 1975), with or without soil extract (Boddy *et al.*, 2000).

The cultures were analysed over 2 days with 3 species run a week later due to dilution levels being initially too high. The cultures were analysed by flow cytometry using thresholds as described in Chapter 1 (Section 1.4.2), imposed via FACStation™ software. Samples were run for four minutes at a flow rate of 100μl min^{-1}, with the majority of cases consisting of 3 replicate samples of 10,000 events for each species. Instrument drift was monitored by analysing Coulter™ Standard Brite™ fluorescent beads, 10 microns in size containing a fluorochrome with a broad band of emission from about 400-700nm. The data were gated at Plymouth Marine Laboratory by omitting events with low red fluorescence signals to exclude any possible noise clusters, such as bacteria or inorganic particles. This produced primarily uni-modal data, with the exception of a small number of multi-modal cultures reflecting, perhaps, cells at different stages of growth. The listmode files were linearly rescaled in the range 0 to 1 for the RBF ANN and –1 to +1 for the MLP ANN (using a tanh transfer function). Without rescaling, the weighted summation of a node can become large, causing saturation of the activation level (0 or 1), producing a zero derivative and halting learning (NeuralWare, Inc, 1991a).

Training and testing files were constructed via random sampling of the listmode files without replacement. This ensures complete autonomy between training and test files, and avoids the possibility of any systematic deviations present within the listmode files.

Table 2.1 Database of 62 PRiME 1 species from five taxonomic groups indicating species name, order and size (n.b. There are two strains of *Emiliania huxleyi*).

Taxonomic Group	Species Name	Order	Size µm
Cryptomonads	*Chroomonas sp.*	Cryptomonadida	8-10
	Chroomonas salina	"	5-12
	Cryptomonas appendiculata	"	15-25
	Cryptomonas calceiformis	"	10-15
	Cryptomonas maculata	"	12-20
	Cryptomonas reticulata	"	18-25
	Cryptomonas rostrella	"	16-25
	Hemiselmis brunnescens	"	5-8
	Hemiselmis rufescens	"	4-9
	Hemiselmis virescens	"	5-8
	Plagioselmis punctata	"	6-9
	Rhodomonas sp.	"	8-13
Flagellates	*Micromonas pusilla*	Prasinomonadida	1-3
	Nephroselmis pyriformis	"	4-7
	Nephroselmis rotunda	"	6-8
	Pyramimonas grossii	"	5-10
	Pyramimonas obovata	"	4-8
	Tetraselmis impellucida	"	11-19
	Tetraselmis suecica	"	6-15
	Tetraselmis verrucosa	"	3-11
	Tetraselmis tetrathele	"	10-16
	Tetraselmis striata	"	6-8
	Chlamydomonas reginae	Volvocida	11-20
	Chlorella salina	"	4-8
	Dunaliella minuta	"	3-12
	Dunaliella primolecta	"	5-12
	Dunaliella tertiolecta	"	6-12
	Stichococcus bacillaris	"	5-8
	Porphyridium pupureum	Rhodomonadida	4-6
	Rhodella maculata	"	7-24
	Ochromonas sp.	Chrysomonadida	3-12
	Pelagococcus subviridis	"	2-3
	Pseudopedinella sp.	"	8-10
Prymnesiomonad	*Chrysochromulina camella*	Prymnesiida	6-12
	Chrysochromulina chiton	"	5-9
	Chrysochromulina cymbium	"	6-10
	Chrysochromulina polylepis	"	6-8
	Emiliania huxleyi 92	"	5-6
	Emiliania huxleyi B11	"	5-7
	Ochrosphaera neopolitana	"	8-10
	Pavlova lutheri	"	4-6
	Phaeocystis pouchetii	"	3-6
	Pleurochrysis carterae	"	10-18
	Prymnesium parvum	"	8-10

Table 2.1 continued.....

Taxonomic Group	Species Name	Order	Size µm
Diatoms	*Amphora coffaeformis*	Bacillariophyceae	10-20
	Chaetoceros calcitrans	"	4-6
	Phaeodactylum tricornutum	"	8-35
	Skeletonema costatum	"	3-5
	Thalassiosira weissflogii	"	12-20
Dinoflagellates	*Amphidinium carterae*	Dinoflagellida	15-20
	Aureodinium pigmentosum	"	7-12
	Gymnodinium vitiligo	"	7-22
	Gymodinium micrum	"	8-15
	Gymodinium simplex	"	6-10
	Gymodinium veneficum	"	9-16
	Gyrodinium aureolum	"	35-45
	Heterocapsa triquetra	"	15-27
	Prorocentrum balticum	"	9-15
	Prorocentrum micans	"	30-40
	Prorocentrum minimum	"	16-18
	Prorocentrum nanum	"	8-10
	Scrippsiella trochoidea	"	30-42

2.7 Experimental Procedure

2.7.1 Training and Testing Files

In order to compare the performance of each paradigm, two sets of 12 species were selected from the PRiME 1 database to construct training and testing files. The first set comprised only Dinoflagellate species (Table 2.2), whose scatter plots of depolarised light and red fluorescence exhibit overlap and indeterminate distinction between clusters (Fig. 2.5a). The second set of species were chosen from the five taxonomic groups, due to a more distinct appearance of clusters from the two parameter scatter plots (Fig. 2.5b), and a greater range of cell sizes (Table 2.2). The training and test files for both sets contained 300 randomly chosen events (patterns) per species. Each event within the data files is labelled with its correct identification, associated with the species class of which it is a member (*i.e.* 1 of N).

An additional test file was created consisting of data for 12 unknown species (*i.e.* species upon which the network had not yet been trained; Section 2.7.4.2), chosen from four of the taxonomic groups (*i.e.* excluding Dinoflagellates), representing both separable and overlapping species (Table 2.3). This was used to evaluate each network's performance with novel data.

2.7.2 MLP training

MLPs were simulated using the Neural Works Professional II Plus (NeuralWare, Inc., 1991a, 1991b) software. The default learning/recall schedule was used for training, employing only the tanh transfer function as literature indicates its superiority over the sigmoidal (Refenes & Alippi, 1991; NeuralWare, Inc., 1991b; Refenes *et al.*, 1993; Harrington, 1993). The initial default values for the learning coefficient were 0.3 and 0.15, for the hidden and output layers respectively, with a momentum coefficient of 0.4 for both layers. Both parameters were decreased as training proceeded. Five different sized networks were trained and tested for each data set, containing 5, 10, 20, 30 and 40 hidden nodes in a single layer. Each network was trained three times from a different initialisation point to avoid a possible poor choice of weights (Section 2.2.3.2). The training for each of the five networks was for 100,000 pattern presentations, which was extended to 400,000 for the network selected as optimum (highest overall performance). This extended period of training had no effect on RMS error and showed no improvement in test data performance.

Table 2.2 Two sets of species, *i.e.* Dinoflagellate Group and Mixed Group, used to construct training and testing files for assessment of the two network paradigms, MLP and RBF.

	Taxonomic Group	Species Name	Order	Size µm
Dinoflagellate Group	Dinoflagellate	*Amphidinium carterae*	Dinoflagellida	15-20
	"	*Aureodinium pigmentosum*	"	7-12
	"	*Gymodinium simplex*	"	6-10
	"	*Gymodinium veneficum*	"	9-16
	"	*Gymnodinium vitiligo*	"	7-22
	"	*Heterocapsa triquetra*	"	15-27
	"	*Prorocentrum balticum*	"	9-15
	"	*Prorocentrum micans*	"	30-40
	"	*Prorocentrum minimum*	"	16-18
	"	*Prorocentrum nanum*	"	8-10
	"	*Scrippsiella trochoidea*	"	30-42
	"	*Gyrodinium aureolum*	"	35-45
Mixed Group	Cryptomonad	*Chroomonas sp.*	Cryptomonadida	8-10
	"	*Hemiselmis virescens*	"	5-8
	"	*Plagioselmis punctata*	"	6-9
	"	*Rhodomonas sp.*	"	8-13
	Diatom	*Amphora coffaeformis*	Bacillariophyceae	10-20
	"	*Chaetoceros calcitrans*	"	4-6
	Dinoflagellate	*Gyrodinium aureolum*	Dinoflagellida	35-45
	"	*Prorocentrum balticum*	"	9-15
	Flagellate	*Micromonas pusilla*	Prasinomonadida	1-3
	"	*Pelagococcus subviridis*	Chrysomonadida	2-3
	"	*Tetraselmis tetrathele*	Prasinomonadida	10-16
	Prymnesiomonad	*Emiliania huxleyi B11*	Prymnesiida	5-7

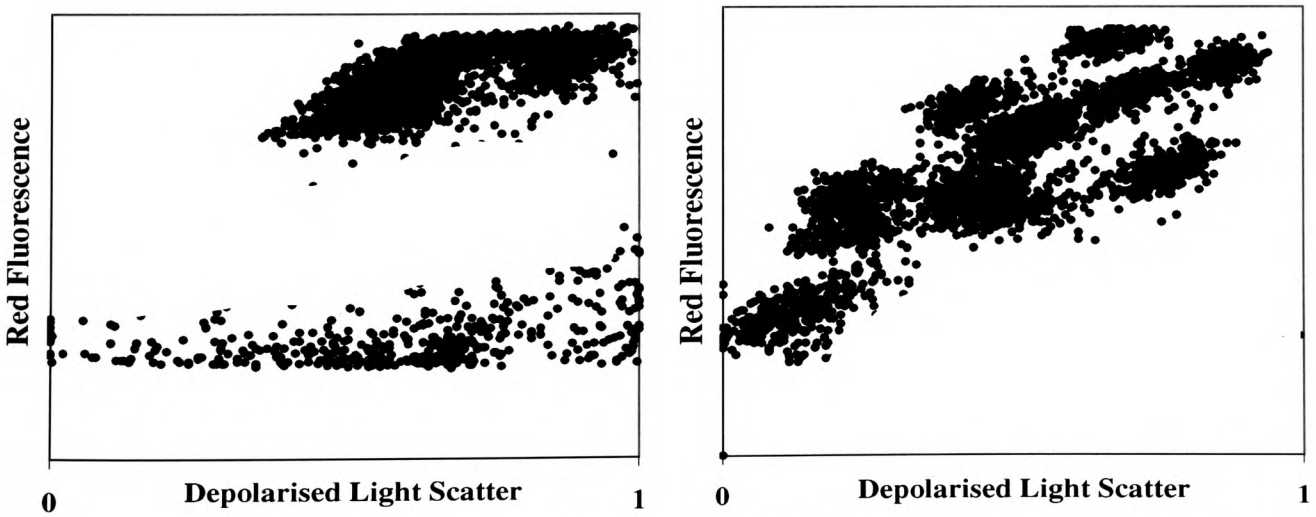


Figure 2.5 Two-dimensional scatter plots showing Depolarised Light Scatter against Red Fluorescence (Height). (a) Dinoflagellate species set (b) Mixed species set. Overlap intensity appears greater in the Dinoflagellate set.

Table 2.3 Novel data used to test the ability of both networks to rejection unknown species.

Taxonomic Group	Order	Species Name	Size µm
Cryptomonads	Cryptomonadida	*Chroomonas salina*	5-12
	"	*Cryptomonas maculata*	12-20
	"	*Hemiselmis rufescens*	4-9
Diatoms	Bacillariophyceae	*Phaeodactylum tricornutum*	8-35
	"	*Skeletonema costatum*	3-5
	"	*Thalassiosira weissflogii*	12-20
Flagellates	Volvocida	*Chlamydomonas reginae*	11-20
	"	*Dunaliella minuta*	3-12
	Prasinomonadida	*Nephroselmis rotunda*	6-8
Prymnesiomonads	Prymnesiida	*Chrysochromulina chiton*	5-9
	"	*Ochrosphaera neopolitana*	8-10
	"	*Pavlova lutheri*	4-6

2.7.3 RBF training

Using the AimsNet software (developed during the Aims project; Automated Identification and Characterisation of Microbial Populations; by Wilkins, 2000). RBFs were simulated applying the Mahalanobis distance. Seven networks were trained and tested for both data sets, containing a total of 12, 24, 36, 48, 60, 84 and 96 asymmetric Gaussian kernels respectively. Placement of kernel centres was done via the LVQ method (Kohonen, 1988; 1990). An optimal subset of the hidden layer nodes was automatically selected via the orthogonal least squares elimination process (Chen *et al.*, 1991). All networks were trained 3 times from different initialisation points, and optimised via 5 iterations of a conjugate directions gradient descent operation, which, as with the MLP, attempts to learn the free parameters in order to minimise the defined mapping error. The optimum network was then trained for a further 10 iterations, which failed to improve identification of the test data.

2.7.4 Testing Procedure

2.7.4.1 Probability Matrices

After training had terminated, performance of each network was assessed using a number of probability measures applied to the labelled test file. The generated results files contain values produced at each of the 1 of N_{out} output nodes, y, for every test pattern, x, presented, with the highest value indicating possible class membership. From these results, matrices were created indicating the probability of correct identification for each class as the leading diagonal values, *i.e.* $M_{ii} = p(y=i|x=i)$, and the probability of a species identified as class j, which actually belongs to class i, as the off diagonal values, *i.e.* $M_{ij} = p(y=j|x=i)$. From this, two additional probabilities can be determined: (1) the overall probability of correct identification, calculated as the mean of the individual probabilities of correct identification, *i.e.* the mean of the values on the leading diagonal; (2) the confidence of identification, $p(x=j|y=j)$, which is the probability that a pattern identified as belonging to class j, actually does belong to class j, calculated as

$$\frac{M_{ii}}{\sum_{i} Mij}$$

(Boddy *et al.*, 1994a), assuming the *a prior* values for each pattern are equal.

2.7.4.2 Rejection of Unknowns

A network's identification success is not the only assessment of how well it performs as a pattern recognition system. In many instances the application for which a neural network has been trained is within a defined region, where any classes present in the testing data have already been incorporated when training the network. However, in many real world applications this is not the case, and an unknown number of ambiguous classes may be present (Morris & Boddy, 1996). Although an approximation of what exists may be sufficient, there will be times when absolute identification is required and any questionable areas must be disregarded, rather than identified incorrectly. This procedure will increase confidence of identification at the expense of overall number of patterns identified.

When considering unknowns it is unrealistic to incorporate an additional class to identify them. Not only would it need to include all biological variations of every species in existence, but also a representation of noise (*e.g.* debris, dead cells, *etc.*). Instead, threshold values are imposed on network parameters, in an attempt to exclude any novel data for whom network generalisation is poor.

For both paradigms two constraints were used to investigate unknown and known data rejection.

1. Rejection if the highest valued output is less than a threshold, T1, where T1 ranges from 0 to 0.9 in intervals of 0.1
2. Rejection if the difference between the two highest outputs is less than a threshold, T2, where T2 ranges from 0 to 0.9 in intervals of 0.1

2.8 Results

Both the MLP and RBF networks discriminated between the mixed species set, with an overall percentage of correct identification of 96.1% for the optimum MLP, and 96.9% for the RBF (Table 2.4). Variation in the number of hidden layer nodes for this set produced only a marginal increase in performance for the RBF, with a 0.6% difference between the minimum and maximum number of nodes. The performance of the MLP rose by 6% when the number of nodes was increased from 5 to 20, after which point it dropped slightly. Correct identification and confidence of identification for individual species was greater than 85% for both paradigms (Table 2.5).

Network performance for the Dinoflagellate group was not as good, with a maximum value of correct identification of 83.2% for the RBF and 77% for the MLP (Table 2.4). Varying the number of hidden nodes from 12 to 96 in the RBF and from 5 to 20 in the MLP, resulted in an increase of 7.6% in both networks. Individual species identification and confidence of identification varied for both paradigms, with the optimum RBF being superior approximately 70% of the time (Table 2.6). For both data sets, the MLP required 20 nodes to perform to the same level as the RBF employing only 12 nodes.

With a rejection criterion imposed on the highest valued output of both paradigms, rejection of known species from the mixed data set remained below 10% for a value of 0.6 and increased to between 28% and 30% for a value of 0.9 (Fig. 2.6a). Rejection of unknown species remained at less than 5% by both networks for a threshold of 0.2. As this threshold was increased to 0.5 the MLP rejected only 12% of unknown species, whereas the RBF rejected 68% (Fig. 2.6a). For both architectures, rejection of known and unknown species from the Dinoflagellate group was approximately zero to a threshold of 0.2, beyond this, rejection of known species through the RBF, and both known and unknown species through the MLP, followed a similar path (Fig. 2.6b). Unknown species rejection by the RBF was slightly higher at all threshold values. As the rejection threshold increased to 0.5, identification dropped for individual species, producing overall success of 74.6% and 65.3% for the RBF and MLP networks respectively (not shown).

When the rejection criterion was set to the difference between the winning node and second highest, at a threshold of 0.2 over 50% of unknown species were rejected from the mixed data set by the RBF, and 23% by the MLP (Fig. 2.7a). Rejection of known species through both networks were similar, with less than 10% rejection at a threshold of 0.4. Rejection of unknown species from the Dinoflagellate set were very close for both architectures, with approximately 23% being rejected at a threshold of 0.1 (Fig. 2.7b). However, known species rejection for both paradigms was also high, with the MLP rejecting approximately 10% more than the RBF at each value. At a value of 0.5, overall identification success were 53.4% and 41.9% for the RBF and MLP respectively.

Confidence of identification increased for most species in both network paradigms, irrelevant of constraint type.

Table 2.4 Overall percentage of correct identification for the Mixed and Dinoflagellate data set, by each paradigm as hidden layer nodes are increased. The optimum networks chosen for further study, are selected from test data performance and indicated by '*M' and '*D' as optimum for the Mixed and Dinoflagellate data sets respectively. Individual species identification for the optimum networks are shown in Table 2.5 (Mixed Species) and Table 2.6 (Dinoflagellates).

Network Paradigm		No. of Hidden Layer Nodes	Overall Correct Identification			
			Mixed Species		Dinoflagellates	
			Training	Test	Training	Test
MLP		5	91.1	90.1	69.7	69.4
		10	95.6	95.5	77.1	75.5
	*M *D	20	96.1	96.1	78	77
		30	95.9	95.7	78.1	77.1
		40	95.9	95.9	78.2	75.9
RBF		12	96.4	96	77.2	76.4
		24	96.8	96.5	81.8	80
		36	97.2	96.6	83.3	82
		48	97.1	96.6	84.1	82
	*M	60	97.7	96.9	84.4	83
		72	97.7	96.8	84.7	83.2
		84	97.8	96.7	84.6	83.1
	*D	96	97.9	96.6	84.7	83.2

Table 2.5 Individual percentage of correct identification (corr) and confidence of identification (conf) for species in the Mixed data set by the optimum networks chosen *i.e.* MLP (96.1%) and RBF (96.9%).

			MLP		RBF	
Species Name	**Order**	**Size μm**	**% Corr**	**Conf**	**% Corr**	**Conf**
Amphora coffaeformis	Bacillariophyceae	10-20	93	97.6	94.3	97.9
Chaetoceros calcitrans	Bacillariophyceae	4-6	98.3	98.7	98.7	99.3
Chroomonas sp.	Cryptomonadida	8-10	94.3	97.6	94.3	97.7
Emiliania huxleyi B11	Prymnesiida	5-7	98	97.7	97	99.3
Gyrodinium aureolum	Dinoflagellida	35-45	97.3	94.5	97	95.8
Hemiselmis virescens	Cryptomonadida	5-8	99.7	97.4	98.3	97
Micromonas pusilla	Prasinomonadida	1-3	100	89.8	100	95.6
Prorocentrum balticum	Dinoflagellida	9-15	89.7	92.8	95.3	90.4
Pelagococcus subviridis	Chrysomonadida	2-3	91.3	98.2	93.3	97.9
Plagioselmis punctata	Cryptomonadida	6-9	95.7	97.3	97.3	98
Rhodomonas sp.	Cryptomonadida	8-13	96.7	95.7	98.3	95.2
Tetraselmis tetrathele	Prasinomonadida	10-16	98	95.8	98.7	99.4

Table 2.6 Individual percentage correct identification (corr) and confidence of identification (conf), for species in the Dinoflagellate data set by the optimum networks chosen *i.e.* MLP (77%) and RBF (83.2%).

			MLP		RBF	
Species Name	**Order**	**Size μm**	**% Corr**	**Conf**	**% Corr**	**Conf**
Amphidinium carterae	Dinoflagellida	15-20	78.8	78.1	85.7	88.1
Aureodinium pigmentosum	"	7-12	80.7	82	85.7	92.5
Gymodinium simplex	"	6-10	94.3	86.8	92.3	91.2
Gymodinium veneficum	"	9-16	31.7	62.9	68.3	69.7
Gymnodinium vitiligo	"	7-22	77.7	59.1	72.3	72.4
Heterocapsa triquetra	"	15-27	83.7	77	82.7	84.6
Prorocentrum balticum	"	9-15	72	80.3	84	79.4
Prorocentrum micans	"	30-40	91.3	70.6	92	74.6
Prorocentrum minimum	"	16-18	55	73.3	67	81.4
Prorocentrum nanum	"	8-10	85.7	81.6	91.7	89
Scrippsiella trochoidea	"	30-42	87.3	92.9	92	93.1
Gyrodinium aureolum	"	35-45	88.7	81.3	90.3	90.4

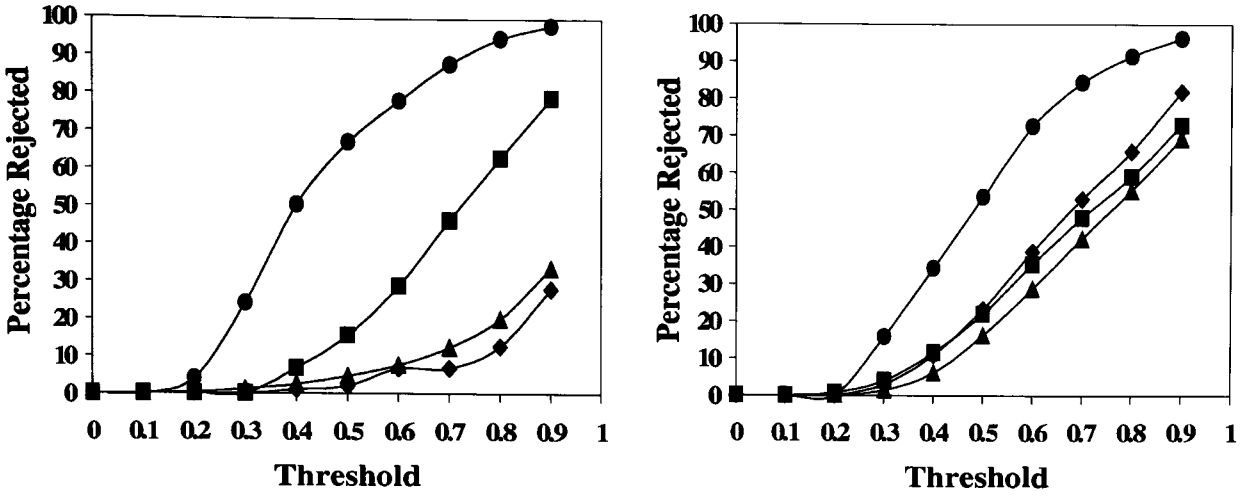


Figure 2.6 Overall percentage of species rejected when a threshold, T1, was imposed upon the highest valued output. • Overall rejection of unknown data by the optimum RBF network, ■ Overall rejection of unknown data by the optimum MLP network, ▲ Overall rejection of known data by the optimum RBF network, ♦ Overall rejection of known data by the optimum MLP network. (a) the mixed species set and (b) the Dinoflagellate set.

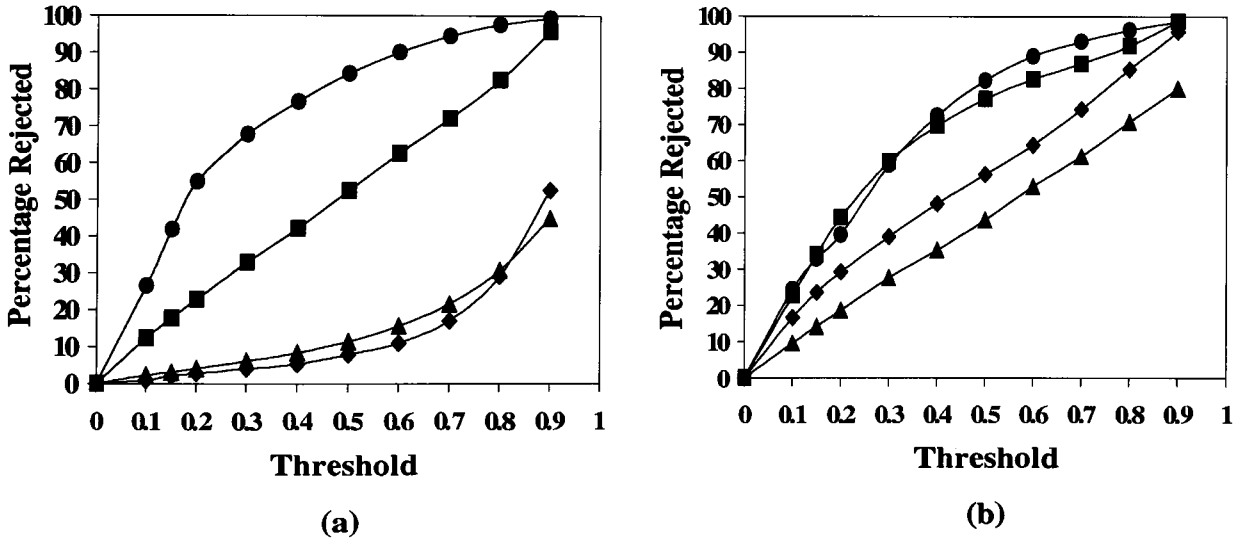


Figure 2.7 Overall percentage of species rejected when a threshold, T2, is imposed upon the difference between the winning and the second highest nodes, • Overall rejection of unknown data by the optimum RBF network, ■ Overall rejection of unknown data by the optimum MLP network, ▲ Overall rejection of known data by the optimum RBF network, ♦ Overall rejection of known data by the optimum MLP network, (a) the mixed species set and (b) the Dinoflagellate set.

2.9 Discussion

The mixed data set exhibits areas of visual distinction between many of the probable clusters (Fig. 2.6b), indicating little overlap between species and, accounting for the high discriminatory power of both paradigms. The performance of the 'optimum' RBF, for the Dinoflagellate data group, using 96 kernel centres, is only 2% higher than a network using 36 centres. In this instance it is chosen as the 'optimum' network as class numbers were few and time is irrelevant. However, in networks trained for a high number of classes, a trade off between a marginal rise in performance and the number of basis functions may take place, in order to reduce computational time and complexity. Additionally, in some cases increasing node numbers can have a more detrimental effect on performance, especially evident in the MLP. The probability of correct identification for individual species varies for both paradigms. However, some correlation can be seen between the networks, where species identified poorly by the RBF also have low identification values by the MLP. A common species to both sets, *Prorocentrum balticum*, indicates the problems in identification when overlap increases. The identification of the species within its own taxonomic group (*i.e.* Dinoflagellate), is at least 10% lower than when present in the mixed data set. This can be attributed to its misidentification with species from its own genus, and is discussed further in Chapter 3. Confidence of identification of individuals is generally higher in the RBF, indicating less cases of misidentification between species than were apparent in the MLP (Tables 2.5 & 2.6). When considering the mixed data set, where visual separation is possible, the hyperplane boundaries of the MLP are as effective as the hyperellipsoids of the RBF. However, the superiority of the RBF is indicated when identifying overlapping data distributions, where its ability to form complex decision boundaries allows it to approximate the data better than the MLP (Section 2.4.2.2). Although introducing a second hidden layer into the MLP will enable disjoint decision regions, it will require an assumption about the network's architecture. An increased number of layers, and therefore nodes, may restrict generalisation by the network, causing unpredictable behaviour, it may increase training times and will not necessarily improve performance (de Villiers and Barnard, 1992; Wilkins *et al.*, 1994b).

The RBF rejects a greater number of unknowns from both data sets, irrelevant of constraint type, with less being rejected in the overlapping data set. The rejection of

known species when the classes are distinct (Fig.2.7a), was relatively small and similar for both networks when the threshold T2 was imposed. However, as data overlap intensifies (Fig. 2.7b), a high number of knowns are rejected by the RBF, and more knowns than unknowns are rejected by the MLP. This results in a very poor overall identification at a very early threshold value. This threshold, although apparently effective with distinct data, must be considered in a real world example where ease of seperability will be low. In this instance, the performance of the constraint is very poor, indicating its unsuitability.

The formation of the decision boundaries in an MLP governs its inadequacy in rejecting novel data (Section 2.4.2.2). For example, Figure 2.8 shows a two-dimensional scatter plot for two of the species in the mixed data set, A and B (*Chroomonas* sp. and *Chaetoceros calcitrans* respectively, although species type is not significant here). The decision regions shown are hypothetical, and serve only to demonstrate the relationship between an imaginary unknown data point *U*, and the possible boundaries created by both paradigms. Identification of a pattern by an MLP, is dependent upon which side of a hyper-plane boundary the pattern falls. As these linear decision boundaries are infinite, any test pattern will have a notable output from the network, despite location in the data space. Although this gives the network the ability to generalise in sparser areas, the output values for unknown patterns (*e.g.U* in Fig. 2.8) may be of similar magnitude to that of the known data. Inevitably, an unknown pattern will be assigned a class membership, and exclusion of the pattern by threshold imposition may be at the expense of known data rejection. Theoretically, surplus nodes could be used to represent closed finite decision boundaries. However, the gradient descent, error reduction strategy of the MLP, means this is not guaranteed unless the network reduces its output error in the process, and as noted, unknown formation of boundaries in empty or sparse areas may have adverse effects on the network (Wilkins *et al.,* 1994b). Conversely, as the hyperellipsoidal (or hyperspherical) boundaries of an RBF are finite and localised, the hidden node outputs decrease to 0 as the distance from the centre of the kernel increases. Therefore, the output value of a basis function (kernel) for a particular pattern, will be in relation to its proximity to the kernel centre, which in the case of a distinct unknown (*e.g.U* in Fig. 2.8) should be negligible. The summation values at the output layer for a novel pattern, will then be of a smaller magnitude to that of the known data points, and a threshold value would reject them without rejecting as many knowns. This distinction also offers an alternative

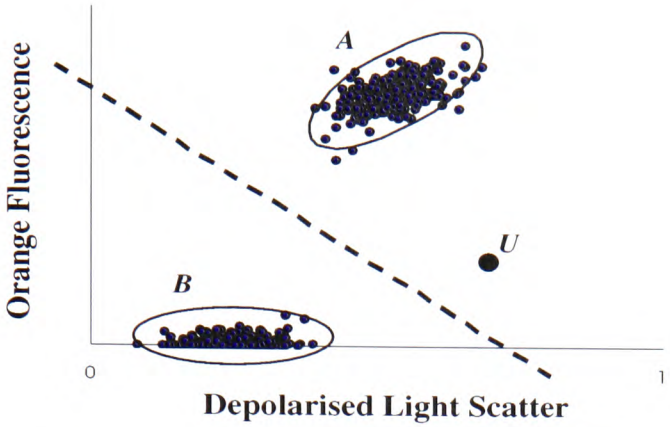


Figure 2.8 Scatter plot showing the position of an imaginary unknown data point (*U*) in relation to the hypothetical boundaries created by the two network paradigms. The dotted line shows a possible location for an infinite hyperplane in the MLP, while the solid lines depict possible finite elliptical boundaries of the RBF.

rejection criterion based on hidden layer node outputs, which will be discussed further (Chapter 3).

The location in data space of actual unknown species, whether they are laboratory grown or field samples, will of course not be as distinct as data point *U*. Therefore, introducing either constraint to a network will inevitably mean exclusion of those known species which fall below the threshold, or the inclusion of unknowns that have some resemblance to the knowns. This will naturally reduce the individual and overall probability of correct identification. This is more evident in the MLP, where identification successes were lower than for the RBF network. An increase in overall confidence of identification is apparent in both networks, where threshold imposition increases confidences for most individual species, due to the exclusion of ambiguous patterns. However, for some species this increase was small, indicating the incorrect identification of a number of unknown species as those present in the data sets; a trait again more evident in the MLP network.

2.10 Conclusion

The aim of this preliminary study was to exemplify the RBFs superiority over the MLP for data of this nature. Initially, the RBF boasts shorter training times than the MLP. Although not an issue with such small data sets, it is a distinct advantage in areas where networks may require constant re-training in real time, a prerequisite of this thesis. Secondly, the phytoplankton data used for this research, represents only a small laboratory grown subset of what is found in the field. In a natural environment, variation and overlap will be greatly increased, with some species exhibiting multi-modal data. The formation of convex continuous decision regions by the MLP, restricts its identification abilities to relatively simple, linearly separable data. However, the RBF is capable of forming complex non-linear decision regions, making it more suitable to model the distribution of this data. Finally, the formation of the decision boundaries has a considerable affect upon a network's ability to reject novel data. Using the more suitable threshold (*i.e.* highest valued output), both networks were capable of excluding unknowns, but the rejection by the MLP was continually lower than that of the RBF, with the reverse being true for knowns. These studies indicate the advantages of the RBF over the MLP for data of this nature, and it is therefore the chosen paradigm for the areas investigated in this thesis.

3 Multi-Class RBF Networks for Phytoplankton Analysis

3.1 Introduction

RBF networks were shown to be superior to MLPs, at least for 12 species, in Chapter 2. This, however, is not a realistic field number and the problem of scaling up is not a simple one. In a field environment the number of species will naturally be greater and data acquisition more difficult. Training data therefore, may not always be available in equal abundance, resulting in the possibility of an imbalanced representation of particular species.

It has been suggested, that an imbalance in event numbers may result in low error convergence of the subordinate class, thereby affecting overall performance (Anand *et al.*, 1993). Thus, this chapter firstly investigates the performance of the RBF, as the number of species and events per species are gradually increased, for both balanced and imbalanced data sets; imbalanced data sets are discussed further in Chapter 4.

Secondly, a more detailed evaluation of the RBF architecture is performed for a large data set. This includes, improving network performance by combining data for species with overlapping flow cytometric signatures, and evaluating the multi-class network's ability to reject unknown species.

3.2 Training Set Size

Although the analysis of large data sets has received some attention from neural network research, it has been applied primarily to data dimensionality (Raudys & Pikelis, 1980; Chandrasekaran & Jain, 1975,1977; Jain & Chandrasekaran, 1982; Fukunaga & Hayes, 1989). High quantity of data is an area still largely under statistical investigation, with only a few algorithms able to cope with massive numbers of events and high-dimensional vectors. In particular fields, much of the data gathered may be redundant and therefore only a suitable subset is required. In these cases the training set comprises only relevant information, selected by, for example, the D-optimality Criterion (Choueiki & Mount-Campbell, 1999). In the case of flow cytometric data, the potentially wide biological variation may require analysis of a large, possibly imbalanced, number of cells from each population.

As class number (species) increases, the possibility that species present in the training data occupy the same or similar areas of the sample space multiplies, making non-

linear partitioning a problem. As long as the classes forming the data set are separable, identification is achieved with high success. However, this only occurs when using a small number of species with distinct flow cytometric signatures (Frankel *et al.*, 1989,1996, Smits *et al.*,1992, Balfoort *et al.*,1992). As the number of classes is increased, overlap can be considerable and the distinction between species (and therefore ease of identification) becomes particularly complex. Thus, scaling up is not easy.

When data size is being investigated, the quantity of data, coupled with the network's architecture, must be considered prior to training. The number of hidden layer nodes can have various effects on a network's ability to generalise. Baum and Haussler (1989) suggest a condition for determining data size for an MLP, based on the number of synaptic weights, W, the number of hidden layer nodes, M and the fraction of errors permitted on test, ε, given by:

$$N \geq \frac{32W}{\varepsilon} \ln\left(\frac{32M}{\varepsilon}\right)$$

However, this equation produces a worst-case scenario (Haykin, 1994), and does not consider the nature or complexity of the data, or the possibility that data may not be available in equal abundance. As each application is unique, a more practical approach through trial and error is advocated.

Many of the species present in the database are represented by approximately 10,000 events. To incorporate this much data for 62 classes (or indeed more) would be nonsensical, as the computational effort and time required would make the procedure unrealistic, and in fact is unnecessary, as is shown (Al-Haddad *et al.*, 2000; Section 3.4). The number of events must allow adequate representation of species variation, in order for the network to generalise on unseen data. This optimum number must be high enough to avoid any memorisation of data, but low enough to avoid any high degree of computation, thereby keeping training times to a minimum.

3.3 Imbalanced Training Sets

When considering imbalanced event numbers, the effect of training data size on the network's performance can be more pronounced. If a species is inadequately represented by event numbers, presentation at every epoch of the same limited amount of data, may cause the network to memorise the particular representation of the subordinate class,

rendering it unable to generalise. When a class is inadequately represented, the basis functions may be unable to approximate its distribution, causing error values to oscillate. In order to compensate for imbalanced training data, Richard & Lippman (1991) suggest adjusting network outputs using training data and correct class probabilities. It has already been noted that with appropriate architecture and algorithm, and sufficient training data, the network outputs estimate Bayesian probabilities (Richard and Lippman 1991). The output is implicitly the *a priori* class probability for class *j*, *i.e.* (*Pr (class j)*), multiplied by the class likelihood, *i.e.* (*Pr (x|class j)*), divided by the unconditional input probability (Chapter 1, Section 1.6.1). Richard and Lippman (1991) suggest that as the *a priori* term is simply a coefficient, it is possible to adjust it to counteract the imbalance producing a corrected identifier. An adjustment can be made dependent upon the ratio of training data frequency to that of test or field data. This scaling is simply performed by first multiplying the network output by the correct class probability and dividing by the training data class probability. This is detailed empirically in Section 3.4.3.

3.4 Experimental Procedure

The following experiments establish the effect on network performance, of both size of training data set and of imbalanced number of events. Data preparation and network training and testing were carried out as described in Chapter 2, Sections 2.6 & 2.7. The number of kernels (hidden layer nodes) defined for both balanced and imbalanced networks, for those networks trained using an LVQ placement strategy, were prior to training, and an optimal subset were automatically selected via the orthogonal least squares procedure (Chen *et al.*, 1991; Chapter 2 - Section 2.4.2.1).

3.4.1 Balanced Event Numbers

For the first study, an initial set of 20 species were selected from the database. This included the Cryptomonads, as research indicated their high identification success (Section 3.8.1.2), and a random selection of separable species. Nine training files were created, each containing the 20 species, with a random selection of 10, 25, 50, 100, 200, 400, 600, 800 and 1000 events per class. For each of the 9 training files, 6 RBF networks were produced, with a total of 10 kernels for the first network, and 1 to 5 kernels per class (increasing in steps of 1) for the remaining 5. A random kernel placement strategy was

used for the first network, while the others employed the LVQ method of placement (Kohonen, 1988, 1990; Chapter 2 - Section 2.4.2.1). All networks used a Mahalanobis distance metric.

Twenty randomly chosen species were added to the first set to create 40 classes, and 9 training files were created as before. The topology of the RBF networks produced were identical to those for the 20 species, with the exception of the first network which contained 20 randomly placed kernel centres.

Finally, a further 20 species were added to create 60 classes and the process repeated with the initial network containing 30 randomly placed centres.

Table 3.1 indicates membership to each of the three sets as A (20 species), B (40 species) or C (60 species).

3.4.2 Imbalanced Event Numbers

For the second study, an optimum number of nodes and events were determined from the results of study one. The same species were used to construct the three individual data sets, containing 20, 40 and 60 classes (species) respectively. For each of the three sets, 30 imbalanced data files were constructed consisting of an x:y ratio of event numbers. Explicitly, x events per species for classes 1 to ½n and y events per species for classes (½n+1) to n, where n is the total number of species and x and y are one of 400, 200, 100, 50, 25 or 10 events per class. To ensure that class representation was not biased through easily discriminable species possibly dominating one half of the data set, the uneven split was then reversed, so a x:y split was then trained as a new network containing a y:x split. Table 3.1 indicates membership to either half of each of the three data sets as A1 or A2 (20 species), B1 or B2 (40 species) or C1 or C2 (60 species). Networks were trained for each data set using 3 hidden layer nodes per class, employing a Mahalanobis distance metric.

For both studies, all networks were trained three times from different initialisation points, and were tested using an independent test file consisting of equal events for all species present in each of the respective training sets (Chapter 2 - Section 2.7.4.1).

Table 3.1 Membership to each of the three data sets, for the balanced event numbers, are indicated as A, B and C for the 20, 40 and 60 species sets respectively. The imbalanced sets are represented by 1 or 2, indicating class membership to either the first or second half of set A, B or C. Individual species identification and misidentification are shown for an optimum RBF network trained on 60 species.

Taxonomic Group and Order	Species Name	Size (μm)	Data Set	% correct i.d.	Species Misidentified >10%
Cryptomonads					
Cryptomonadida	*Chroomonas sp.*	8-10	A1 B1 C1	93.5	
"	*Chroomonas salina*	5-12	A1 B1 C1	94	
"	*Cryptomonas appendiculata*	15-25	A1 B1 C1	98.5	
"	*Cryptomonas calceiformis*	10-15	A1 B1 C1	89	
"	*Cryptomonas maculata*	12-20	A1 B1 C1	94	
"	*Cryptomonas reticulata*	18-25	A1 B1 C1	95	
"	*Cryptomonas rostrella*	16-25	A1 B1 C1	98.5	
"	*Hemiselmis brunnescens*	5-8	A1 B1 C1	64	*H. rufescens (33%)*
"	*Hemiselmis rufescens*	4-9	A1 B1 C1	64.5	*H. brunnescens (30%)*
"	*Hemiselmis virescens*	5-8	A2 B2 C1	93	
"	*Plagioselmis punctata*	6-9	A2 B2 C2	92.5	
"	*Rhodomonas* sp.	8-13	A2 B2 C2	87.5	
Flagellates					
Prasinomonadida	*Micromonas pusilla*	1-3	A2 B2 C1	99.5	
"	*Nephroselmis pyriformis*	4-7	B2 C2	70	*N. rotunda (23%)*
"	*Nephroselmis rotunda*	6-8	B2 C2	55.5	*N. pyriformis (31%)*
"	*Pyramimonas grossii*	5-10	C2	71	
"	*Pyramimonas obovata*	4-8	C2	68	*P. lutheri (11%)*
"	*Tetraselmis impellucida*	11-19	A2 B2 C2	94	
"	*Tetraselmis suecica*	6-15	B2 C2	88.5	
"	*Tetraselmis verrucosa*	3-11	C2	64	
"	*Tetraselmis tetrathele*	10-16	A2 B2 C2	95.5	
"	*Tetraselmis striata*	6-8	B2 C2	74.5	
Volvocida	*Chlamydomonas reginae*	11-20	B1 C1	89	
"	*Chlorella salina*	4-8	C1	54	*C. polylepis (12%)*
"	*Dunaliella minuta*	3-12	B1 C1	62.5	
"	*Dunaliella primolecta*	5-12	B1 C1	88.5	
"	*Dunaliella tertiolecta*	6-12	B1 C1	82.5	
"	*Stichococcus bacillaris*	5-8	C2	63	
Rhodomonadida	*Porphyridium pupureum*	4-6	A2 B2 C2	96	
"	*Rhodella maculata*	7-24	A2 B2 C2	91.5	
Chrysomonadida	*Ochromonas* sp.	3-12	C2	39.5	*P. parvum (10.5%)* *S. costatum (24%)*
"	*Pelagococcus subviridis*	2-3	A2 B2 C2	86.5	
"	*Pseudopedinella* sp.	8-10	C2	79.5	
Prymnesiomonads					
Prymnesiida	*Chrysochromulina camella*	6-12	B1 C1	89	
"	*Chrysochromulina chiton*	5-9	C1	60.5	*C. polylepis (20%)*
"	*Chrysochromulina cymbium*	6-10	C1	32.5	*C. polylepis (14%)* *C. chiton (13%)* *O. neopolitana (15%)*

Table 3.1 continued....

Taxonomic Group and Order	Species Name	Size (μm)	Data Set	% correct id	Species Misidentified >10%
Prymnesiomonads					
Prymnesiida	*Chrysochromulina polylepis*	6-8	C1	63.5	*C. chiton (17.5%)*
"	*Pleurochrysis carterae*	10-18	B2 C2	90	
"	*Emiliania huxleyi* B11	5-7	A1 B1 C1	99.5	
"	*Emiliania huxleyi* 92	5-6	C1	78.5	
"	*Ochrosphaera neopolitana*	8-10	C2	45.5	
"	*Pavlova lutheri*	4-6	C2	72	*Pyramimonas obovata (14%)*
"	*Phaeocystis pouchetii*	3-6	C2	56.5	*Chlorella salina (11%)*
"	*Prymnesium parvum*	8-10	C2	75.5	
Diatoms					
Bacillariophyceae	*Chaetoceros calcitrans*	4-6	B1 C1	87	
"	*Phaeodactylum tricornutum*	8-35	A2 B2 C2	94.5	
"	*Skeletonema costatum*	3-5	C2	80	
"	*Thalassiosira weissflogii*	12-20	C2	93.5	
Dinoflagellates					
Dinoflagellida	*Aureodinium pigmentosum*	7-12	C1	86.5	
"	*Gymnodinium micrum*	8-15	B1 C1	74	
"	*Gymnodinium simplex*	6-10	C1	62.5	
"	*Gymnodinium veneficum*	9-16	B1 C1	46.5	*G. vitiligo (32%)*
"	*Gymnodinium vitiligo*	7-22	B1 C1	63	*G. veneficum (21%)*
"	*Gyrodinium aureolum*	35-45	B1 C1	88.5	
"	*Heterocapsa triquetra*	15-27	B2 C1	79.5	
"	*Prorocentrum balticum*	9-15	B2 C2	66.5	
"	*Prorocentrum micans*	30-40	B2 C2	74	
"	*Prorocentrum minimum*	16-18	B2 C2	58	
"	*Prorocentrum nanum*	8-10	C2	63	
"	*Scrippsiella trochoidea*	30-42	B2 C2	42	*D. minuta (16.5%)* *D. tertiolecta (17.5%)*

3.4.3 Compensation for Imbalanced Event Numbers

In the subsequent analysis, the training data class probabilities were estimated as the occurrence of events for each class in the data set. For example, for the 20 class data set, with 10 events per class for the first 10 species and 400 events for the remaining 10, the total number of events is given by:

$$\begin{matrix}\textit{Events per class} \\ \textit{for first } 10 \textit{ classes}\end{matrix} \times \frac{\textit{classes}}{2} + \begin{matrix}\textit{Events per class for} \\ \textit{remaining } 10 \textit{ classes}\end{matrix} \times \frac{\textit{classes}}{2}$$

$$\textit{i.e. } 10\times\frac{20}{2}+400\times\frac{20}{2}=4100$$

The training class probability for the classes with 10 events each is therefore,

$$\frac{\textit{no. of events per class}}{\textit{total no. of events}} = \frac{10}{4100}$$

and for those with 400 events per class, $\frac{400}{4100}$

The probabilities for the remaining combinations of event numbers are shown in Table 3.2. Richard and Lippman (1991) note that for certain applications the correct class probability can be discovered from the relevant statistics, generated either heuristically or empirically, depending on the area. In this instance, the correct class probabilities are estimated assuming classes are represented by equal event numbers, *i.e.* 1/20, 1/40 or 1/60, for the 20, 40 and 60 class data sets respectively. The adjustments were performed for each of the three data sets using the combinations of events indicated in Table 3.2.

3.5 Results

3.5.1 Balanced Event Numbers

The primary observation was the decrease in overall success as the number of species increased from 20 to 60 (Fig. 3.1a, b, & c). For all three data sets, when the number of events was small, the difference between performance of training and test data was high. For 10 events per class identification success was between 10% and 20% higher for the training data, dropping to a difference of 3%-6% as the events were increased to 50 per class. From 100 to 1000 events the 20 class networks exhibit little improvement, with a difference of 1% to 2% between training and test data for all numbers of kernel centres (92%-94% correct; Fig. 3.1a). With the exception of the network trained using one node

Table 3.2 Combinations of event numbers per class for each half of the imbalanced data sets as well as respective training data class probabilities, used to adjust network outputs.

Classes	Events per class	20 Classes		40 Classes		60 Classes	
		Total Events per set	Training Probs	Total Events per set	Training Probs	Total Events per set	Training Probs
1 to ½n	10	4100	$\frac{10}{4100}$	8200	$\frac{10}{8200}$	12300	$\frac{10}{12300}$
(½n+1) to n	400		$\frac{400}{4100}$		$\frac{400}{8200}$		$\frac{400}{12300}$
1 to ½n	25	4250	$\frac{25}{4250}$	8500	$\frac{25}{8500}$	12750	$\frac{25}{12750}$
(½n+1) to n	400		$\frac{400}{4250}$		$\frac{400}{4850}$		$\frac{400}{12750}$
1 to ½n	50	4500	$\frac{50}{4500}$	9000	$\frac{50}{9000}$	13500	$\frac{50}{13500}$
(½n+1) to n	400		$\frac{400}{4500}$		$\frac{400}{9000}$		$\frac{400}{13500}$
1 to ½n	100	5000	$\frac{100}{5000}$	10000	$\frac{100}{10000}$	15000	$\frac{100}{15000}$
(½n+1) to n	400		$\frac{400}{5000}$		$\frac{400}{10000}$		$\frac{400}{15000}$
1 to ½n	200	6000	$\frac{200}{6000}$	13000	$\frac{200}{13000}$	18000	$\frac{200}{18000}$
(½n+1) to n	400		$\frac{400}{6000}$		$\frac{400}{13000}$		$\frac{400}{18000}$

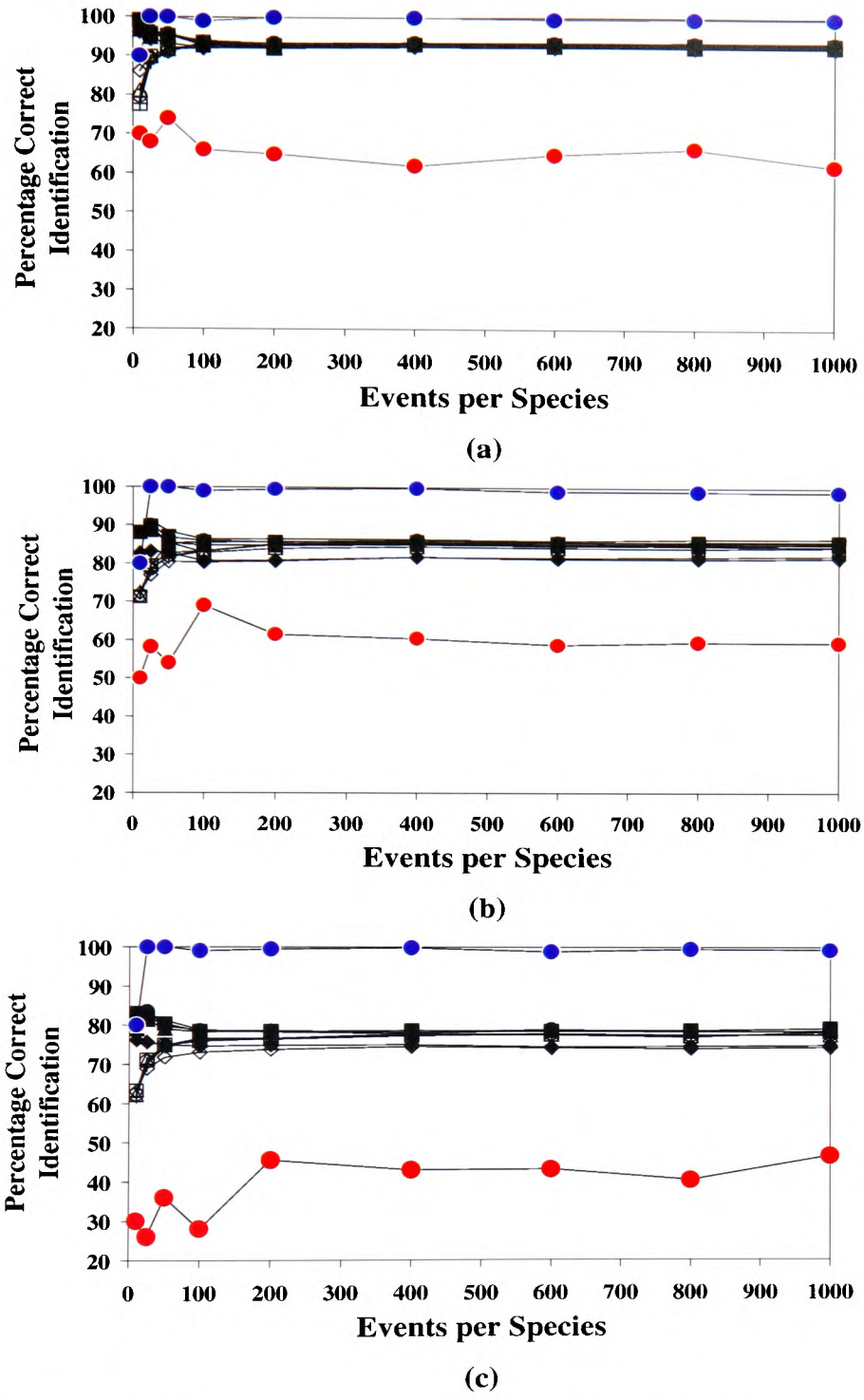


Figure 3.1 Overall percentage correctly identified for the three sets indicating both training, test and extreme species results for varying node numbers. ◆ 1 node (Training), ◇ 1 node (Test), ▲ 2 nodes (Training), △ 2 nodes (Test), ■ 3 nodes (Training), □ 3 node2 (Test), ● 4 nodes (Training), ○ 4 nodes (Test), – 5 nodes (Training), | 5 nodes (Test). (a) 20 species ● *Micromonas pusilla,* ● *Hemiselmis brunnescens*, (b) 40 species ● *Micromonas pusilla* ● *Prorocentrum minimum*, (c) 60 species ● *Micromonas pusilla* ● *Ochrosphaera neopolitana.*

per class (≈81% correct), a similar trend can be seen for the 40 class networks for events greater than or equal to 200 (84%-86% correct; Fig. 3.1b), and also for the 60 class networks with the uniformity beginning just past 300 events per class (≈74% correct for one node per class and ≈77% for 2 or more; Fig. 3.1c).

For all three data sets the species at the extreme ends of the identification range are shown, with *Micromonas pusilla* indicated as the most easily distinguishable in all cases, with a minimum identification of 95% correct, for all event numbers except 10.

3.5.2 Imbalanced Event Numbers

Overall identification success for the 20 species data set was higher than that of the 40 and 60 species sets (not shown). With the exception of y<50, identification of species 11-20 were high and consistent (Figs. 3.2a, b, c, d, e & f – N.B. Graphical representation is on event number not class numbers, in contrast to balanced events). As the event numbers for x dropped to 50 and below, performance of species 1-10 decreased. For all sets of networks the identification for species 1 to ½n decreased as the events and identification of species (½n+1) to n increased. Identification for the training data was always higher than that of the test data, with an approximate difference of 1-2% for training sets with events greater than 100, which increased to between 5% and 10% as the events dropped to 10. The results for balanced events for each of the three data sets all coincided with the positive or negative gradient of identification success, depending upon which half of the data set was being considered. The number of hidden layer nodes remaining were between 46 and 56 for the 20 species networks, 96 and 106 for the 40 species networks and between 126 and 136 for the 60 species network. The results of extreme species identification are indicated graphically showing best and worst cases for both halves of each data set (Figs. 3.3a, b, c, d, e & f).

3.5.3 Compensation for Imbalanced Event Numbers

Adjusting network outputs dramatically improves identification success for test data, exhibiting a large improvement in the subordinate classes, at the expense of a slight decrease in identification success of the dominant classes (Fig. 3.4).

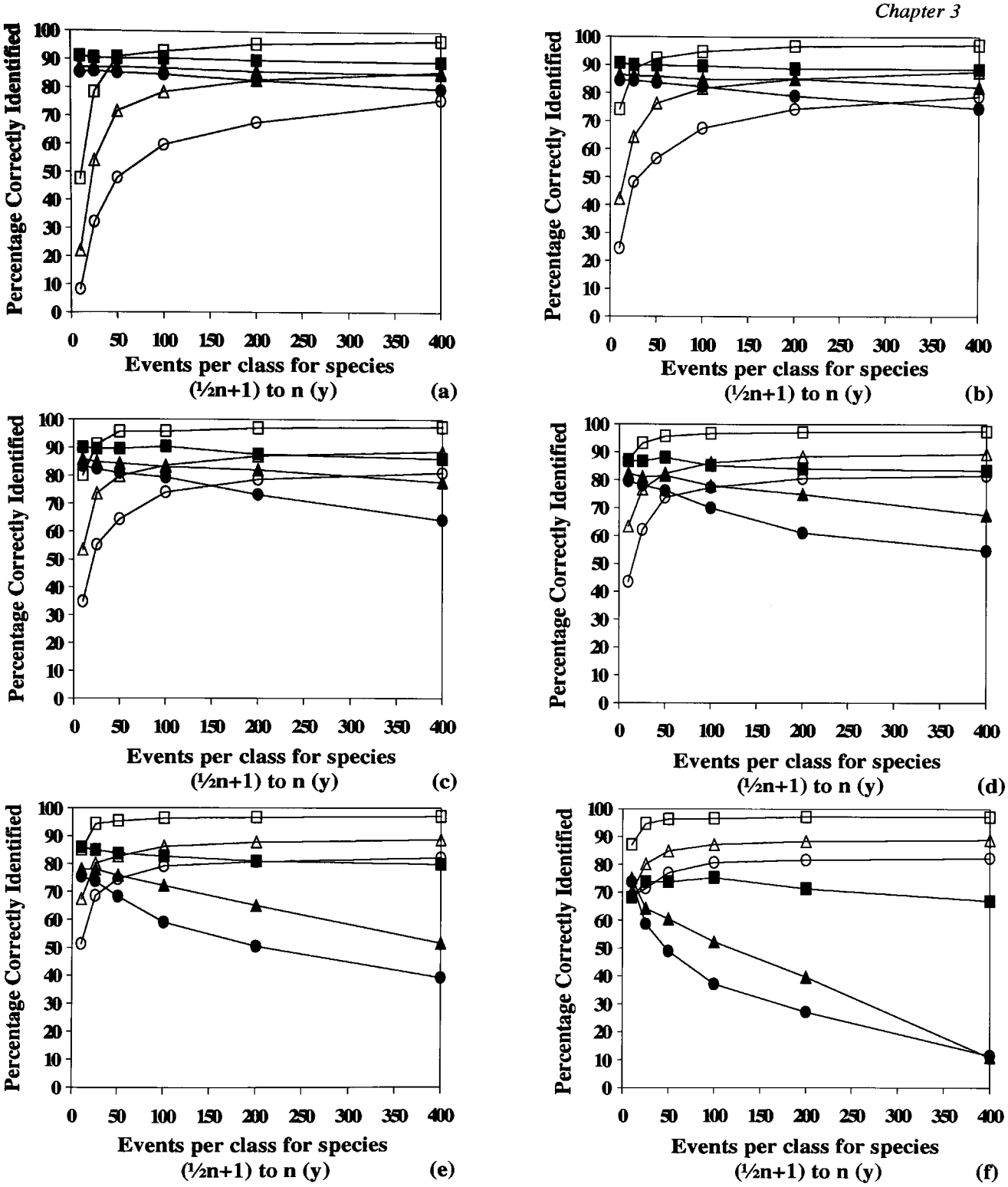


Figure 3.2 Percentage of test data correctly identified as imbalanced event numbers are altered for classes 1 to ½n (x) and (½n+1) to n (y), for each of the three data sets; ■ 20 Species (1-10), □ 20 Species (11-20), ▲ 40 Species (1-20), △ 40 Species (21-40), ● 60 Species (1-30), ○ 60 Species (31-60), (a) x = 400 events per class, (b) x = 200 events per class, (c) x = 100 events per class, (d) x = 50 events per class, (e) x = 25 events per class, (f) x = 10 events per class.

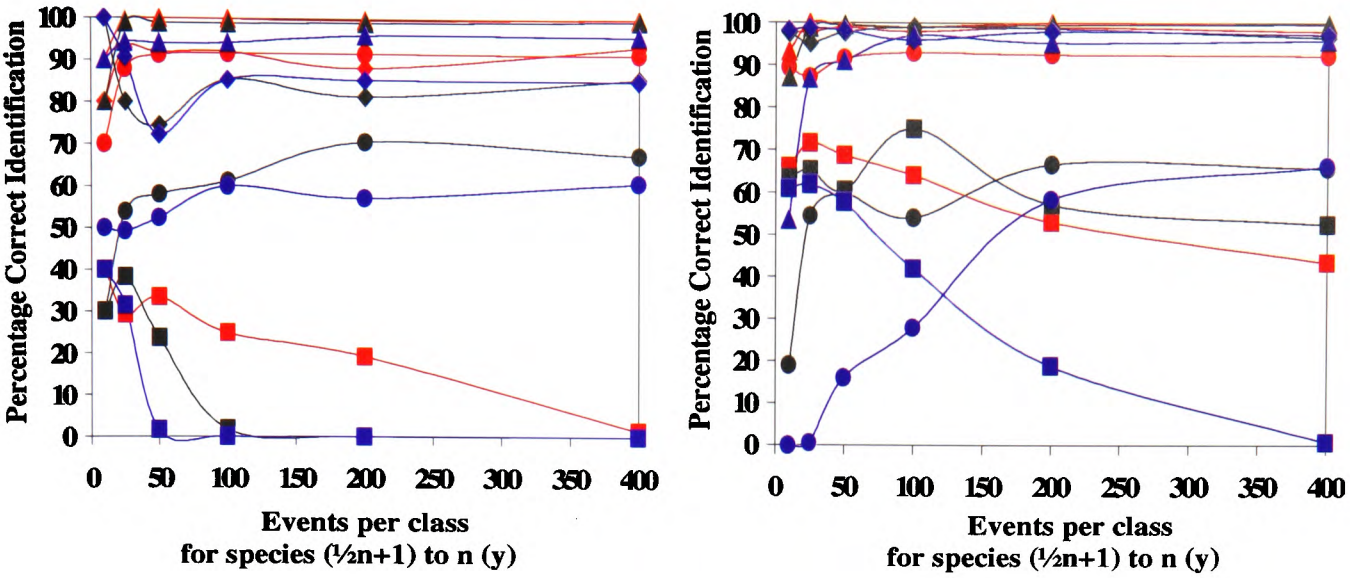


Figure 3.3 The identification of extreme species for both halves of each data set are indicated with the best for classes 1 to ½ n as ◆, worst for 1 to ½n as ■, best for (½n + 1) to n as ▲ and worst for (½n + 1) to n as ●, for the 40 species data set. The same results are plotted in red for the 20 species data set and blue for the 60 species data set. (a) 10 events in classes 1 to ½n. ◆ *Cryptomonas reticulata,* ■ *Hemiselmis rufescens,* ▲ *Tetraselmis impellucida,* ● *Pelagococcus subviridis,* ◆ *Emiliania huxleyi B11,* ■ *Gymnodinium micrum,* ▲ *Micromonas pusilla,* ● *Nephroselmis rotunda,* ◆ *Emiliania huxleyi B11,* ■ *Chrysochromulina chiton,* ▲ *Rhodella maculata,* ● *Ochromonas* sp. (b) 100 events in classes 1 to ½n. ◆ *Emiliania huxleyi B11,* ■ *Hemiselmis brunnescens,* ▲ *Micromonas pusilla,* ● *Pelagococcus subviridis,* ◆ *Cryptomonas rostrella,* ■ *Hemiselmis rufescens,* ▲ *Micromonas pusilla,* ● *Prorocentrum minimum,* ◆ *Cryptomonas appendiculata,* ■ *Chrysochromulina cymbium,* ▲ *Phaeodactylum tricornutum,* ● *Ochrosphaera neopolitana.*

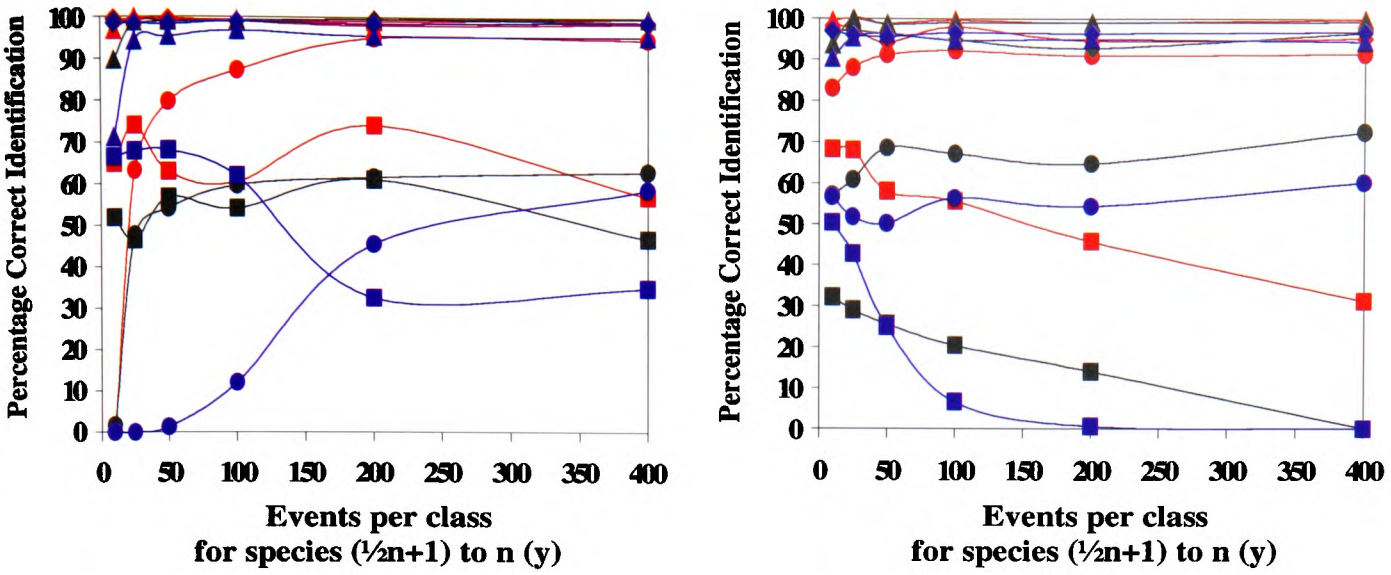


Figure 3.3 The identification of extreme species for both halves of each data set are indicated with the best for classes 1 to ½ n as ◆, worst for 1 to ½n as ■, best for (½n + 1) to n as ▲ and worst for (½n + 1) to n as ●, for the 40 species data set. The same results are plotted in red for the 20 species data set and blue for the 60 species data set. (c) 200 events in classes 1 to ½n. ◆ *Emiliania huxleyi* B11, ■ *Hemiselmis rufescens*, ▲ *Micromonas pusilla*, ● *Plagioselmis punctata*, ◆ *Cryptomonas appendiculata*, ■ *Gymnodinium veneficum*, ▲ *Micromonas pusilla*, ● *Prorocentrum minimum*, ◆ *Cryptomonas appendiculata*, ■ *Chrysochromulina cymbium*, ▲ *Tetraselmis tetrathele*, ● *Ochrosphaera neopolitan*. (d) 25 events in classes 1 to ½n. ◆ *Cryptomonas appendiculata*, ■ *Hemiselmis brunnescens*, ▲ *Micromonas pusilla*, ● *Pelagococcus subviridis*, ◆ *Emiliania huxleyi* B11, ■ *Gymnodinium veneficum*, ▲ *Micromonas pusilla*, ● *Nephroselmis pyriformis*, ◆ *Emiliania huxleyi* B11, ■ *Chrysochromulina cymbium*, ▲ *Tetraselmis tetrathele*, ● *Ochromonas* sp.

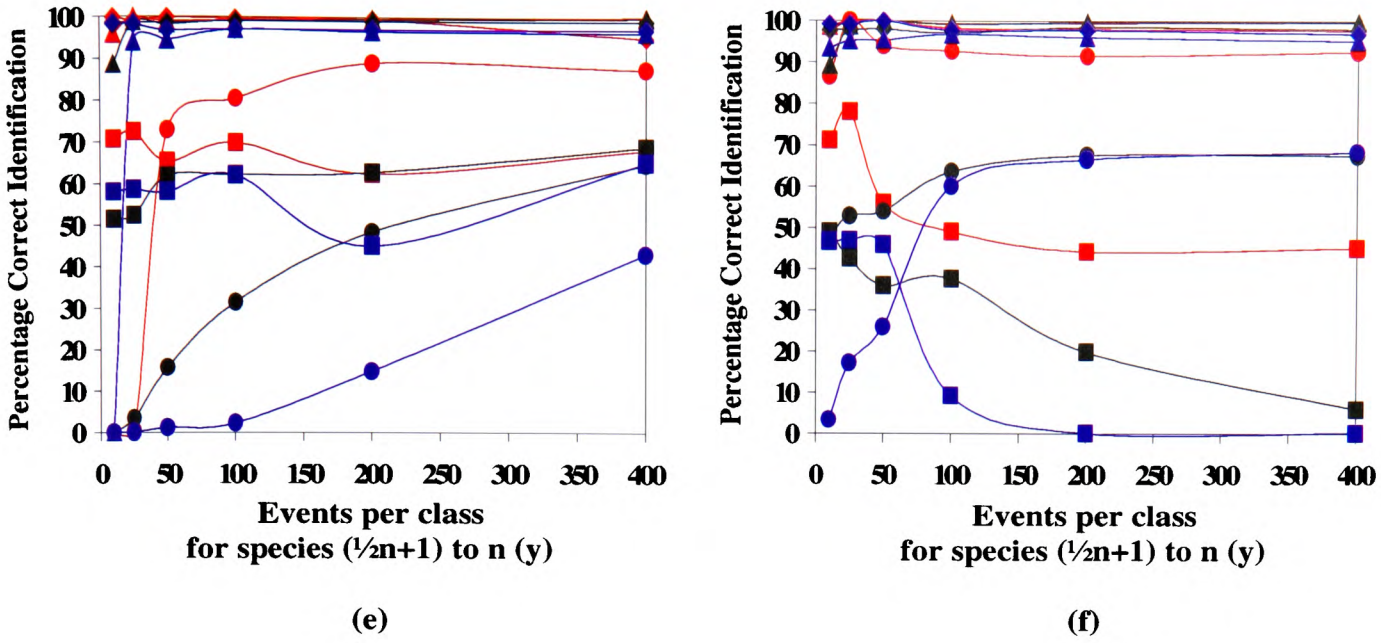


Figure 3.3 The identification of extreme species for both halves of each data set are indicated with the best for classes 1 to ½ n as ◆, worst for 1 to ½n as ■, best for (½n + 1) to n as ▲ and worst for (½n + 1) to n as ●, for the 40 species data set. The same results are plotted in red for the 20 species data set and blue for the 60 species data set. (e) 400 events in classes 1 to ½n. ◆ *Emiliania huxleyi* B11, ■ *Hemiselmis brunnescens,* ▲ *Micromonas pusilla,* ● *Plagioselmis punctata,* ◆ *Cryptomonas appendiculata,* ■ *Gymnodinium veneficum,* ▲ *Micromonas pusilla,* ● *Scrippsiella trochoidea,* ◆ *Emiliania huxleyi* B11, ■ *Gymnodinium veneficum,* ▲ *Tetraselmis tetrathele,* ● *Ochrosphaera neopolitana.* (f) 50 events in classes 1 to ½n ◆ *Emiliania huxleyi* B11, ■ *Hemiselmis brunnescens,* ▲ *Micromonas pusilla,* ● *Pelagococcus subviridis,* ◆ *Emiliania huxleyi* B11, ■ *Gymnodinium veneficum,* ▲ *Micromonas pusilla,* ● *Prorocentrum minimum,* ◆ *Micromonas pusilla,* ■ *Chrysochromulina cymbium,* ▲ *Tetraselmis tetrathele,* ● *Ochrosphaera neopolitana.*

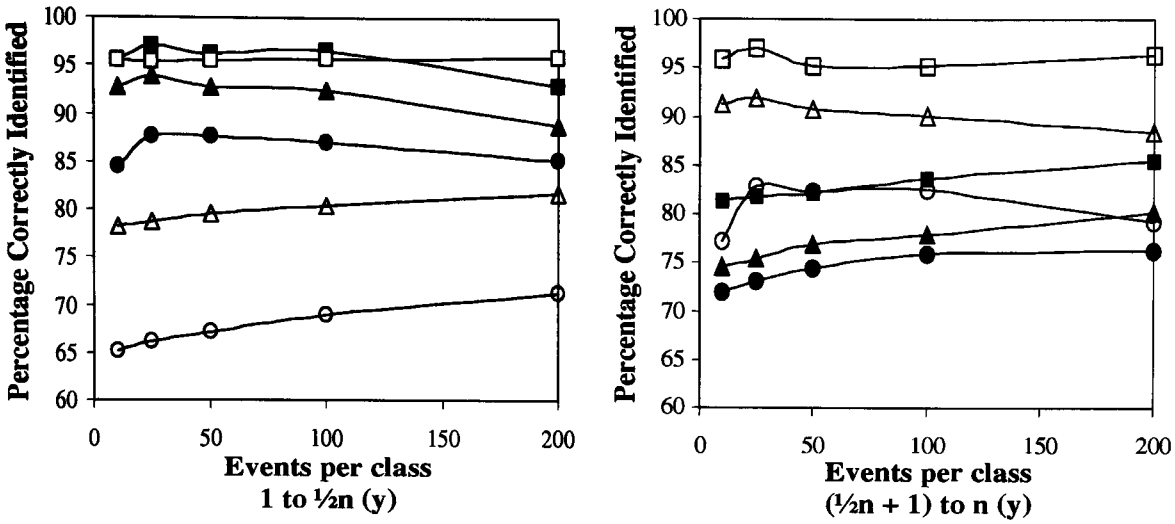


Figure 3.4 Performance identification for each half of the data sets after adjustments. 20 species, ■ classes 1-10, □ classes 11-20; 40 species, ▲ classes 1-20, Δ classes 21-40; 60 species, ● classes 1-30, ○ classes 31-60. (a) y = 400 for classes ½n + 1 to n (b) y = 400 for classes 1 to ½n.

3.6 Discussion

3.6.1 Balanced Event Numbers

As the number of classes increases, overall success drops. This is expected of data with such complex and overlapping characteristics. The difference in identification success, between the training and test data using 10 events per class, is extreme for each of the three studies, with the training data at least 10% higher. This indicates poor generalisation, where the network has memorised the training data therefore reducing its ability to recognise unseen patterns. This phenomenon is apparent to a lesser extent when the events are raised to 50 per class. Compared to the 40 and 60 class sets, the overlap of data points in hyper-dimensional space for 20 species is at a minimum, and the performance of the network using one node for each of the 20 classes is as good as those using 2 or more (Fig.3.1a). This appears to imply a simple linear discriminant function may be sufficient for this identification problem, however, this would only be true if all species distributions were Gaussian (which is not the case), and in fact as class number and overlap increases the margin drops to between 3% and 5%, indicating one node per class is inadequate (Fig. 3.1b & c). Using anything less than one node per class requires the inefficient random placement strategy, which for the 20 class data set greatly reduces identification to 57%. Beyond 200 events (disregarding the networks using one node per class), each of the overall identifications for the three data sets follow a relatively similar path, implying that an increased number of events and nodes does not necessarily mean an increase in performance. This would suggest, that much of the variation that is present in the laboratory cultured species can be covered by using 300 events per class, depending upon the number of classes being analysed. The high individual identification of some species reflects the separability and distinction of their optical characteristics. Increasing event numbers for those species for whom identification is poor, does not always improve performance, indicating that some may never be completely separable. Others, such as *Micromonas pusilla*, appear so distinct that only a small number of events are required for adequate identification.

3.6.2 Imbalanced Event Numbers

As with the balanced data sets (and for the same reason), the overall success decreased as the number of classes increased. The poor results in the balanced data sets for

50 events and less, and the high margin between training and test data, indicated loss of generalisation by the networks. This memorising of the training data was again reiterated in the results for the imbalanced sets, where the difference for 50 events and below was always between 3% and 6%, with the higher end of the scale evident for the 60 species data sets. Similar trends are apparent in the identification success of all species, dependent upon which half of the data set they are from. The identification of species 1 to ½n for each number of events x, dropped as the number of events for species (½n+1) to n increased, for several reasons. As the biological variation of species in the second half of the data set increases with event number, the networks' ability to generalise on unseen data from these classes (½n+1) to n improves. With this increase there is greater chance of misidentification by species 1 to ½n, thereby reducing the identification of the first half of the data set, and possibly overall success. If the mean location of an inadequately represented cluster is false, due to insufficient event numbers or poor data, kernel placement towards that mean (via LVQ) will not necessarily be characteristic of the distribution of the class. Subsequently, as an unseen pattern is presented, it may locate a distance away from the poorly defined distribution centre and provoke an activation from a basis function within its proximity, representing a different, possibly dominant class.

For the 20 class species set, performance for classes 1 to ½n was always lower than for classes (½n+1) to n. This was not a result of the number of events used, but merely the split of particular species into either half of the data set. For example, *Hemiselmis brunnescens* and *Hemiselmis rufescens* (both in A1) have low individual identifications of 62% and 60% respectively, due to misidentification with each other, therefore lowering the overall value. This occurrence can be seen in the extreme species identification, where there appears to be a large difference between best and worst, in each half of the three data sets. For the 40 and 60 species data sets, similar identifications were indicated for both halves of the data set at the points where events are equal, indicating a relatively even distribution of individual species identification. For the reasons already noted, as the balance tips in these two sets, the half represented by a greater number of events has the higher identification.

3.6.3 Compensation for Imbalanced Event Numbers

As the adjustments using class and training probabilities are approximations of training to test data ratio, they will naturally correct the performance rate of the under-represented data. In this instance, event frequency for individual classes, for both training and testing data are known, allowing easy evaluation and application of the method. However, in a field environment or laboratory cultured mixture, this is rarely the case and assuming equal correct class probabilities will generally be incorrect. Additionally, once a network is trained, its employment lies in its ability to identify species, and proportion of species, in an unknown sample. If correct class frequencies are required for output adjustment to identify the sample, then obviously sample content is already known and the procedure seems pointless. With no knowledge available, it is not sensible to adjust output values on the assumption that poorly identified classes are under-represented. The intense overlap of some species implies that complete separability, and subsequent high identification, may never be achieved and adjusting outputs to compensate may produce ambiguous results.

Although the method is suitable in areas such as medical research, where *a priori* data may be obtained from statistical records, in the area investigated here it is unclear how to employ it directly. However, it may be beneficial in a situation where the cost of misidentification is high, and a few false positive identifications are acceptable. For example, when detecting toxic species the number of misidentities could be reduced by scaling network outputs accordingly.

3.7 Summary

The studies indicate, that with a multi-class neural network, adequate training patterns are required to cover the biological variation of each species, especially in cases of high class numbers where overlap is intense. Although a balanced data set is preferable, some species identify well, despite being represented by fewer event numbers. For example, a species that lies outside the main cluster of N species is distinguishable by its hyper-dimensional distance from the group. If the number of variants of the single strain are few, but distinct, a LVQ placement strategy will locate a kernel centre within the cluster, allowing its distribution to be modelled. Despite representation by a smaller number of events than the main cluster (providing they are not too few so as to provoke

memorisation), identification of the species should be high because of its separable characteristics. In this case, the number of events need only cover its biological variation. This appears true of *Micromonas pusilla* (Flagellate), which identifies to 95% with only 100 events, against 400 events for each of the remaining 59 classes. In a balanced data set this species constantly identifies to at least 98% correct. The consistency of the identification when less events are used, can be attributed to its variant characteristics, making it distinct from other species, including its own taxonomic group. *Micromonas pusilla* is not only small (1-3μm), but unlike other species, a plot of orange fluorescence against time of flight produces an empty data set, indicating minimum measures of phycoerythrin.

This is also evident when training a network on 400 events for the 1- ½n classes and 1000 events for the (½n + 1) to n classes (results not shown). When tested on 1000 unseen events for each class, the identification is approximately 8% higher overall for those species represented by classes (½n+1) to n, than classes 1 to ½n. On closer investigation the species that identify to >85% when the events are balanced, are still within approximately 2%-5% of this value when an imbalance is introduced, irrelevant of which side of the data set they are from. However, the species for whom identification is <75%, suffer more when using lower event numbers, benefiting as the number increases. This indicates the importance of representation of overlapping species whose flow cytometric signatures are less distinct.

For most of the studies in this research 300, 400 or 500 events per species have been chosen as adequate to cover biological variation, without the risk of memorisation. The difference in performance, if any, when an increased number of events are used is negligible and would require greater computational time and intensity.

3.8 Analysis of 62 Phytoplankton Species by RBF

Further studies were carried out to produce a more detailed investigation of the RBFs performance for a high number of overlapping classes.

Preparation of the flow cytometric data, and network training and testing was as described in Chapter 2 (Sections 2.6 & 2.7.3). As with the analysis of training data size (Section 3.4), the node numbers defined in the following sections are prior to network

training and the optimal subset were determined automatically, via the orthogonal least squares algorithm (Chen *et al.*, 1991; Chapter 2 - Section 2.4.2.1).

3.8.1 Experimental Procedure

3.8.1.1 Identification to Taxonomic Group Level

The first study examined identification to taxonomic group level. Each of the 5 taxonomic categories were represented by 2500 events, taken equally from the species belonging to each respective group. Ten networks were trained, five using a Euclidean distance metric and five using a Mahalanobis distance metric. Hidden nodes were increased from 2 to 10 per class, in steps of 2. An independent test set was constructed to assess performance.

3.8.1.2 Comparison of Distance Metrics with Large Data Sets

The second study compared the performance of the RBF networks to species level, using the full compliment of data and two different distance metrics. Training and test files were created containing 62 classes (species) of 500 events per class. Ten networks were trained all starting with 6 nodes per class, five using the Euclidean distance metric and five using the Mahalanobis distance metric. The use of five identical networks for each distance metric was to assess consistency of performance and optimisation from random initialisation points.

3.8.1.3 Number of Hidden Layer Nodes

The third study examined the effect of varying the number of hidden layer nodes, with each distance metric, for identification to species level. Eight networks were trained on the 62 class data set containing 500 events per species. Four of the networks used a Euclidean distance metric and four used a Mahalanobis distance metric. Hidden layer nodes were increased from 2 to 8 per class, in steps of two, for both sets of four networks.

3.8.1.4 Principal Component Analysis

The fourth study applied Principal Component Analysis to the seven-dimensional 62 class data, prior to training. Each class contained 500 events. Three principal

components were defined, and used to train a 6 node per class network, employing a Mahalanobis distance metric. A similar test file was constructed to asses performance.

3.8.1.5 Species Combinations

The fifth study examined the effect on network identification when three grouping schemes were introduced. The first analysis investigated performance when all species within a genus were combined into separate groups (genera). This produced 37 classes containing 500 events each, with some groups containing only one species. The second part of this study combined those species, within a genus, whose mutual misidentification, from the optimum network in Section 3.8.1.2, was greater than 5%. This formed 50 classes with 500 events per class. The third area involved the production of a confusion dendrogram from the same optimum network (Fig. 3.5). This is constructed through analysis of the misidentification matrix and shows a progressive natural grouping by the network. This starts with species whose mutual misidentification was high at the left of the diagram, down to those for which confusion was lower. As merging progresses, the networks performance increases until all species are grouped as one, and the networks overall identification is 100%. The degree of confusion between two groups of taxa is determined by summing all probabilities that a pattern a_i (i=1 to n species), belonging to a species from group A, is misidentified as a pattern b_j (j=1 to m species), belonging to a species from group B, as follows:

$$\sum_i \sum_j p(a_i)p(b_j \mid a_i) + \sum_j \sum_i p(b_j)p(a_i \mid b_j) = \frac{1}{m+n} \sum_i \sum_j M_{ij} + M_{ji}$$

Three points were chosen on the dendrogram as indicated in Figure 3.5 and the corresponding species combined, producing 54, 50 and 40 classes of 500 events per class. All networks in study five were initialised with 6 hidden layer nodes per class, employing a Mahalanobis distance metric.

With the exception of the second study (Section 3.8.1.2), all networks were trained 3 times from different initialisation points.

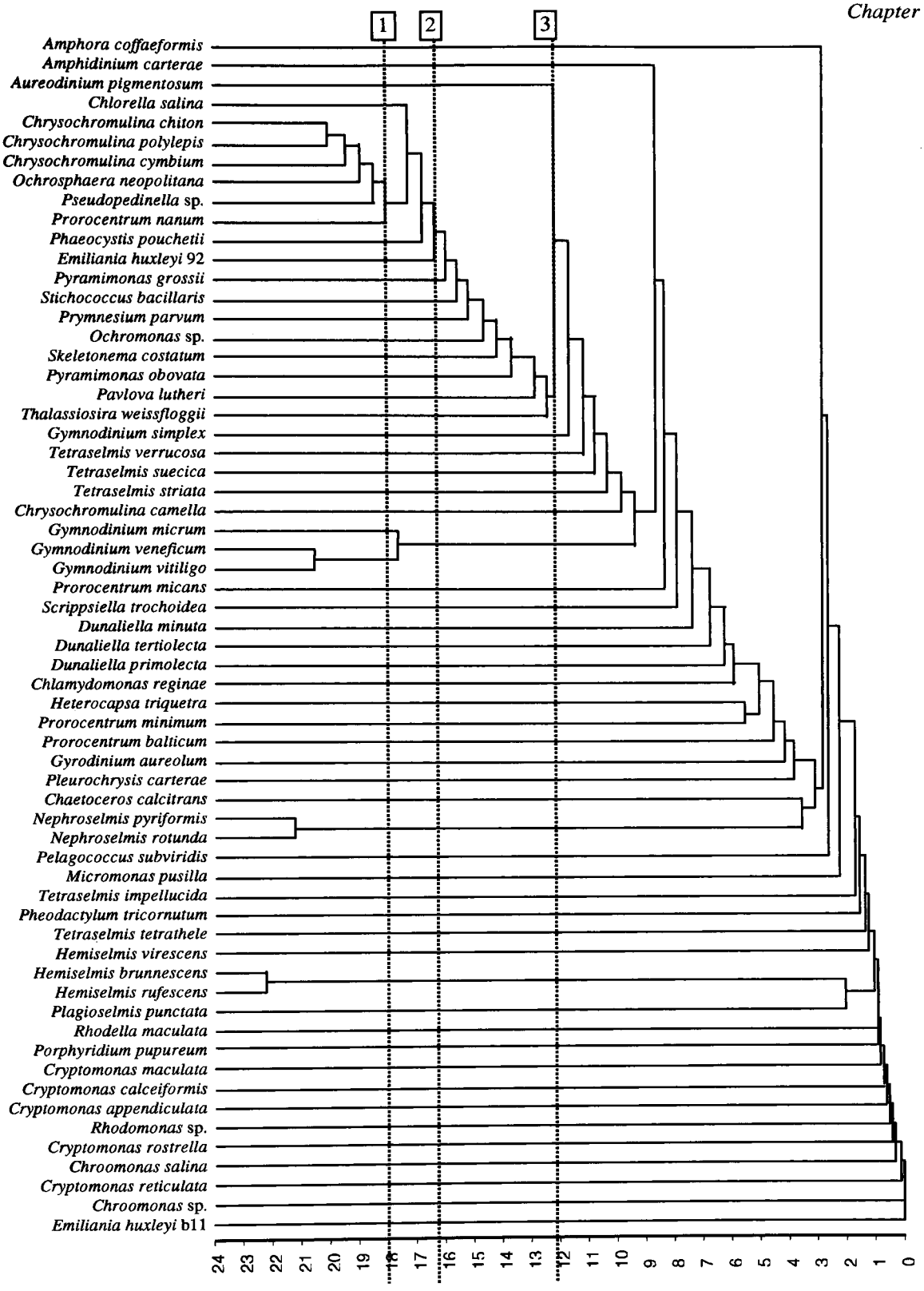


Figure 3.5 Dendrogram showing the order in which respective species were clustered (Section 3.8.1.5) using the optimum network from Section 3.8.1.2. Clustering proceeds from left to right with the ordinate axis showing the percentage of misidentified data remaining at each clustering stage. The positions marked 1, 2 and 3 on the dendrogram indicate the 54, 50 and 40 group sets respectively, for species combination.

3.8.2 Results

3.8.2.1 Identification to Taxonomic Group Level

As the number of hidden layer nodes was increased from 2 to 8, overall performance for both distance metrics increased, with networks employing the Mahalanobis distance metric consistently 3-4% higher than those employing the Euclidean distance metric. Increasing nodes beyond this had a detrimental effect on performance. The optimum network had an overall success of 87.7% and used 8 nodes per class, employing a Mahalanobis distance metric. Group misidentity was most common in the Flagellates, where approximately 15% were misidentified as Prymnesiomonads and more than 7% each as Diatoms and Dinoflagellates (Table 3.3). Confidence of identification for all taxonomic groups was greater than 80%.

3.8.2.2 Comparison of Distance Metrics with Large Data Sets

The networks using a Mahalanobis distance metric had an overall performance of approximately 4% higher, than those trained employing the Euclidean distance metric (Table 3.4). The 5 replicate networks employing a Mahalanobis distance metric had an overall mean of 77.7% correct. The number of asymmetric hidden layer nodes remaining, ranged from 135 to 146, compared to 133 to 154 radially symmetric nodes. The difference in performance for individual species varied. With the exception of three species, the networks employing a Mahalanobis distance metric, had individual identifications of up to 20% higher than those using a Euclidean distance metric. Individual identification in the optimum network, ranged from 41% to 99%, with misidentity existing both within and between groups. With the exception of *Hemiselmis brunnescens* and *Hemiselmis rufescens*, of which misidentity was mutal, the Cryptomonads had identifications and confidence of identification greater than 86%. Prymnesiomonad identification ranged from 46% to 97%, with mutal misidentity between the genus *Chrysochromulina*, and misidentity of *Ochrosphaera neopolitana* and *Phaeocystis pouchetii* with species from different groups. Diatom identification is quite high for all species at an average of 87.6%. Flagellates and Dinoflagellates exhibit mutual, and non-mutual misidentity with species from various groups.

Table 3.3 Misidentification matrix for network trained to identify to taxonomic group level with an overall performance of 87.7%

Taxonomic Group	Cryptomonads	Diatoms	Dinoflagellates	Flagellates	Prymnesiomonads
Cryptomonads	**98.1**	0.1	0.9	0.9	0
Diatoms	0	**94.4**	0.9	3.3	1.4
Dinoflagellates	0.8	0.7	**87.2**	6.5	4.7
Flagellates	0.7	7.4	7.2	**70**	14.8
Prymnesiomonads	0.2	2.2	4.6	4.3	**88.7**
Confidence	**98.3**	**90.1**	**86.5**	**82.4**	**80.9**

Table 3.4 Mean identification and standard error of mean, for individual species across the 5 networks trained for the Mahalanobis (Maha) and Euclidean (Euc) distance metrics.

Species name	Taxonomic Group	Order	Size µm	Mean Identification		Standard Error	
				Maha	Euc	Maha	Euc
Chroomonas salina	Cryptomonads	Cryptomonadida	5-12	92.4	86.0	0.15	0.98
Chroomonas sp.		"	8-10	94.9	91.4	0.21	0.77
Cryptomonas appendiculata		"	15-25	98.0	96.6	0.14	0.67
Cryptomonas calceiformis		"	10-15	92.8	92.2	0.40	0.32
Cryptomonas maculata		"	12-20	90.8	89.0	0.44	1.74
Cryptomonas reticulata		"	18-25	94.6	94.4	0.31	0.35
Cryptomonas rostrella		"	16-25	99.4	99.4	0.04	0.04
Hemiselmis brunnescens		"	5-8	64.6	47.4	1.10	2.23
Hemiselmis rufescens		"	4-9	58.6	65.9	1.98	1.81
Hemiselmis virescens		"	5-8	96.1	94.7	0.17	0.24
Plagioselmis punctata		"	6-9	91.6	84.5	0.70	1.33
Rhodomonas sp.		"	8-13	92.4	88.1	0.78	0.76
Ochromonas sp.	Flagellates	Chrysomonadida	3-12	60.3	52.2	1.67	2.07
Pelagococcus subviridis		"	2-3	87.9	84.6	0.71	0.47
Pseudopedinella sp.		"	8-10	74.1	71.1	1.06	1.04
Micromonas pusilla		Prasinomonadida	1-3	98.9	96.5	0.15	0.64
Nephroselmis pyriformis		"	4-7	70.5	67.8	1.38	1.60
Nephroselmis rotunda		"	6-8	50.6	43.0	1.75	2.56
Pyramimonas grossii		"	5-10	68.4	65.9	1.01	1.00
Pyramimonas obovata		"	4-8	65.0	63.8	0.39	0.42
Tetraselmis impellucida		"	11-19	93.4	88.5	0.74	2.05
Tetraselmis striata		"	6-8	71.8	72.6	1.55	2.22
Tetraselmis suecica		"	6-15	86.6	81.0	0.62	0.71
Tetraselmis tetrathele		"	10-16	94.9	94.8	0.31	0.10
Tetraselmis verrucosa		"	3-11	64.8	41.8	2.35	5.81
Porphyridium pupureum		Rhodomonadida	4-6	95.2	95.2	0.04	0.16
Rhodella maculata		"	7-24	93.5	89.9	0.51	0.75
Chlamydomonas reginae		Volvocida	11-20	91.4	91.3	0.40	0.14
Chlorella salina		"	4-8	53.9	42.8	1.99	2.43
Dunaliella minuta		"	3-12	67.3	59.2	1.19	1.23
Dunaliella primolecta		"	5-12	85.0	82.6	0.45	0.90
Dunaliella tertiolecta		"	6-12	84.0	80.2	0.87	1.18
Stichococcus bacillaris		"	5-8	66.0	48.5	1.13	4.88
Chrysochromulina camella	Prymnesiomonads	Prymnesiida	6-12	86.2	83.6	0.39	0.32
Chrysochromulina chiton		"	5-9	62.4	58.6	0.72	0.60
Chrysochromulina cymbium		"	6-10	43.2	33.0	1.28	2.11
Chrysochromulina polylepis		"	6-8	60.6	59.3	1.33	1.60
Emiliania huxleyi 92		"	5-6	81.1	79.8	0.78	0.92
Emiliania huxleyi B11		"	5-7	97.3	96.4	0.15	0.17
Ochrosphaera neopolitana		"	8-10	41.6	43.5	1.45	0.84
Pavlova lutheri		"	4-6	78.4	73.3	0.53	0.77
Phaeocystis pouchetii		"	3-6	61.1	56.7	0.71	1.27
Pleurochrysis carterae		"	10-18	89.5	83.1	0.97	0.90
Prymnesium parvum		"	8-10	80.1	76.1	0.61	1.34
Amphora coffaeformis	Diatoms	Bacillariophyceae	10-20	88.2	85.1	0.17	0.34
Chaetoceros calcitrans		"	4-6	87.1	85.4	0.39	0.55
Phaeodactylum tricornutum		"	8-35	92.6	92.6	0.24	0.67
Skeletonema costatum		"	3-5	74.5	67.3	0.82	0.80
Thalassiosira weissflogii		"	12-20	91.6	90.4	0.58	0.51
Amphidinium carterae	Dinoflagellates	Dinoflagellida	15-20	75.0	67.6	0.90	1.42
Aureodinium pigmentosum		"	7-12	87.0	84.8	0.56	0.29
Gymnodinium micrum		"	8-15	72.2	65.1	1.07	4.20
Gymnodinium simplex		"	6-10	64.3	62.7	1.80	2.02
Gymnodinium veneficum		"	9-16	43.8	23.2	2.25	3.82
Gymnodinium vitiligo		"	7-22	67.5	70.8	1.34	0.73
Gyrodinium aureolum		"	35-45	85.6	86.7	0.67	1.35
Heterocapsa triquetra		"	15-27	75.6	73.6	1.10	0.46
Prorocentrum balticum		"	9-15	71.4	61.5	1.15	2.08
Prorocentrum micans		"	30-40	80.2	78.5	0.33	0.78
Prorocentrum minimum		"	16-18	59.8	56.6	0.99	1.27
Prorocentrum nanum		"	8-10	56.1	52.6	1.92	1.88
Scrippsiella trochoidea		"	30-42	51.2	44.8	2.13	3.04

3.8.2.3 Number of Hidden Layer Nodes

Increasing the number of hidden layer nodes per class, for networks employing either distance metric, had no marked effect on the identification of the test data (Table 3.5). Performance on the training data was approximately 1-1.5% better.

3.8.2.4 Principal Component Analysis

Using the first 3 principal components produced a very poor result of 54.8% successful identification (data not shown).

3.8.2.5 Species Combinations

All three studies combining species improved overall identification. Grouping all species within a genus and grouping species within a genus only if misidentity is high (>5%), produced comparable results of 84% and 83.4% respectively (Table 3.6). The dendrogram grouping produced identifications of 83.5% (1), 85.6% (2) and 88.7% (3) for points 1, 2 and 3 indicated on Figure 3.5.

3.8.3 Discussion

3.8.3.1 Identification to Taxonomic Group level

Network performance to taxonomic group level is good when considering the overall identification. The main component of this relatively high result are the Cryptomonads which, with the exception of the genus *Hemiselmis,* are a fairly distinctive group. This is primarily due to the presence of Phycoerythrin, a cellular pigment indicated by orange fluorescence, which is more pronounced in the Cryptomonads. The Flagellates are the least well identified, having high misidentity with the Prymnesiomonads and to a lesser extent, the Diatoms and Dinoflagellates. Mutual misidentification with the Prymnesiomonads was not equal and overall identification of this group was 19% higher. This implies kernels located in the overlapping areas of these two groups are representative of the Prymnesiomonad clusters, therefore misidentifying any Flagellate patterns that fall within this region of the sample space, as Prymnesiomonads.

The overall identification to taxonomic group level is probably as high as is possible here. Any identification system will be unable to accurately model an underlying pattern between same class events, when their signatures suggest a pattern does not exist. This situation arises in some species, where the groupings produced from morphometric

Table 3.5 Percentage of test data correctly identified as node numbers for networks employing both Mahalanobis and Euclidean distance metrics are increased. (3 nodes show results from the optimum balanced network from Section 3.5.1)

	Performance of Distance Metric	
No. Hidden Layer Nodes	**Euclidean**	**Mahalanobis**
2	72.7%	77.4%
3	73.2%	77.7%
4	73.9%	77.2%
6	73.7%	77.3%
8	72.9%	77.6%

Table 3.6 Overall identification and confidence of identification produced from combining species at various levels.

Groups	Number of Groups	Overall identification	Confidence of identification
All Species Separate	62	77.7	77.5
Genus' grouped	37	84	83.8
5% Misidentity within a Taxonomic group	50	83.4	83.6
Dendrogram groupings 1	54	83.5	83.4
Dendrogram groupings 2	50	85.6	85.5
Dendrogram groupings 3	40	88.7	88.6

similarities are not necessarily reflected in their flow cytometric signatures. This certainly appears true of the Flagellates. These traditional taxonomic groupings may not be the ideal primary division for identification of flow cytometric data, but suggests a classification centred around the network's interpretation of what is similar may be more appropriate. This is addressed further in Chapter 5.

3.8.3.2 Comparison of Distance Metrics with Large Data Sets

The standard error of mean of the five replicates, indicates slight fluctuations in the identification of some species. This can be attributed to boundary placement being marginally different each time a new network is trained, where the slight change in location may find some data points on the opposite side. The standard error of mean is higher for the Euclidean distance metric approximately 71% of the time, indicating more sensitivity in boundary movement. The spatial extent and orientation of the Mahalanobis measure makes it a much more suitable distance metric for these data, allowing information regarding class distribution to be incorporated. As these data generally have a non-isotropic variance-covariance structure, the Mahalanobis distance metric is more appropriate than the Euclidean, verified by higher identification success. This is apparent when comparing the spatial orientation of a species identified to a much higher extent by the Mahalanobis distance, such as *Hemiselmis brunnescens*, to one equally identified by both, such as *Cryptomonas rostrella* (Fig.3.6). Using radially symmetric kernels, empirically, requires an exponential increase in the number of hidden units (Haykin, 1994). This can have a considerable drain on computational memory and time. Since the kernel size depends upon the spatial distribution of the data, the asymmetric Gaussians, possessing greater orientation qualities than the symmetric, alleviate this problem. Far fewer adjustable hyperellipsoids than hyperspheres are required to adequately model the data. Although the symmetric kernels can be widened to encompass a possible cluster, this will increase the chance of boundary overlap and misidentification, thus having a negative affect on network performance.

The identification of such a large number of species is an advancement over previous works where class numbers were small with little overlap (Chapter 1). Approximately 65% of the species in the optimum network (Table 3.7) were identified to at least 70%, with confidences of the same or higher values for 70% of classes. Overall

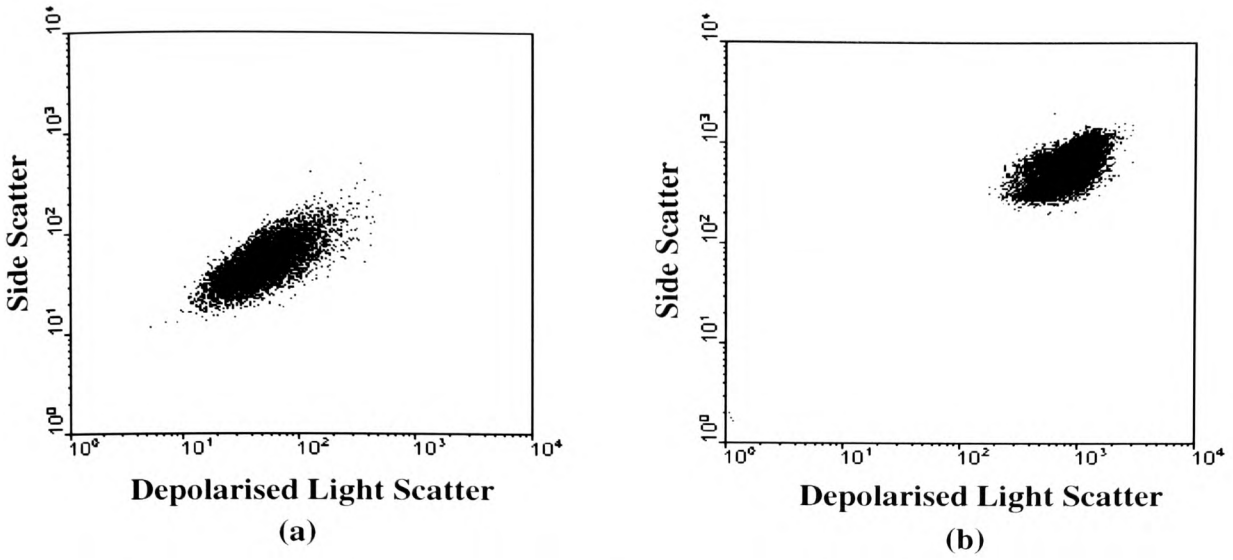


Figure 3.6 Spatial orientation in two dimensions of (a) *Hemiselmis brunnescens* and (b) *Cryptomonas rostrella. Hemiselmis brunnescens* has a non-isotropic variance-covariance structure, and is identified to a much higher extent by the Mahalanobis distance than the Euclidean, due to its elongated ellipsoidal distribution. The more isotropic *Cryptomonas rostrella* is equally identified by both distance metrics.

performance is lowered by a number of poorly identified species, such as *Chrysochromulina cymbium*, which has 20% misidentification as other species within its genus. The same occurs with *Gymodinium veneficum* misidentifying 34% as other *Gymodinium* strains. However, not all misidentifications are within a genus, others are across groups such as *Ochrosphaera neopolitana*, which misidentifies with a number of species in each of the five groups. This does not necessarily indicate poor generalisation or performance of the network. Some species may in fact never be distinguishable by any pattern recognition system, in which case it may only be possible to identify a species as one of two, rather than a definite individual.

3.8.3.3 Number of Hidden Layer Nodes

Increasing the number of hidden layer nodes had little effect upon the performance of the networks. The most marked increase was a 1% rise when going from 2 to 4 nodes per class for the Euclidean distance metric indicating again, to a small extent, the requirement of more nodes to cover the spatial orientation of the clusters. As node numbers increase, the drop in performance and the widening of the margin between training and test data identification, indicates the onset of overfitting by the networks.

3.8.3.4 Principal Component Analysis

In reducing the dimensionality of the data to 3 principal components, the poor results imply a loss in discriminatory information. This indicates the importance of the full multivariate representation, for an improved indication of class membership.

3.8.3.5 Species Combinations

When comparing the results from study one (Table 3.3), identification to taxonomic group level, to the average values for the 5 taxonomic groups from study two (Table 3.7), some discrepancies are apparent. The identification to taxonomic group level for the Cryptomonads was higher than the average value calculated from the groups individual species identifications. This difference can be attributed to overlap and misidentity within the taxa, when identified to species level. For example, the average calculation was lowered by the mutual misidentity of the species *Hemiselmis brunnescens* and *Hemiselmis rufescens*. The same is apparent for the genera *Chrysochromulina* and *Gymnodinium*, both

Table 3.7 Correct identification and confidence of identification of individual species from the optimum network trained on 62 phytoplankton species. Overall identification 77.7% (Section 3.8.1.2).

Species Name and Group	Corr	Conf	Species Name and Group	Corr	Conf
Cryptomonads			**Prymnesiomonads**		
Chroomonas sp.	95.2	97.9	*Chrysochromulina camella*	85.4	76.7
Chroomonas salina	92.4	96.9	*Chrysochromulina chiton*	61.2	60.6
Cryptomonas appendiculata	97.6	95.9	*Chrysochromulina cymbium*	46.2	56.1
Cryptomonas calceiformis	93.6	95.1	*Chrysochromulina polylepis*	60	56.8
Cryptomonas maculata	90.8	91.3	*Emiliania huxleyi 92*	80.8	73.5
Cryptomonas reticulata	95	97.5	*Emiliania huxleyi B11*	97.6	95.1
Cryptomonas rostrella	99.4	94.6	*Ochrosphaera neopolitana*	45.6	55.5
Hemiselmis brunnescens	65	67.4	*Pavlova lutheri*	77.6	70.6
Hemiselmis rufescens	64.4	68.8	*Phaeocystis pouchetii*	59.4	62.4
Hemiselmis virescens	95.8	95.8	*Pleurochrysis carterae*	92.2	83.8
Plagioselmis punctata	92	86.5	*Prymnesium parvum*	79.8	70.9
Rhodomonas sp.	93.4	94.2			
Average	**89.55**	**90.16**	**Average**	**71.44**	**69.2**
Flagellates			**Diatoms**		
Chlamydomonas reginae	91.8	76.9	*Amphora coffaeformis*	88	90.9
Chlorella salina	52.2	58.8	*Chaetoceros calcitrans*	87.6	83
Dunaliella minuta	67.4	75.1	*Phaeodactylum tricornutum*	93.4	90.8
Dunaliella primolecta	85.2	82.7	*Skeletonema costatum*	76.2	76
Dunaliella tertiolecta	82.4	75.7	*Thalassiosira weissflogii*	92.8	74.5
Micromonas pusilla	99.4	81.5			
Nephroselmis pyriformis	71	62.6	**Average**	**87.6**	**83**
Nephroselmis rotunda	54	61.8			
Ochromonas sp.	57.4	67.5	**Dinoflagellates**		
Pelagococcus subviridis	87	90.6	*Amphidinium carterae*	77.8	72.6
Porphyridium pupureum	95.2	97.7	*Aureodinium pigmentosum*	88.2	73.4
Pseudopedinella sp.	76	69.5	*Gymodinium micrum*	71.2	62.6
Pyramimonas grossii	67.4	73.6	*Gymodinium simplex*	69	66.9
Pyramimonas obovata	64	65.4	*Gymodinium veneficum*	41.4	68.8
Rhodella maculata	93	94.5	*Gymnodinium vitiligo*	66.4	63.1
Stichococcus bacillaris	67.6	77.3	*Gyrodinium aureolum*	86	92.1
Tetraselmis impellucida	94.8	93.5	*Heterocapsa triquetra*	72.4	79.7
Tetraselmis striata	76	71.4	*Prorocentrum balticum*	70	73.5
Tetraselmis suecica	87	84	*Prorocentrum micans*	81.2	59.9
Tetraselmis tetrathele	94.6	89.2	*Prorocentrum minimum*	61.6	77.6
Tetraselmis verrucosa	60.2	71.7	*Prorocentrum nanum*	56.4	70.9
			Scrippsiella trochoidea	51.2	66.7
Average	**77.31**	**77.19**	**Average**	**68.68**	**71.3**

lowering the average calculation of their respective taxonomic group. These occurrences would imply, that in this instance, it may be easier to group the overlapping species and identify a pattern as either X or Y, rather than choosing one, thereby improving overall performance. The increase in identification when grouping species from the same genus, or species within a genus whose mutual misidentification was greater than 5%, indicates that for some of these species placement into genera by morphology is supported by similarities in their flow cytometric signatures.

Although imposing groupings based on genus may be an obvious procedure, an analysis of the results of study two indicate that not all misidentifications are within a taxonomic group or genus. For example, the 5 strains of the genus *Cryptomonas* all identify with at least 90% success, with confidences of 91% and above. Similarly, two out of the five species in the genus *Tetraselmis*, particularly *Tetraselmis impellucida* and *Tetraselmis tetrathele*, identify to 94% with high confidences. In contrast, there is at least 6% mutual misidentification between *Pseudopedinella* sp. (Flagellate) and *Prorocentrum nanum* (Dinoflagellate). *Phaeocystis pouchetti* (Prymnesiomonad) constantly has at least a 10% misidentification as *Chlorella salina* (Flagellate), with the reverse misidentity only 2.5%, implying not all mutual misidentifications are necessarily equally weighted. These misidentifications are further illustrated in the dendrogram. The three groupings chosen from the dendrogram, gave slightly better results than combining species based on genus or taxa. The networks selection of groupings is more successful than forcing morphometric groupings not supported by flow cytometry (Chapter 1, Fig. 1.3a & 1.3b). This point is illustrated in the groups formed at position 2 on the dendrogram and the combination of species within a genus whose mutual misidentity >5%. Both selections produce 50 groups but the former has a 2.2% greater overall success. Although producing an identification as one of N possible species can be at the expense of detail, knowledge to genus or *user-defined* group level may be a requirement, or sufficient to narrow down the choices. An additional method of identification, such as increased discriminatory parameters or microscopy, can then be employed with a much smaller number of known possibilities.

3.8.4 Rejection of Species as Unknown

The importance of network confidence, at the expense of overall identification through rejection of unknowns, was discussed in Chapter 2 (Section 2.7.4.2). However, as class numbers increase, the greater the possibility of overlap between the flow cytometric signatures of a known and novel species, thus making the process more difficult. This is assessed below.

3.8.4.1 Experimental Procedure

Eleven completely novel species (Table 3.8), *i.e.* species upon which the network had not been trained, were added to the original test file from study one (500 events per class). The optimum network from this study was selected and three different criteria for rejecting novel data were imposed.

1. Rejection if the difference between the two highest outputs is less than a threshold *T1*, where *T1* was varied from 0 to 0.9 in intervals of 0.1
2. Rejection if the highest valued output is less than a threshold *T2*, where *T2* was varied from 0 to 0.9 in intervals of 0.1
3. Rejection if the maximum hidden layer node output is less than a threshold *T3*, where *T3* was varied from 0 to 0.9 in intervals of 0.1

3.8.4.2 Results

With a threshold imposed upon the difference between the two highest node outputs, overall success drops rapidly as *T1* increases (Fig. 3.7a). At *T1* = 0.3 correct identification is at 53%, with 80.7% of unknowns being rejected and 44% of knowns. Rejection criterion based on maximum output node value (Fig. 3.7b) and maximum hidden layer node output (Fig. 3.7c) perform better, with the latter producing a fairly steady overall identification until *T3* = 0.6, and then dropping rapidly. The former drops much sooner past *T2* = 0.3. Although rejection of unknowns at this value is high, *i.e.* 80%, as the threshold increases to 0.5, the number of knowns rejected is also quite high, at approximately 32%. The threshold imposed upon hidden layer node output, rejects a lower number of unknown species, *i.e.* 60% at *T3* = 0.7, but less known species (12%).

Table 3.8 Eleven new species selected to assess the RBF networks ability to reject novel species when original class number is high.

Taxonomic Group	Species Name	Order	Size µm
Diatom	*Chaetoceros affinis*	Bacillariophyceae	10->100
	Chaetoceros debilis	"	10->100
	Chaetoceros radicans	"	10->100
	Surirella sp.	"	10->100
Dinoflagellate	*Alexandrium tamarense*	Dinoflagellida	28-40
	Alexandrium lusitanicum	"	25-40
Prymnesiomonad	*Imantonia*	Prymnesiida	2-4
	Platychrysis	"	8-10
	Dicrateria inornata	"	3-5
	Chrysotila lamellosa	"	4-7
Flagellate	*Nannochloris atomus*	Volvocida	2-4

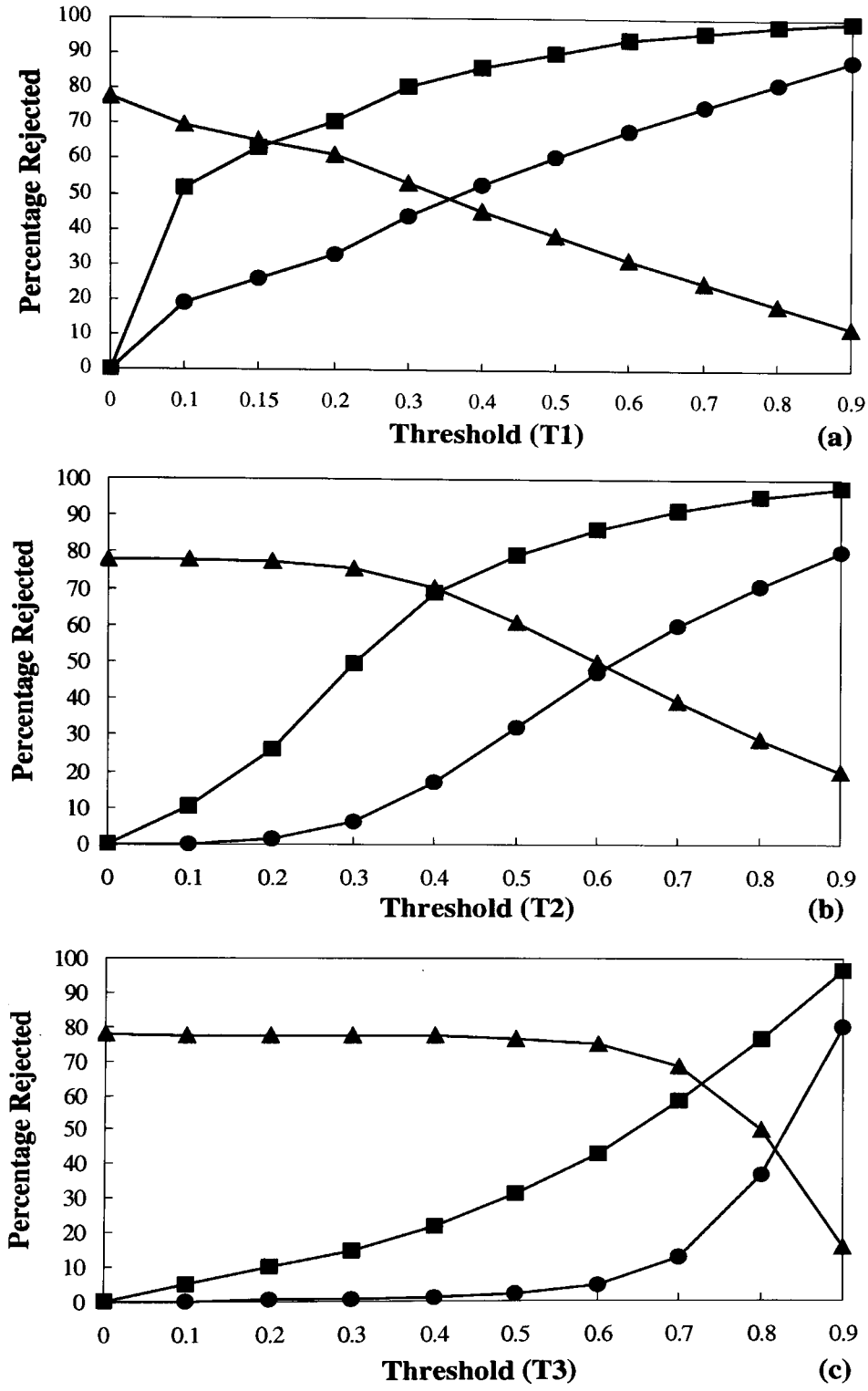


Figure 3.7 Overall percentage of species rejected as unknown employing three thresholds ● Overall rejection of known species ■ Overall rejection of unknown species ▲ Overall percentage of correct identification (a) Difference between highest and second highest node is less than a threshold *T1* (b) Highest valued output node is less than a threshold *T2* (c) Maximum hidden layer node output is less than a threshold *T3*.

3.8.4.3 Discussion

Although the imposition of a threshold on output node difference (*T1*) produces a high rejection of unknown species, its similar affect on known species makes it a poor criterion to use. The complexity and similarity of the spatial distribution of some classes results in relatively high values for a number of output layer nodes. A significant difference between the highest and second highest node is very rare, causing early rejection of species at a very low threshold.

With increased classes, the thresholds imposed on the hidden layer node output and maximum valued output are more appropriate. However, despite having a high rejection of unknowns, the maximum valued output threshold (*T2*) has a relatively high rejection of knowns, though not as high as *T1*. Conversely, the threshold on the maximum hidden layer node (*T3*), rejects a much lower number of knowns but not as many unknowns. The high number of knowns rejected by *T2* indicates low output values for many of these species, resulting in quite a rapid drop in overall identification. Although overall percentages are depicted, individual species' contribution to these results are considerable. For example, Figure 3.8 shows the rejection of the species at the extreme ends of the scale, for both novel and known data for *T3*. *Alexandrium lusitanicum*, an unknown species, is rejected almost immediately by this threshold, indicating its considerable difference to any species in the original database. However, *Imantonia* sp. has a very low rejection level due to its 69% misidentification as *Pelagococcus subviridis*, at a threshold of $T3 = 0.6$. Although this known species has a high individual identification (87%) its confidence of identification at this threshold is very low (45%), attributed to unknown species being retained and misidentified as it. Rejection of the known species *Gyrodinium aureolum* indicates that class distribution about its kernel is quite wide, with data points being found a distance away from the cluster centre. This is unlike the close proximity of *Pseudopedinella* sp., to its kernel centre, for which rejection is very low.

Although novel data will be rejected through imposition of thresholds, for some, this rejection will be weak (*e.g. Imantonia* sp.) due to similarities to known species. A knowledge of possible inhabitants would be preferable in a field area, thereby eliminating possibilities that may not be present in certain locations.

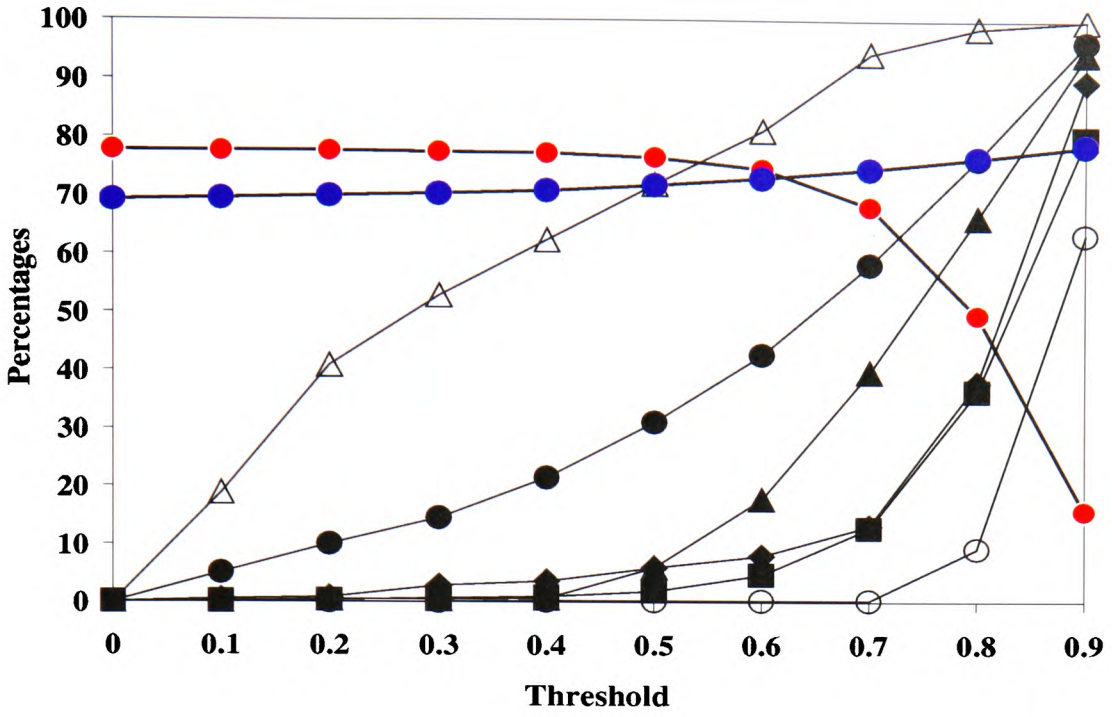


Figure 3.8 Correct identification and rejection percentages when a threshold is imposed upon the maximum hidden layer node output. ● Overall percentage correct identification; ● Overall confidence of correct identification; ● Percentage of unknown species rejected; ■ Percentage of known species rejected; Rejection of individual species, ○ *Pseudopedinella* sp. (Known), ▲ *Gyrodinium aureolum* (Known), △ *Alexandrium lusitanicum* (Unknown), ♦ *Imantonia* sp. (Unknown).

3.9 Conclusion

The optimum architecture of the RBF network, as well as varying event numbers for phytoplankton flow cytometric data, has been illustrated here. Although 500 events per class were used in the analysis of networks trained on 62 species (Section 3.8), the primary results of this chapter imply that biological variation can be adequately covered, in this instance, by less event numbers.

The overall identification of the RBF network trained on 62 classes, is comparative to previous work (Chapter 1, Section 1.7). For example, a correct identification of 75% was achieved by an MLP network trained on 40 species of phytoplankton (Boddy *et al.*, 1994a), and an RBF network, employed by Wilkins *et al.* (1999), identified 34 species to 92% correct. However, in both these cases the data was represented by eleven optical parameters, and the latter used both laboratory cultured and field samples producing more distinct data classes (Boddy *et al.*, 2000). Greater correlation can be seen with Boddy *et al.* (2000), where a RBF network trained on seven-dimensional data successfully identified 72 species to an overall of 70% correct.

Further improvement in network performance can be achieved by combining those species for whom mutual misidentification is high. The increase in overall identification resulting from this, reflects the lack of correlation between the flow cytometric signatures of some species and their respective morphological characteristics. Combining species will also lower rejection of borderline knowns that lie on the boundary overlap of two similar species. In some cases no system will be able to adequately differentiate between overlapping species. When this arises, it seems appropriate that the species concerned are in fact similar enough to be considered as a flow cytometric group. If required, further discrimination can then be achieved via an alternative method. These results direct the initiation of an alternative grouping system for this area of research. This is discussed further in Chapter 5.

Although rejection of unknown species can increase network performance, it may also be desirable to then add the new species to the existing database. Species addition is a difficult process using the original multi-class network, and requires the introduction of an alternative approach, presented in Chapter 4.

4 Alternative Multiple Neural Network Architecture

4.1 Introduction

The advantages and performance of the original multi-class RBF, within a biological field, have been demonstrated for phytoplankton identification in Chapter 3. However this architecture has a number of limitations. The phytoplankton community is diverse and immense, and the introduction of a new species to the database is a natural occurrence. Each time a new species is encountered and added, the network requires complete retraining, involving long optimisation procedures to be carried out under the supervision of scientists familiar with ANNs. This makes the multi-class network an inflexible architecture that cannot operate in real time.

This chapter introduces an alternative multiple network architecture for the identification of phytoplankton species. The approach is flexible, rapid and can be used by non-computer scientists.

Experiments are documented that compare the alternative multiple network approach with the original multi-class architecture, with regards to training times, performance, ease of use, species combinations, exclusion of unknowns and their subsequent addition to the identification system.

4.2 Restraints of the Original Multi-Class Network Architecture

The efficiency of any identification/classification system, neural or statistical, is dependent upon a number of factors, including training time, real time analysis, ease of use and system requirements. However the primary influence on performance is the quality, and in some cases, the quantity of the training data.

For identification of organisms as diverse as phytoplankton, all variations of flow cytometric signature for a particular species must be represented. As the complexity of such a data set increases, so does the architecture of the network required to model it. The increase in number of nodes and weights to achieve optimal performance, can have a heavy effect on memory and will inevitably increase training time. In many applications this is irrelevant. Once an optimum level has been achieved, the network has the capacity to identify phytoplankton cells, however, this is limited to species contained in the original training data. When the network is presented with unseen data, any unknowns with significant differences from known species, can be discarded by the imposition of a

suitable threshold (Chapters 2 & 3). Once the unknowns have been rejected and identified, it may be necessary to include them in the database for future encounters. In this situation the original multi-class network is now rendered useless, and a new network will need to be constructed and retrained using the updated data set. This process will need to be repeated each time a new species is added.

In a laboratory, or on board ship, this is not a feasible process for real time operation, where the network needs to be easily and quickly re-trainable whenever a new species is encountered. The primary users will not necessarily be familiar with ANNs and a large network will require complex optimisation to achieve good performance. If training data is sparse, as is the case with many field applications, there is a risk of overparametisation with increasing network size, thus causing inadequate modelling of the data distribution and poor generalisation.

Additionally, the target population for analysis may be a subset of the actual database, where only certain species are required, and a network particular to the entire data set may therefore be unnecessary, and may be sub-optimal. This flexibility, real time training and dynamic selection of species, is impossible using the original multi-class architecture. A different approach, introducing a combination of multiple networks, novel to the identification of phytoplankton, has therefore been developed.

4.3 Combinatorial Neural Networks

Some of the limitations of the original architecture mentioned above, have been overcome by combining neural network models in various configurations. Many researchers have found that they improve not only performance and training time, but in some cases reduce complexity. These combinatorial structures can be a fusion of different, or the same network algorithms, existing in hierarchical or nested arrangements, connected in series or parallel.

Some methods combine unsupervised and supervised learning, where the former partitions the input space into subsets through feature extraction, and a number of supervised networks are then trained individually on the subsets of data. Yang *et. al.*(1996) used an Adaptive Resonance Theory network (Carpenter & Grossberg, 1987a, 1987b) to classify non-stationary gas from odorous environments, while an MLP identifies the gas or odorous mixtures. Similarly, Raghavan *et al.* (1991) used a collection of

Adaptive Resonance Theory networks for generic feature extraction and a number of back-propagation MLPs to identify the finer more specific features.

Other hierarchical/multiple combinations consist of interconnections of individual networks, each trained for a specific task. Hierarchical configurations tend to be sequential, where the results of a preceding network are passed onto the next. Examples of these are evident in many areas of research. Juell and Marsh (1996) used a hierarchical structure of four back-propagation networks to identify images of faces. The three *'child'* networks were trained individually to identify a mouth, nose or eyes respectively, while the *'parent'* network was trained to recognise a face, providing all three features were present. Mehdi *et. al.*(1994) used a similar approach to separate normal cells from abnormal cells, where the first of three back-propagation networks performed the primary separation of normal or mildly dysplastic cells from moderate or severely dysplastic, and the two remaining networks partitioned the two subgroups of data into one of four classes. Higher numbers of networks can be seen in the production of the Artificial Neural Network Short-Term Load Forecaster or ANNSTLF (Khotanzad *et. al.* 1997). Developed for electric load forecasting, this paradigm utilises multiple back-propagation networks each focusing on a particular aspect of the training data. Other approaches use multi-layered hierarchical networks, like that of Namphol *et. al.* (1996), where a nested training algorithm is employed to partition images into sections and each particular segment is then processed simultaneously.

Whilst some of the sequential and parallel processes converge to a basic one output solution, others require more complex analysis involving combinations of outputs from multiple identifiers. Basic solutions include a voting winner takes all system (Huang *et al.*, 1997), averaging the outputs or combining networks using AND/MIN logic gates to generate output statistics (*e.g.* Shazeer 1992). These procedures assume that the individual networks within the system are trained to an equal optimum level. Other methods employ areas such as fuzzy logic (*e.g.* Wang *et. al.* 1998) or principal component regression (*e.g.* Zhang, 1999) to fuse multiple network outputs.

The methodologies documented above are a selection of combination procedures, all of which perform optimally for their particular application. Although many of them are individually trained they are not all completely independent. Sequential processes will depend upon the outcome of a preceding network's performance. Many of the individual

networks are trained for a specific task or characteristic of the application. These processes, much like decision trees and statistical hierarchical approaches, imply that accurate identification is dependent upon all network *'questions'* being *'answered'* correctly, and the redundancy of one network, due to perhaps unavailable or poor information, makes the procedure inaccurate or impossible. A poor decision (due to poor training or memorisation of data) at a top level network, renders all subsequent decisions incorrect. Partitioning of the input space by unsupervised means requires some measure of known similarity (Chapter 5), as do those methods using a statistical *a priori* approach. Some are complex, lacking in flexibility, are application dependent and will require complete re-training with a new data set if a novel category is introduced.

To address the limitations of the original multi-class architecture and offer an alternative approach in combinatorial networks for phytoplankton identification, a multiple network structure has been suggested to incorporate large data sets, direct weighted node placement, rapid training times and easy addition of a new species by non-computer scientists.

4.4 Alternative Multiple Neural Network Architecture

The alternative multiple approach introduced in this research, is a variation on many of the general combinatorial network structures. The architecture consists of N simple identification Radial Basis Function networks (single species networks), each one responsible for one of the N individual species being considered (Fig. 4.1). Each of the single species networks, consist of a seven node input layer and one hidden node layer. Unlike many traditional combinatorial neural networks structures, identification by each of the individual networks is for a particular species only, however, as a one class RBF is meaningless, the single species networks have a two node output layer. Training data, therefore, consists of two classes, mapping onto the two network outputs. Class A, the class (species) of interest, contains x events for the species in question and Class B, the background class, contains a random selection of the remaining species (in this case 61). Each of the N networks are trained on the seven input parameter data, with the respective species being the class of interest and the output indicating membership as either species or background.

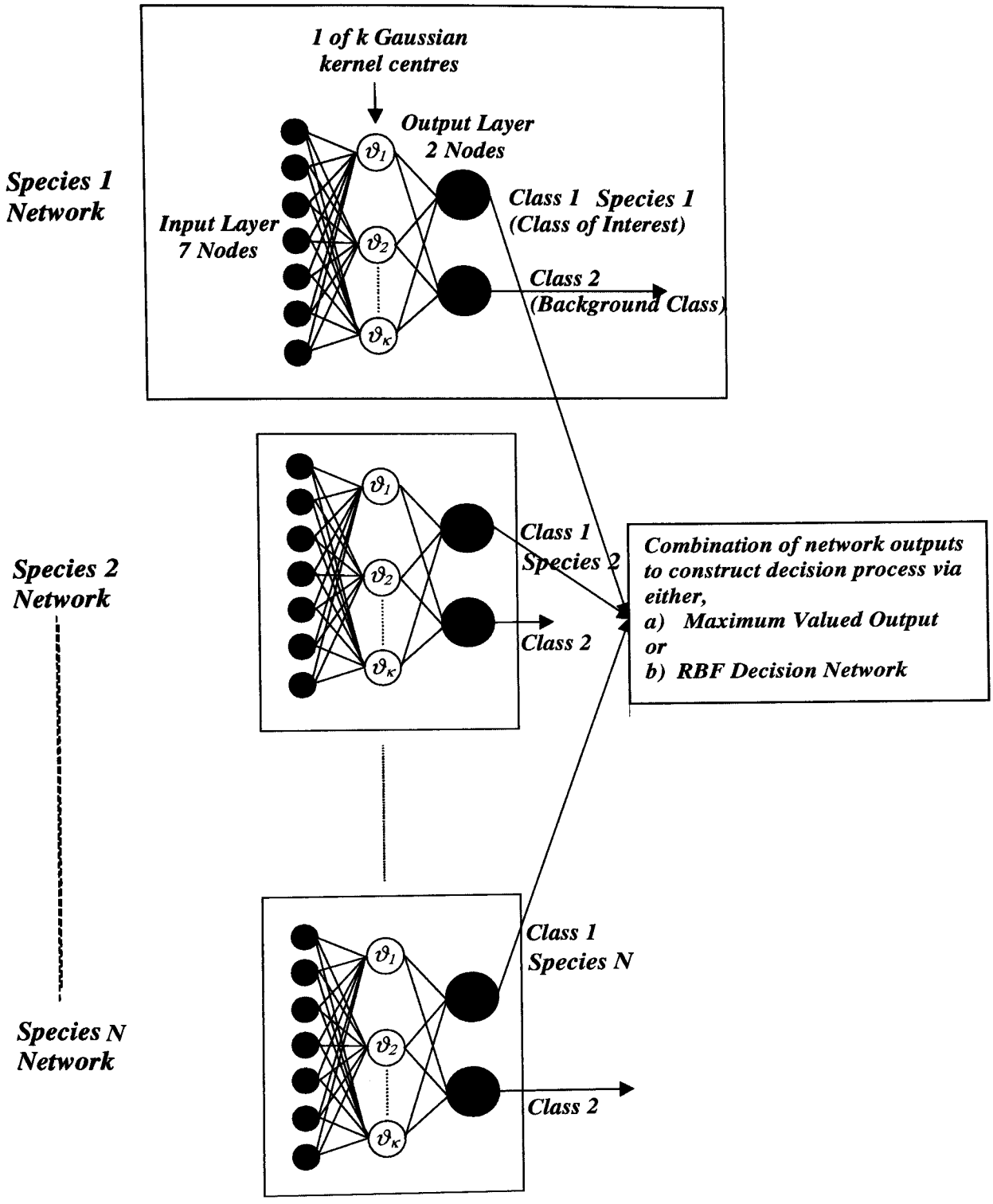


Figure 4.1 Alternative multiple network architecture showing combination of N single species networks for N species. Each single species network consists of a 7 node input layer for the seven-dimensional optical data, an individual hidden layer and a 2 node output layer. Outputs from single species networks are then passed on to the combinatorial stage where an decision on pattern identification is made.

When training against a background class (Section 4.5) the amount of data that constitutes it can be unbounded. Even in laboratory cultured phytoplankton, it is unrealistic to obtain sufficient event numbers to balance the class of interest against the combination background class, and it has been shown that this is in fact unnecessary with regards to biological variation; which can be represented sufficiently using 300 events per species (Chapter 3). Thus the ratio of event numbers for Class A to that of Class B may be unequal. With a heavily imbalanced data set, where the class of interest (Class A) is generally engulfed within the mass of background data, RBF identification will perform best if hidden nodes are placed directly at the position vector of representative data patterns from the class of interest. Training involves the random selection of a subset of data points to act as kernel centres, employing a Euclidean distance metric. Node placement can be weighted towards and directly into either class (*i.e.* m_A kernel centres are placed at the position vectors of m_A nodes from class A, similarly for class B). Employing a Gaussian kernel function, the normalisation parameter is calculated as the root mean square distance between a kernel centre and its corresponding data points (Chapter 2, Section 2.4.2.2). Matrix operations are performed in the output layer to find the optimum weights (Chapter 2, Section 2.4.2.3).

In a two class imbalanced data set, identification will tend towards the dominant class, where the likelihood that a data point, within close proximity of a centre, is from the background class, increases as negative examples abound. Identification of the subordinate class is very specific to data points that have been well represented by the hidden nodes. Naturally, ideal representation will come from placing as many centres as there are points and using each point as a kernel centre. This is of course unacceptable as it will only force the network to memorise the training data, making it unable to generalise and computationally intense. Any pattern unrepresentative of the class of interest will be misidentified as class B. The sheer abundance of data in the background class amplifies this outcome and, whether there are 1 or 20 nodes placed directly into the background class, has been found to have negligible affect on its identification. However, increasing nodes in class B progressively reduces the number of class A patterns identified correctly, and as class A is the class of interest, 1 hidden node is found sufficient to represent the background class. Once the N single species networks have been trained and tested, the

outputs are combined via one of the procedures below and a decision upon class membership of an unseen pattern is made.

4.4.1. Maximum Valued Output

This is a simple method, employing a winner takes all procedure. The output from each of the N networks, for a particular pattern, are considered and compared, and the network having the highest value is deemed the winner. The pattern is then assigned the species for whom the network has been trained.

4.4.2 RBF Network Decision

Using the original multi-class network structure to identify a 62 class data set, gives varied levels of performance for different species. For example, the strains of Cryptomonas have distinct optical characteristics producing identifications of 95% and above, whereas species for whom overlap is great, have values as low as 41% (*Gymodinium veneficum*). When the class of interest of a single species network is a species for whom identification was found to be poor by the original multi-class network, the number of events correctly identified by the single species network may be low or possibly zero (Section 4.7.1). A set of seven-dimensional optical characteristics for a particular species represents the flow cytometric signature associated with that species. Providing there are no outliers and the data set is representative, a trained network will produce output values that lie within a common range for a particular species. Once the network is trained and the optimum architecture determined, a test pattern from the species in question will be translated from its seven-dimensional input to an output value for the class of interest. This pattern, having gone through the same network transformations as its representative training data, will produce an output value within the same common range. After all the single species networks have been trained, each one will produce a distinctive output for a particular pattern applied to it, where ideally, the output produced from the n^{th} single species network for species n will be 1, and the remaining N-1 outputs will be zero (this is of course never the case as continuous values are produced). Combining these N outputs, will form a signature vector representative of the species for whom the original seven-dimensional optical pattern belonged. Applying a data set of j events per species through a consistent architecture of N single species networks, will

produce j, N-dimensional patterns distinctive to each individual species. This process forms a new set of data with characteristic signature vectors, that can be used as inputs to train an individual RBF decision network, with N inputs and N outputs.

4.4.3 Bayesian a posteriori Probabilities

A process of combining outputs from multiple classifiers was used by Singer and Lippmann (1992), to obtain scaled word likelihoods by the combination of RBF network outputs which are normalised by *a priori* class probabilities. This requires the network inputs to be independent.

The multiple network approach introduced here, is not a combination of multiple classifiers each trained to recognise the same classes. Although scaling the outputs of the single species networks and employing Bayes theorem to combine them will allow an evaluation of probability, it requires *a priori* knowledge. Ideally, balanced training data frequencies are employed, and this coupled with an assumption of equal *a priori* probabilities, will in this instance repeat an iterative update process of calculating *a posteriori* and subsequently adjusting *a priori*, that does nothing more than approximate the relationship between the actual outputs, thereby producing results identical to those of the maximum valued output method.

Although the approach is not employed here, its benefit may lie in an area where prior knowledge is available, or when scaling towards identification is required to avoid damaging false negatives.

4.5 Background Class – Content and Quantity

When only one species is of interest everything else is considered background. It is unrealistic to attempt to train a network to discriminate one species against every known strain of phytoplankton and every possible form of debris. Chapter 3 has already indicated the high identification of particular species, where their separability was demonstrated against an environment of 61 others, although their distinction may be such that the content of the remaining classes is irrelevant. This is not the case for those species for whom identification is poor, due to class overlaps. In some areas of research involving multiple networks, each individual network is trained for all classes and the outputs of each network, or a selection of the networks, combined to form a decision (Hashem, 1997). In other cases the individual networks are trained on two or more classes, but less than the

total number (Wilson, *et. al.* 1996). The training data sets for the individual networks in these instances, comprise more than one known class of interest, each containing a balanced number of events, where identification of every class within a data set, for each individual network, is important. In the alternative multiple approach presented here, the content of the background class is varied and shown, for this data, to have little affect on the final identification of the species particular of the individual networks.

The quantity of background data is also a factor. When considering identification for one species against an unbounded plethora of others, if full biological variation of the background class is necessary, there will inevitably be an imbalance in event numbers. From investigation researched in Chapter 3, the number of events for a balanced set of 62 species (to adequately represent biological variation) was between 100 and 300. For an imbalanced data set, the lack of identification of the subordinate class indicates that generalisation ability and primary error reduction are concentrated in the dominant class, where the greater numbers of events allows the hidden layer nodes to better model its distribution. Data duplication is one method used to overcome this problem (Foody, 1995). Although it does not add new information regarding characteristics or underlying trends of the data, it will increase the representation of the subordinate class, possibly improving the networks approximation of its distribution. Data replication and weighted node allocation have been investigated further (Section 4.4).

4.6 Experimental Procedure

Data preparation and pre-processing was as described in Chapter 2 (Section 2.6).

4.6.1 Single Species Training and Testing Files

To assess the approach with varying ratios of Class A event number to Class B event number, six sets of 62 single species training files were constructed. Each file contained x events for the species of interest (Class A), and a combination of y events per species for the remaining 61 (Table 4.1, sets A-F). In order to assess data repetition, two additional sets of 62 files contained duplicate data. The first comprised 1000 events for Class A, *i.e.* 500 repeated twice, and 500 per species in the background class. The second contained 1000 events repeated 30.5 times, and again 500 events per remaining species in class B, producing a balanced set (Table 4.1, sets G & H).

Table 4.1 Event numbers for each of the 8 sets of 62 single species network training files. Overall identification success shown for each set of single species networks with increasing numbers of hidden layer nodes (hln) in the class of interest, employing a Euclidean distance metric (Section 4.6.2).

Data Set	Class of Interest	Background Class		Overall identification of Class of Interest		
		Events per Species	Total Events	3 hln	10 hln	20 hln
A	400	400	24400	38.21	51.24	55.73
B	500	500	30500	36.29	49.44	56.00
C	500	8 or 9	500	93.8	N/A	N/A
D	1000	500	30500	43.21	66.33	67.99
E	500	250	15250	44.20	67.34	68.20
F	500	1000	61000	30.15	37.79	46.01
G	1000 (500 repeated twice)	500	30500	40.12	67.46	69.55
				5 hln		
H	30500 (1000 repeated 30.5 times)	500	30500	91.02	95.24	N/A

An independent test set of 500 events per species was created to assess network performance as documented in Chapter 2 (Section 2.7.4.1).

4.6.2 Single Species Network Training

For each of the data sets, A, B, D, E, F and G, three different architectures of 62 single species networks were trained. These used a random kernel placement strategy, placing either 3, 10 or 20 nodes directly into class A, all employing a Euclidean distance metric. As discussed, 1 hidden node is placed into class B for all networks. Data sets C and H, for whom event numbers were balanced, were trained using equal nodes in each class. A single architecture of 62 single species networks, with 3 hidden layer nodes per class (A and B), was trained for data set C, and two sets of 62 single species networks with 5 and 10 nodes per class were trained for class H. All networks were trained three times from a different initialisation point.

4.6.3 Maximum Valued Output

The results from each set of 62 single species networks were analysed in parallel. The maximum valued output was calculated for each individual pattern across the set, and the class associated with it assigned the winner.

4.6.4 RBF Network Decision

4.6.4.1 Training and Testing Files

From each structure of 62 single species networks (of which there are 21), the outputs were combined to create 4 training files containing 400, 300, 200 and 100, 62 parameter events (84 training files in total). This was to assess the number of events required to cover the variation of the 62 parameter signature vectors, used as input data for training the RBF decision networks.

The outputs of each set of single species networks, for an independent test file, were combined to form 21 files to assess performance of the particular RBF decision networks.

4.6.4.2 Network Training

Each of the four training files created for each of the 21 sets of single species networks, were used to train 6 RBF decision networks using 3, 4 and 5 hidden layer nodes respectively, 3 using a Euclidean distance metric and 3 using a Mahalanobis distance metric to compare (504 RBF decision networks in total). Initial experiments (results not shown) comparing 1, 3, 5 and 10 iterations of optimisation (Chapter 2), exhibited negligible improvement in network performance, but a considerable increase in training time, particularly those employing a Mahalanobis distance metric taking 35 hours to train for 10 iterations. Therefore, 1 iteration was used to train all networks, each trained three times from a different initialisation point.

4.7 Results

4.7.1 Single Species Networks

As the node numbers increased from 3 to 20, so did the overall identification of the class of interest (Table 4.1). This improvement in performance is also evident when the ratio of events between the two classes is reduced. The data sets containing balanced and repeated balanced events, produced identification values almost as high as that of Class B (97%-100%). However, as the number of nodes and degree of repetition increases, so does the margin between training and test data identification success. When considering individual results for a set of balanced single species networks many of the identifications of the class of interest were >80%, however, when the imbalance is greater the identification is more varied (Table 4.2). Using 3 hidden layer nodes in the class of interest left a number of species with few or no test patterns assigned to them. As the number increases to 10, the identification in some of these badly allocated classes also increases, but rarely higher than 40%. Training using 20 hidden layer nodes has little effect on many of these weaker species, improving some only marginally and in certain cases reducing identification.

4.7.2 Maximum Valued Output

Using only 3 hidden layer nodes for the class of interest in the single species networks produced the poorest results, with values below 70% (Table 4.3). Improvements were evident from networks employing 10 and 20 hidden layer nodes, ranging between

Table 4.2 Table indicating the results of three sets of 62 single species networks trained using 500 events in the class of interest (Class A) and 500 per species in the background class (Class B). The networks were trained using a Euclidean distance metric and 3, 10 and 20 hidden layer nodes respectively for the class of interest. The mean and standard deviation for the classes of interest are also shown. (Mean of class B ranged between 0.98 – 0.99 and standard deviation <0.06)

Taxonomic Group / Species Name	3 H L N			10 H L N			20 H L N		
	% Idied	Mean	S.D.	% Idied	Mean	S.D.	% Idied	Mean	S.D.
Cryptomonad									
Chroomonas sp.	88.6	0.756	0.232	89.0	0.821	0.264	87.6	0.825	0.294
Chroomonas salina	88.4	0.740	0.225	88.0	0.840	0.292	84.0	0.802	0.299
Cryptomonas appendiculata	90.2	0.758	0.206	91.4	0.857	0.243	91.8	0.878	0.248
Cryptomonas calceiformis	75.8	0.581	0.197	82.2	0.809	0.319	76.8	0.776	0.333
Cryptomonas maculata	63.2	0.541	0.247	76.0	0.710	0.335	75.8	0.720	0.356
Cryptomonas reticulata	91.2	0.791	0.226	87.0	0.843	0.312	86.8	0.849	0.301
Cryptomonas rostrella	92.0	0.917	0.257	90.4	0.898	0.259	90.6	0.894	0.267
Hemiselmis brunnescens	50.6	0.458	0.159	44.4	0.466	0.194	51.0	0.477	0.193
Hemiselmis rufescens	23.2	0.388	0.125	35.6	0.433	0.185	44.0	0.458	0.228
Hemiselmis virescens	88.6	0.794	0.242	87.0	0.828	0.274	87.8	0.835	0.272
Plagioselmis punctata	93.2	0.813	0.210	90.0	0.834	0.261	86.2	0.817	0.286
Rhodomonas sp.	90.2	0.733	0.205	84.8	0.810	0.316	82.2	0.812	0.333
Diatom									
Amphora coffaeformis	66.2	0.495	0.172	73.8	0.687	0.337	68.0	0.700	0.354
Chaetoceros calcitrans	56.8	0.544	0.215	74.0	0.654	0.259	67.2	0.653	0.307
Phaeodactylum tricornutum	76.4	0.625	0.230	74.0	0.719	0.318	84.0	0.720	0.251
Skeletonema costatum	3.0	0.320	0.161	25.0	0.352	0.195	35.8	0.450	0.283
Thalassiosira weissflogii	40.0	0.482	0.196	61.2	0.571	0.288	68.0	0.649	0.323
Dinoflagellate									
Amphidinium carterae	0.0	0.200	0.065	15.8	0.357	0.148	46.4	0.446	0.258
Aureodinium pigmentosum	22.2	0.402	0.120	51.6	0.495	0.279	47.0	0.496	0.238
Gymnodinium micrum	0.0	0.170	0.064	16.0	0.310	0.171	34.6	0.379	0.214
Gymnodinium simplex	1.6	0.306	0.114	5.6	0.297	0.111	25.4	0.376	0.170
Gymnodinium veneficum	0.0	0.228	0.067	0.6	0.296	0.118	18.8	0.361	0.162
Gymnodinium vitiligo	7.2	0.386	0.109	43.6	0.440	0.165	39.6	0.446	0.153
Gyrodinium aureolum	57.0	0.458	0.225	65.8	0.626	0.336	75.8	0.691	0.322
Heterocapsa triquetra	0.0	0.252	0.083	50.4	0.499	0.240	51.6	0.536	0.291
Prorocentrum balticum	0.0	0.283	0.111	29.6	0.392	0.244	51.6	0.491	0.279
Prorocentrum micans	69.0	0.533	0.224	61.2	0.567	0.308	59.4	0.595	0.304
Prorocentrum minimum	0.0	0.205	0.084	40.2	0.382	0.248	41.2	0.398	0.279
Prorocentrum nanum	0.0	0.265	0.075	16.2	0.341	0.162	43.6	0.443	0.234
Scrippsiella trochoidea	0.0	0.342	0.097	59.6	0.525	0.245	60.8	0.580	0.299

Table 4.2 continued.....

Taxonomic Group / Species Name	3 H L N			10 H L N			20 H L N		
	% Idied	Mean	S.D.	% Idied	Mean	S.D.	% Idied	Mean	S.D.
Flagellate									
Chlorella salina	0.0	0.141	0.045	0.0	0.241	0.134	26.4	0.323	0.212
Chlamydomonas reginae	79.2	0.577	0.210	75.6	0.670	0.292	78.4	0.768	0.353
Dunaliella minuta	0.0	0.216	0.080	45.6	0.418	0.263	43.4	0.425	0.282
Dunaliella primolecta	68.8	0.561	0.213	76.0	0.637	0.219	70.0	0.610	0.269
Dunaliella tertiolecta	42.4	0.451	0.128	68.8	0.564	0.228	67.2	0.655	0.313
Micromonas pusilla	94.4	0.761	0.143	84.6	0.785	0.273	86.6	0.784	0.263
Nephroselmis pyriformis	12.6	0.375	0.117	25.4	0.369	0.161	37.0	0.433	0.160
Nephroselmis rotunda	0.0	0.309	0.100	25.8	0.380	0.159	26.4	0.380	0.167
Ochromonas sp.	0.0	0.127	0.062	0.0	0.178	0.083	2.4	0.252	0.140
Pseudopedinella sp.	0.0	0.227	0.062	13.6	0.366	0.128	28.2	0.380	0.170
Pelagococcus subviridis	75.8	0.598	0.205	79.2	0.753	0.332	78.8	0.774	0.340
Porphyridium pupureum	92.2	0.910	0.285	93.4	0.912	0.265	91.8	0.899	0.267
Pyramimonas grossii	0.0	0.184	0.059	25.2	0.368	0.181	30.4	0.372	0.193
Pyramimonas obovata	0.0	0.225	0.085	0.0	0.275	0.123	26.2	0.353	0.193
Rhodella maculata	91.6	0.918	0.270	92.0	0.904	0.275	93.0	0.945	0.239
Stichococcus bacillaris	0.0	0.239	0.090	29.4	0.030	0.029	42.6	0.437	0.280
Tetraselmis impellucida	84.2	0.770	0.317	82.0	0.833	0.351	84.4	0.854	0.310
Tetraselmis striata	0.0	0.214	0.097	0.0	0.252	0.115	46.6	0.465	0.283
Tetraselmis suecica	0.0	0.243	0.088	68.2	0.563	0.251	67.0	0.590	0.269
Tetraselmis tetrathele	88.6	0.801	0.226	88.2	0.834	0.261	84.0	0.785	0.258
Tetraselmis verrucosa	1.2	0.315	0.109	39.2	0.426	0.186	45.6	0.464	0.225
Prymnesiomonad									
Chrysochromulina camella	0.0	0.319	0.085	52.8	0.493	0.222	47.4	0.551	0.308
Chrysochromulina chiton	0.0	0.265	0.102	1.6	0.301	0.115	29.4	0.374	0.217
Chrysochromulina cymbium	1.0	0.210	0.104	3.2	0.259	0.123	8.6	0.307	0.147
Chrysochromulina polylepis	1.0	0.286	0.112	21.8	0.361	0.178	32.2	0.393	0.190
Emiliania huxleyi 92	0.0	0.229	0.070	33.2	0.412	0.210	47.6	0.499	0.276
Emiliania huxleyi B11	87.4	0.893	0.293	92.6	0.909	0.277	92.6	0.912	0.267
Ochrosphaera neopolitana	0.0	0.143	0.045	0.0	0.187	0.087	1.8	0.190	0.130
Pavlova lutheri	14.0	0.356	0.105	34.4	0.431	0.147	48.6	0.497	0.249
Phaeocystis pouchetii	0.0	0.175	0.076	0.0	0.224	0.107	9.6	0.282	0.158
Pleurochrysis carterae	80.8	0.725	0.244	81.4	0.780	0.310	88.8	0.817	0.253
Prymnesium parvum	0.0	0.300	0.101	21.8	0.358	0.164	45.4	0.479	0.262

Table 4.3 Results of decision processes, RBF and maximum valued output, from the combinations of the various sets of trained single species network outputs. Architecture shown for the RBF decision networks which employed a Euclidean distance metric and 300 events per class.

Data Set	Single Networks Events		Single Networks Nodes		RBF Identification Network			Maximum Output	
	Class A	Class B	Class A	Class B	Nodes	% Corr	Conf	% Corr	Conf
A	400	24400	3	1	3	76.2	76	69.3	69.7
					4	76.1	76		
					5	77.1	76.2		
			10	1	3	76.4	76.5	74.7	74.5
					4	76	76.1		
					5	75.8	76		
			20	1	3	77.2	77.4	76.1	76.1
					4	76.2	76.5		
					5	75.9	76.2		
B	500	30500	3	1	3	76.2	76.6	687	69.9
					4	76.1	76		
					5	75.8	75.9		
			10	1	3	77.4	77.5	74.2	74.3
					4	77.2	77.4		
					5	76.3	76.7		
			20	1	3	77.3	77.6	76.7	76.6
					4	75.1	75.9		
					5	74.9	75		
C	500	500	3	3	3	77.1	77.9	77	77.1
					4	78.5	77.8		
					5	78.2	77.8		
D	1000	30500	3	1	3	76.2	76.4	69.2	69.3
					4	76.1	76		
					5	75.9	75.8		
			10	1	3	76.9	77.2	74.9	75.8
					4	76.8	77.2		
					5	77.3	77.5		
			20	1	3	77.9	77.4	76.7	76.8
					4	77.6	77.6		
					5	76.8	77		
E	500	15250	3	1	3	76.6	76.4	69.2	69.3
					4	76.5	76.1		
					5	76.1	76		
			10	1	3	77.9	78	77	76.8
					4	77.7	78.2		
					5	77.2	77.3		
			20	1	3	77.6	77.2	76.7	76.8
					4	77	77.6		
					5	76.8	77		
F	500	61000	3	1	3	76	75.9	69	68.7
					4	76.1	76		
					5	77.1	76.2		
			10	1	3	77.1	77	73.9	73.5
					4	77	76		
					5	76.9	76.1		

Table 4.3 continued....

Data Set	Single Networks				RBF Identification Network			Maximum Output	
	Events		Nodes						
	Class A	Class B	Class A	Class B	Nodes	% Idied	Conf	% Idied	Conf
F	500	61000	20	1	3	77.3	77.6	75.1	75.6
					4	75.1	75.9		
					5	74.9	75		
G	1000 (500 *2)	30500	3	1	3	76.4	76.5	69.3	69.1
					4	76.3	76.2		
					5	75.9	76		
			10	1	3	77.9	78.5	77.1	77.3
					4	77.9	78		
					5	77.6	77.7		
			20	1	3	77.5	78	77.9	78
					4	78	78.2		
					5	77.6	77.7		
H	30500 (1000 *30.5)	30500	5	5	3	77.8	77.6	77	76.9
					4	77.5	77.6		
					5	77.2	77.1		
			10	10	3	78	78	77.6	78.7
					4	78.1	77.9		
					5	78	77.5		

approximately 75% and 78% for both identification and confidence, with the greater number of nodes again having the bigger margin between training and test data. The methods give individual identification values of 20% and above for all classes, with the lower values representing species for whom allocation was minimal by the single species networks. For example, *Chlorella salina* and *Ochrosphaera neopolitana* are identified 42.8% and 20.2% correct respectively (Table 4.4), both of which have zero allocation from their respective single species networks.

4.7.3 RBF Network Decision

Using a Mahalanobis distance metric in the identification network, had little effect on the overall performance when compared to that of a Euclidean distance metric. The maximum increase achieved between all architectures was 1%, with a high increase in training time (Section 4.11.3). Using 100 events per class produced a difference of 3-4% between training and test data. An average correct overall identification of 77.3% was produced from those networks trained using 200, 300 and 400 events per class. Overall identification and confidence of identification are shown for those networks employing a Euclidean distance metric with 300 events per class (Table 4.3). No significant improvement was evident when node numbers were increased from 3 to 5, where in some cases 5 nodes had a slight detrimental affect on performance (1 node per class identifies approximately 72% correct – results not shown). There also appeared to be slight overparametisation with 5 hidden layer nodes, increasing the division between training and test data results. Identification of individual species varies between groups with highs of 99%, mainly in the Cryptomonads and with *Chrysochromulina cymbium* the poorest at 35.3% (Table 4.4).

4.8 Discussion

4.8.1 Single Species Networks

Performance of the single species networks varies considerably when the imbalance ratio is high. Many of the poorly identified species are from the same genus or have exhibited overlap in the past, indicating the possibility of misidentification with the background class. In comparison to the original multi-class architecture, the identification of individual species from the single species networks are low, but all are of a similar

Table 4.4 Identification and confidence of identification for individual species, by the two decision processes for combining the outputs of one of the sets of single species networks. The single species networks were trained using 500 events for the class of interest with 10 hidden layer nodes and 500 events per species for the background class. The RBF identification network used 300 events per class and 3 hidden layer nodes, employing a Euclidean distance metric. Overall identification was 77.4% for the RBF decision network and 74.2% for the maximum valued output method.

Taxonomic Group and Species Name	Decision network		Maximum valued output		Taxonomic Group and Species Name	Decision network		Maximum valued output	
	Corr	Conf	Corr	Conf		Corr	Conf	Corr	Conf
Cryptomonads					**Prymnesiomonads**				
Chroomonas sp.	95.0	94.5	95.2	96.7	*Chrysochromulina camella*	81.2	80.9	82	61.1
Chroomonas salina	90.0	94.7	93	94.5	*Chrysochromulina chiton*	58.5	59.0	56	56.5
Cryptomonas appendiculata	99.5	84.6	99.2	92.8	*Chrysochromulina cymbium*	35.3	54.3	31	44.6
Cryptomonas calceiformis	91.8	96.0	94	94.3	*Chrysochromulina polylepis*	63.8	59.6	63.6	54.9
Cryptomonas maculata	86.5	96.4	88.2	97.7	*Emiliania huxleyi 92*	80.3	73.1	78.4	62.0
Cryptomonas reticulata	91.6	95.8	90.8	98.7	*Emiliania huxleyi B11*	97.3	97.3	96.6	98.5
Cryptomonas rostrella	99.0	99.5	99	99.2	*Ochrosphaera neopolitana*	51.3	58.1	20.2	58.3
Hemiselmis brunnescens	54.6	58.2	57.6	58.4	*Pavlova lutheri*	76.0	69.8	85.2	58.5
Hemiselmis rufescens	52.3	60.4	52.6	58.5	*Phaeocystis pouchetii*	58.3	58.1	39.2	56.4
Hemiselmis virescens	97.0	89.7	97.6	89.0	*Pleurochrysis carterae*	86.2	91.0	89.8	88.2
Plagioselmis punctata	98.3	84.8	95	88.9	*Prymnesium parvum*	79.0	69.7	78.8	62.0
Rhodomonas sp.	93.1	98.7	94.2	96.3					
Average	**87.4**	**87.8**	**88.0**	**88.7**	**Average**	**69.8**	**70.1**	**65.5**	**63.78**
Flagellates					**Diatoms**				
Chlamydomonas reginae	87.0	93.7	89.8	80.2	*Amphora coffaeformis*	83.3	87.3	86	82.8
Chlorella salina	43.2	59.7	42.8	62.4	*Chaetoceros calcitrans*	83.2	85.3	86.6	79.6
Dunaliella minuta	71.7	81.7	65.8	76.0	*Phaeodactylum tricornutum*	91.5	88.0	91.8	83.6
Dunaliella primolecta	86.2	80.5	92.0	77.4	*Skeletonema costatum*	63.5	77.2	62.4	73.7
Dunaliella tertiolecta	78.3	80.1	78.8	79.9	*Thalassiosira weissflogii*	80.7	78.4	87.8	70.5
Micromonas pusilla	98.2	81.6	98.0	85.5					
Nephroselmis pyriformis	75.0	61.9	53.0	62.2	**Average**	**80.4**	**83.2**	**82.9**	**78.08**
Nephroselmis rotunda	42.0	67.9	47.2	55.9					
Ochromonas sp.	64.2	41.8	56.2	52.5	**Dinoflagellates**				
Pelagococcus subviridis	85.7	88.9	87.2	84.5	*Amphidinium carterae*	74.5	60.8	71.4	51.2
Porphyridium pupureum	95.8	97.3	95.0	99.0	*Aureodinium pigmentosum*	78.7	76.5	80.8	67.2
Pseudopedinella sp.	75.2	66.1	65.8	65.0	*Gymnodinium micrum*	58.2	59.5	60.2	57.2
Pyramimonas grossii	72.0	68.6	65.2	59.1	*Gymnodinium simplex*	62.2	68.6	53.2	68.2
Pyramimonas obovata	58.7	61.2	42.4	66.5	*Gymnodinium veneficum*	38.8	59.1	21.0	63.3
Rhodella maculata	98.5	84.4	98.2	99.4	*Gymnodinium vitiligo*	68.8	60.7	70.2	53.3
Stichococcus bacillaris	63.0	70.8	63.4	61.1	*Gyrodinium aureolum*	89.5	88.2	88.2	86.5
Tetraselmis impellucida	95.8	94.3	95.0	96.0	*Heterocapsa triquetra*	80.2	75.4	76.0	72.7
Tetraselmis striata	77.2	70.8	50.2	83.1	*Prorocentrum balticum*	77.0	82.4	75.6	79.7
Tetraselmis suecica	85.7	85.4	84.8	72.2	*Prorocentrum micans*	80.7	63.4	81.2	67.1
Tetraselmis tetrathele	95.5	92.1	95.0	88.0	*Prorocentrum minimum*	60.3	73.1	59.6	75.6
Tetraselmis verrucosa	66.8	74.1	64.8	65.9	*Prorocentrum nanum*	75.2	74.4	72.4	60.1
					Scrippsiella trochoidea	75.5	78.2	68.4	72.6
Average	**76.9**	**76.3**	**72.9**	**74.8**	**Average**	**70.7**	**70.7**	**67.6**	**67.3**

magnitude below their respective identification by the multi-class network, thus preserving the relationship between good, average and poorly identified classes. Although the kernels representing the class of interest are placed directly into the class, this placement is random, and with no optimisation the centre's are representing the data within their radius. Naturally, with a greater imbalance between event numbers, there is a high possibility that the data surrounding the kernels are from the background class. Consequently, despite placement of only one centre directly into this class, it will inevitably be represented by the nodes placed at position vectors of the class of interest. Identification will be highest for those species with optical parameters that exhibit tight distinct clusters in hyper-dimensional space. The reason being, kernels placed directly into the class of interest will have a better chance of representing the species, if presented primarily with data points from that class. With a more diverse looser cluster, the input vector that falls within a centres radius, has a greater chance of membership to the background class. As expected, as the imbalance ratio between the dominant and subordinate class numbers decreases, the individual and overall identification increases.

4.8.2 Output Combinations

Overall identification and confidence of identification are higher by the RBF decision network than through the Maximum Valued Output method (Table 4.3). The range of values produced by the latter is wider than that of the decision network. This variance appears dependant upon the architecture of the single species networks and, to a lesser extent, event numbers. As the number of hidden layer nodes in the class of interest increase, there is a greater chance of patterns from this class falling within the radius of the kernels and producing a higher output. This is evident from the results of the individual species networks, where an increase in identification through node addition produces a high mean output value and has a widening effect on the standard deviation. This improvement with additional node numbers is good, until the difference between training and test data performance becomes too wide, indicating reduction in the network's generalisation and the introduction of memorisation. This has a direct effect on the maximum valued output method. Despite individual species having low or zero identification by their respective single species networks, they are identified, although in some cases poorly, by the maximum valued output method. This indicates that the range

of output values from the class of interest, depicted by the mean to be less than 0.5, are in fact greater than the output produced by other single species networks when presented with the same input pattern.

The relationship between the architecture of the single species networks and the RBF decision network is much less important. Identification by the network is not only higher than the other method, but the range of overall values produced by varying architectures, of both the RBF decision network and the single species networks is minimal. This would indicate, that the signature input vector produced by the single species networks for a particular pattern, has little dependence upon the architecture of the networks, or the number of events. The small difference between Mahalanobis and Euclidean distance metrics for overall identification, would imply that the clusters are more spherical than elliptical and therefore do not require the more complex modelling of the Mahalanobis distance metric. The lack of improvement in increasing the event numbers from 200 to 400 per class, indicates coverage of the 62 parameter signatures within 200 events.

4.9 Conclusion

There are a number of approaches that can be considered to improve the individual performance of the single species networks. The affect of increasing event numbers and hidden layer nodes has already been demonstrated. Using further optimisation procedures, such as LVQ kernel placement, or a Mahalanobis distance metric would also have positive effects. However, the importance of the single species networks lies in how their results influence the combining procedure for final identification, and not the individual identification of species by the networks themselves.

Although the individual identification of many species are comparable in both methods (Table 4.4), overall identification by the RBF decision network is always higher and, more importantly, it is consistent. Performance of the maximum valued output method is dependent upon the architecture of the single species networks and the data file structure. It is also envisaged that addition of new species, increasing overlap, will have a direct effect on this method.

These areas are not an issue when combining the outputs through the RBF decision network, where the architecture of the single species networks and their training files has

negligible influence on final performance. Providing there is consistency across the architectures of the single species networks, the range of the outputs used as signature input vectors for the RBF should be preserved. This negates the need for each network to perform to the same equal level, but only to their own individual optimum.

4.10 Multiple Network Architecture - Evaluation

In order to compare and evaluate the efficiency of the multiple network approach to that of the original multi-class architecture, a number of studies were performed, some of which were similar to those in Chapter 3. The single species networks for set B (Section 4.7) were considered as a sufficient worst case, where biological variation of the background class is fully represented.

4.10.1 Species Combinations

The outputs produced for particular classes, from each of the 62 single species networks, were combined into individual training files for a RBF decision network. As with the multi-class architecture, three areas of overlap were considered. Initially, the 62 parameter outputs for each particular genus were combined into separate groups, producing 37 classes of 300 events each, with some groups containing only one species. Secondly, those species within a genus, whose mutual misidentification from the optimum network for set B was greater than 5%, were combined, forming 55 classes with 300 events per class. Finally, a confusion dendrogram (Chapter 3, Section 3.8.1.5) from the same optimum network was produced (Fig. 4.2). The three points chosen combined species into 55, 46 and 37 groups, with 300 events per class.

All networks used 3 hidden layer nodes per class, employed a Euclidean distance metric and were trained three times from different initialisation points. An independent test set of 500 events per species was constructed to assess performance.

4.10.2 Rejection of Unknowns

The importance of unknown species rejection and its influence on performance has been discussed in Chapter 2 (Section 2.7.4.2). To assess the multiple network's ability to reject unknowns, 12 completely novel species (500 events per class) were added to the test file (Table 4.5). The trained 62 single species networks (set B) were presented with the file

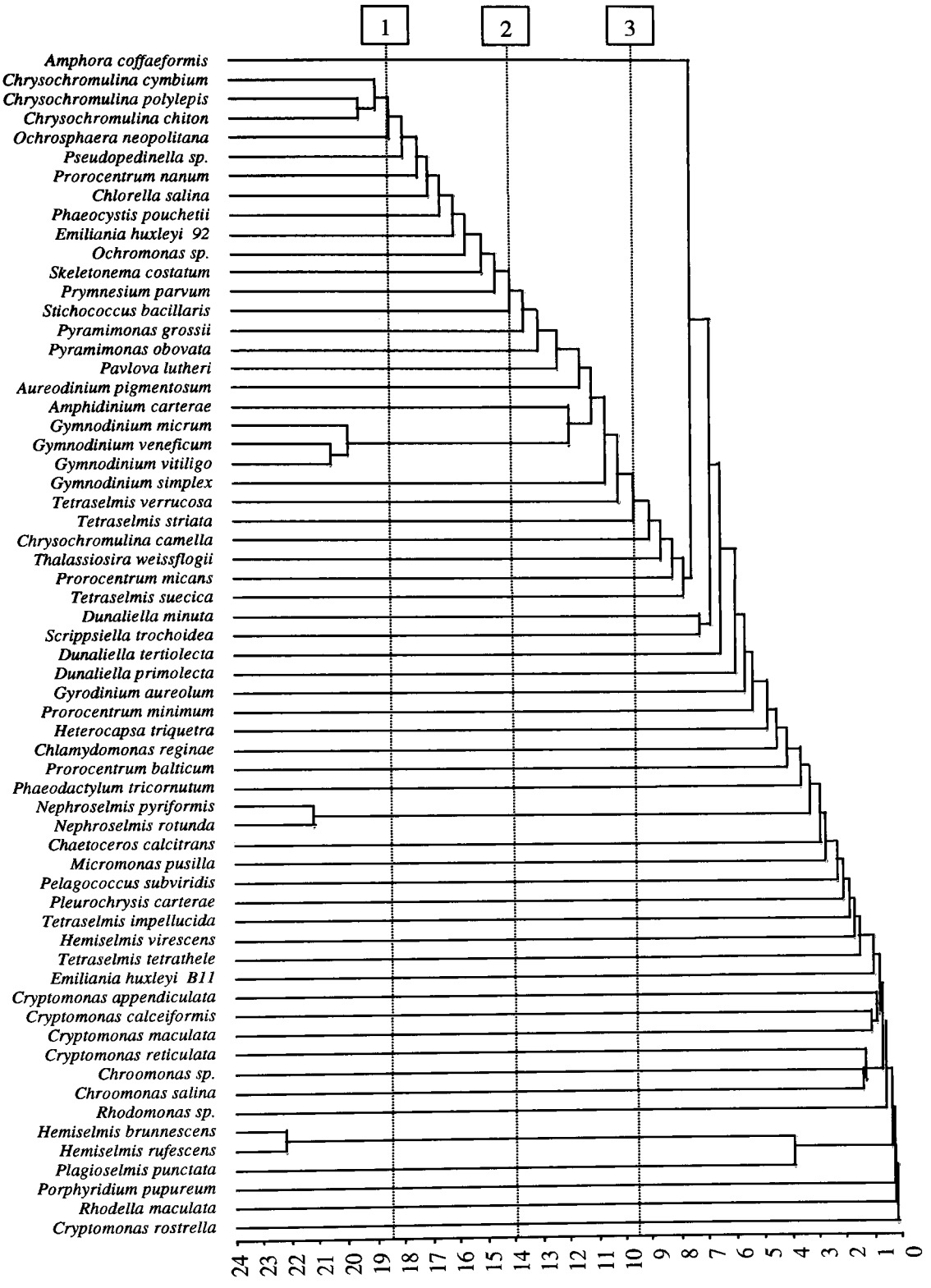


Figure 4.2 Dendrogram showing the order in which respective species were clustered Clustering proceeds from left to right with the ordinate axis showing the percentage of misidentified data remaining at each clustering stage. The positions marked 1, 2 and 3 on the dendrogram indicate the 55, 46 and 37 groups respectively for species combination.

Table 4.5 Twelve species of phytoplankton used as unknowns to assess the ability of the alternative multiple network architecture to reject novel species. Species also used to assess addition of classes to the existing database.

Taxonomic Group	Species Name	Order	Size μm
Diatom	*Chaetoceros affinis*	Bacillariophyceae	10->100
	Chaetoceros debilis	"	10->100
	Chaetoceros radicans	"	10->100
	Surirella sp.	"	10->100
Dinoflagellate	*Alexandrium lusitanicum*	Dinoflagellida	25-40
	Alexandrium tamarense	"	28-40
Flagellate	*Nannochloris atomus*	Volvocida	2-4
Prymnesiomonad	*Chrysotila lamellosa*	Prymnesiida	4-7
	Dicrateria inornata	"	3-5
	Imantonia	"	2-4
	Isochrysis galbana	"	4-8
	Platychrysis	"	8-10

containing the additional species, and a 62 parameter input file was constructed from the combined outputs. Three criteria for rejecting novel data were imposed upon the RBF decision network trained in the previous study on the data set B (Section 4.6.4).

1. Rejection if the maximum hidden layer node output is less than a threshold *T1*, where *T1* was varied from 0 to 0.9 in intervals of 0.1
2. Rejection if the difference between the two highest outputs is less than a threshold *T2*, where *T2* was varied from 0 to 0.9 in intervals of 0.1
3. Rejection if the highest valued output is less than a threshold *T3*, where *T3* was varied from 0 to 1 in intervals of 0.1

4.10.3 Dynamic Selection of Species

When comparing this research to previous work, the number of phytoplankton species assessed here is greater, making identification more difficult. This is still only a fraction of the phytoplankton community, which increases not only because of different species, but variation within individual species. However, a user may not require all species present in the database for a particular identification. For example, certain species may be known not to exist in a particular body of water, and may therefore be eliminated from the analysis, possibly reducing misidentifications.

As the single species networks are independent of each other, once trained, a subset can be chosen and their outputs combined. These can then be used as input data to train an RBF decision network specific to the particular species. Hence, the input file will comprise signature vectors whose dimensional value is equal to the number of single species networks originally chosen. As this allows dynamic control over the number selected, a user has the option of including extra single species network outputs in the input file, therefore increasing vector dimension and possibly adding discriminatory information, if needed.

In order to compare performance and training time to the original multi-class network, both architectures were trained with 5 to 60 classes increasing in steps of 5, as well as the full compliment (62 classes). As the database is constructed of overlapping and distinct species, random selection of individuals may be biased, therefore two data sets were constructed for each step of 5 to 62 classes. Using the dendrogram (Fig. 4.2), species were selected from left to right (high values of misidentification) to construct overlapping

data sets and from right to left (low values of misidentification) to construct separable data sets. Three hidden layer nodes employing both Euclidean and Mahalanobis distance metrics and 300 events per class, were used for each architecture to allow full comparison. Only 1 iteration was used for all to ensure comparability between time and performance.

4.10.4 Addition of Novel Species

The primary advantage of the multiple network approach is the ease of addition of a new species. To investigate this, 12 additional single species networks were trained using the twelve unknown species as classes of interest (Table 4.5). In each case, 500 events were used for class A and 30500 events in total for class B, chosen randomly from the original species database. The networks were trained using 10 hidden layer nodes placed directly into class A and one node in class B. An unseen test set of 500 events per species, including both the original 62 species and the 12 novel species, was applied to each of the 74 trained single species networks, 62 of which had already been trained from the initial study. The outputs from the networks were combined into 74 files, containing 74 parameters and 500 events. Three RBF decision networks were trained with 300 events and 3, 4 and 5 hidden layer nodes per class, employing a Euclidean distance metric. An independent test set was constructed, and the outputs from applying it to the 74 trained single species networks were combined to test the performance of the RBF decision networks. As a comparison, an original multi-class network was trained with 500 events per class for the 74 classes. Training was initiated with 6 hidden layer nodes per class employing a Mahalanobis distance metric. Maximum valued output performance was also determined.

4.11 Results

4.11.1 Species Combinations

Overall identification when species were grouped into their respective genera was 78.38%, compared to a value of 85.25% produced from the same number of groups constructed from the dendrogram (Table 4.6). Species whose mutual misidentification within a genus was greater than 5%, identified to 80.22%. The 46 and 55 dendrogram groupings produced an identification of 82% and 81% respectively. Overall confidence of correct identification was improved in all cases by the species combinations.

Table 4.6 Overall identification and confidence of identification by the RBF decision network produced from combining species.

Groups	Number of Groups	Overall identification	Confidence of identification
All Species Separate	62	77.4	77.22
Genus' grouped	37	78.38	81.17
5% Misidentity within a Taxonomic group	55	80.22	81.53
Dendrogram groupings 1	55	81	82.13
Dendrogram groupings 2	46	82.05	86.12
Dendrogram groupings 3	37	85.25	90.86

4.11.2 Rejection of Unknowns

Rejection criteria, based on maximum hidden layer node outputs, produced very poor results for rejection of both known and unknown species (Fig. 4.3a). No rejection is evident until the threshold reaches 0.6, and progression to 0.9 rejects more knowns than unknowns. The difference between the two highest output node values gives better results, but still rejects a high number of knowns at a low threshold (Fig. 4.3b). At a value of 0.3, 34% of knowns are rejected and 68% of unknowns. The threshold imposed on the highest valued output provides the best rejection criteria, with only 12% of knowns rejected at a threshold of 0.4, compared to 60% of unknowns (Fig. 4.3c). Overall performance drops for all thresholds as the number of knowns rejected increases, whereas confidence of identification increases, as both misidentified knowns and unknowns are rejected. Extreme species rejections vary, with the exception of *Isochrysis galbana* being the least rejected unknown, due to its high misidentity with *Hemiselmis virescens*.

4.11.3 Dynamic Selection of Species

Low training times were recorded for both the original multi-class networks employing Mahalanobis and Euclidean distance metrics, and the alternative multiple network approach using a Euclidean distance metric (Fig. 4.4). Training times were high for the multiple networks architecture employing a Mahalanobis distance metric, with 62 classes taking just over 5 hours. With all networks, identification of small numbers of separable species produces fairly similar identification values, starting at 95% and above, for 5 classes and decreasing as the overlap increases with additional species (Fig. 4.5).

Naturally, as depicted by the dendrogram, identification of the overlapped data sets are less than for that of the separable species. The original multi-class architecture, using a Euclidean distance metric, gave the poorest performance, producing overall identification of 72% for 62 species. For separable species, the original multi-class network exhibits a 2% to 4% difference in performance of the two distance metrics and a 3% to 5% difference for the overlapping species sets. Using the multiple networks architecture, the difference, for both the separable and overlapped data sets, is between 0 and 2%, with an average of 0.9% difference. Overall identification for the 62 species using the alternative multiple network architecture is 77.1% and 77.6%, for the Euclidean and Mahalanobis distance metrics respectively, compared to 71.7% and 76% by the original multi-class architecture.

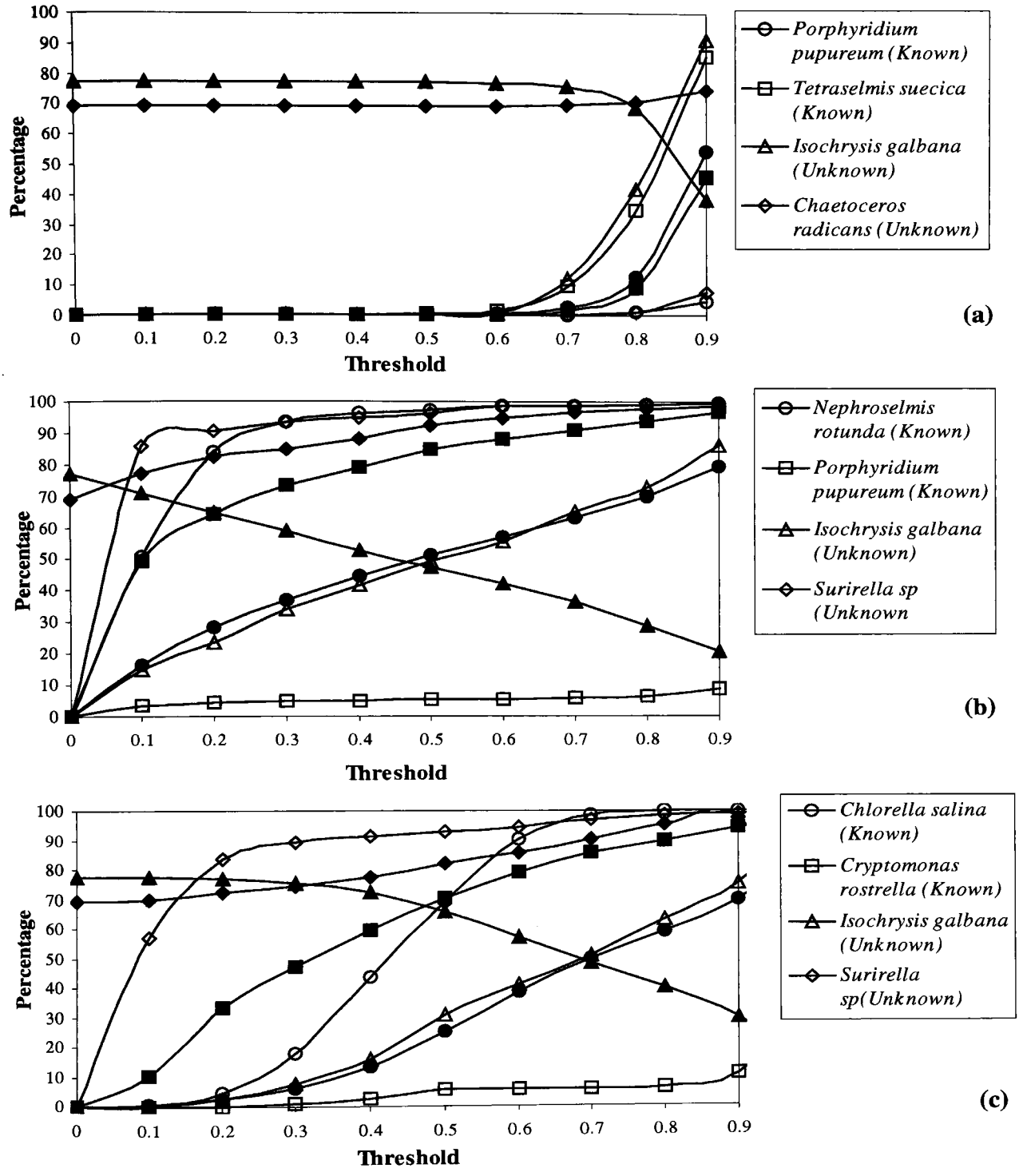


Figure 4.3 Percentage of known and unknown species rejected, as well as overall identification and confidence of identification for the three rejection criteria. Extreme rejections for both known and unknown species is also shown. ● Overall rejection of known species ■ Overall rejection of unknown species ▲ Overall percentage of correct identification ◆ Overall confidence of correct identification (a) Maximum hidden layer node is less than a threshold *T1*. (b) Difference between highest and second highest node is less than a threshold *T2* (c) Highest valued output node is less than a threshold *T3*.

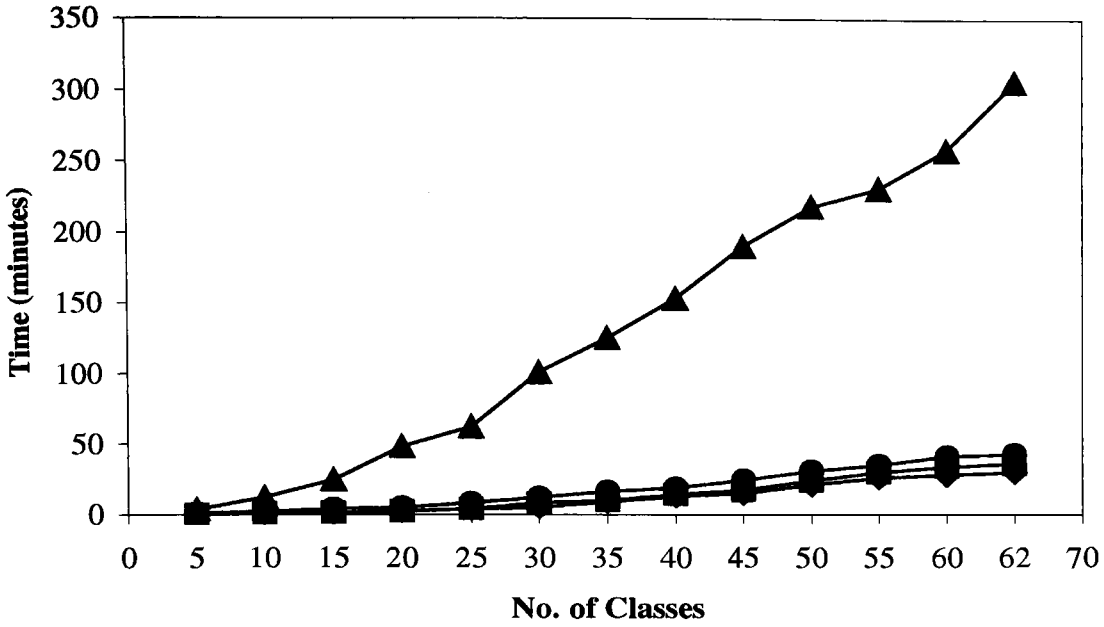


Figure 4.4 Training times for both network architectures employing each distance metric using 3 hidden layer nodes and 300 events per class. Alternative multiple architecture employing ▲ Mahalanobis distance metric ● Euclidean distance metric. Original multi-class architecture employing ■ Mahalanobis distance metric ◆Euclidean distance metric.

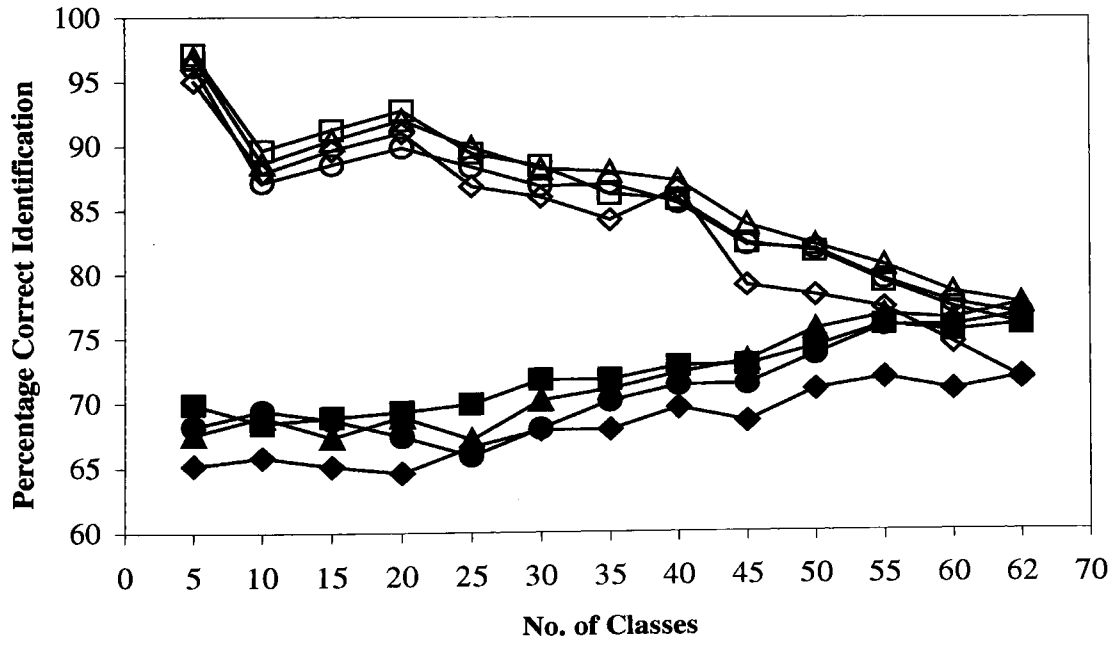


Figure 4.5 Percentage correct identification for the separable (empty markers) and overlapping data (solid markers) for both architectures (n.b. scale is exaggerated). Alternative multiple architecture employing ▲ Mahalanobis distance metric ● Euclidean distance metric. Original multi-class architecture employing ■ Mahalanobis distance metric ◆ Euclidean distance metric.

4.11.4 Addition of Novel Species

For 7 out of the 12 unknowns, identification by the single species networks was less than 60% (Table 4.7). The RBF decision network, however, identified 10 of the species at approximately 60% or above, with the exception of *Platychrysis* and *Chaetoceros affinis*, which identified to 29.2% and 33.8% respectively (Table 4.8). *Platychrysis* misidentifies as *Chaetoceros debilis* (36%) and *Chrysotila lamellosa* (12.5%), as well as various others. This is mutual, to a lesser extent, with *Chaetoceros debilis* misidentifying as *Platychrysis* 15.5%. *Chaetoceros affinis* misidentifies as *Chaetoceros radicans* 16.5% and as *Amphora coffaeformis* 20%, reducing the confidence of this original species. Various other species misidentify between both knowns and unknowns, reducing confidences and identifications. However, the greatest misidentities were between species for whom overlap was already high, for example, *Hemiselmis brunnescens* with *Hemiselmis virescens*, and *Nephroselmis pyriformis* with *Nephroselmis rotunda*. The overall identification of the network was 75.3% with a confidence of identification of the same value, compared to 72% correct and confidence of 72% by the maximum valued output method and 76% correct and confidence of 75.1% by the original multi-class network (results not shown).

4.12 Discussion

4.12.1 Species Combination

As expected, the grouping together of certain species where overlap is great will naturally improve network performance. The difference in identification of genus combinations to the groupings produced at position 1 on the dendrogram, indicates that the generated 62 parameter inputs, like the genus groupings of the seven-dimensional optical data (Chapter 3, Section 3.8.3.5), are not necessarily where the overlap lies. The confidence of identification is significantly improved as the misidentity of species is reduced, due to their coupling with overlapping classes.

Dendrogram groupings shown here are very similar to that of the dendrogram produced for the multi-class network (Chapter 3) and, although there are a few minor discrepancies, the degrees of overlap are almost identical, even if the order is slightly different. Whilst again illustrating the lack of similarity between the flow cytometric

Table 4.7 Results of the single species networks trained for the additional 12 species using 500 events in the class of interest and 500 per species in the background class. The networks were trained using a Euclidean distance metric with 10 hidden layer nodes for the class of interest. The mean and standard deviation of the network output values for the classes of interest are also shown. (Mean output value of class B ranged between 0.98 – 0.99 and standard deviation of output values <0.06)

Taxonomic group	Species name	Percentage identified	Mean	Standard Deviation
Diatom	*Chaetoceros affinis*	0	0.161	0.092
	Chaetoceros debilis	37.6	0.392	0.221
	Chaetoceros radicans	39.2	0.401	0.275
	Surirella sp.	57	0.518	0.226
Dinoflagellate	*Alexandrium lusitanicum*	77.6	0.755	0.333
	Alexandrium tamarense	61.2	0.591	0.310
Flagellate	*Nannochloris atomus*	82.6	0.714	0.235
Prymnesiomonad	*Chrysotila lamellosa*	3.8	0.192	0.139
	Dicrateria inornata	37.8	0.418	0.239
	Imantonia	47.6	0.467	0.158
	Isochrysis galbana	83	0.717	0.246
	Platychrysis	0	0.229	0.117

Table 4.8 Individual identifications and confidence of identification by the decision RBF network trained using the original 62 species and the 12 additional species (shaded). The single networks used 500 events for the class of interest with 10 hidden layer nodes and 500 events per species for the background class. The RBF decision network used 3 hidden layer nodes and 300 events per class (Overall identification of 75.3%). All networks employed a Euclidean distance metric.

Taxonomic group and Species Name	Corr	Conf	Taxonomic group and Species name	Corr	Conf
Prymnesiomonads			**Cryptomonads**		
Chrysochromulina camella	85.5	82.5	*Chroomonas sp.*	96.2	95.7
Chrysochromulina chiton	61.2	60.4	*Chroomonas salina*	91.5	97.7
Chrysochromulina cymbium	37.2	51.3	*Cryptomonas appendiculata*	99.0	94.4
Chrysochromulina polylepis	61.3	60.5	*Cryptomonas calceiformis*	91.0	95.0
Emiliania huxleyi 92	81.8	73.4	*Cryptomonas maculata*	93.2	92.1
Emiliania huxleyi B11	96.8	82.8	*Cryptomonas reticulata*	96.0	96.0
Ochrosphaera neopolitana	46.5	55.1	*Cryptomonas rostrella*	100.0	95.8
Pavlova lutheri	81.8	68.6	*Hemiselmis brunnescens*	24.0	65.3
Phaeocystis pouchetii	56.0	58.8	*Hemiselmis rufescens*	81.0	54.8
Pleurochrysis carterae	87.5	91.8	*Hemiselmis virescens*	92.7	78.5
Prymnesium parvum	85.0	68.1	*Plagioselmis punctata*	94.0	86.7
Chrysotila lamellosa	58.7	62.6	*Rhodomonas sp.*	92.2	91.6
Dicrateria inornata	71.3	77.0			
Imantonia	76.2	70.4	**Average**	**87.6**	**87.0**
Isochrysis galbana	83.5	81.3			
Platychrysis	29.2	50.6	**Diatoms**		
			Amphora coffaeformis	82.8	77.3
Average	**68.7**	**68.5**	*Chaetoceros calcitrans*	89.5	87.3
			Phaeodactylum tricornutum	91.5	83.1
Flagellates			*Skeletonema costatum*	70.2	72.4
Chlamydomonas reginae	89.8	89.6	*Thalassiosira weissflogii*	81.2	79.0
Chlorella salina	54.3	61.2	*Chaetoceros affinis*	33.8	63.6
Dunaliella minuta	73.8	78.3	*Chaetoceros debilis*	63.5	57.7
Dunaliella primolecta	87.7	81.5	*Chaetoceros radicans*	70.2	67.8
Dunaliella tertiolecta	87.7	77.9	*Surirella sp.*	88.2	79.5
Micromonas pusilla	91.7	79.0			
Nephroselmis pyriformis	58.0	59.6	**Average**	**74.5**	**74.2**
Nephroselmis rotunda	50.2	61.9			
Ochromonas sp.	58.2	43.1	**Dinoflagellates**		
Pelagococcus subviridis	83.7	76.3	*Amphidinium carterae*	74.5	64.9
Porphyridium pupureum	94.8	97.9	*Aureodinium pigmentosum*	83.5	72.9
Pseudopedinella sp.	77.5	72.0	*Gymnodinium micrum*	61.3	58.7
Pyramimonas grossii	71.7	71.4	*Gymnodinium simplex*	57.8	71.1
Pyramimonas obovata	50.7	64.4	*Gymnodinium veneficum*	49.7	57.3
Rhodella maculata	98.3	90.4	*Gymnodinium vitiligo*	55.0	59.7
Stichococcus bacillaris	55.8	69.8	*Gyrodinium aureolum*	98.5	90.0
Tetraselmis impellucida	95.0	91.7	*Heterocapsa triquetra*	81.3	73.0
Tetraselmis striata	75.5	70.8	*Prorocentrum balticum*	66.0	80.0
Tetraselmis suecica	82.8	79.9	*Prorocentrum micans*	76.8	74.3
Tetraselmis tetrathele	95.8	91.1	*Prorocentrum minimum*	67.2	80.3
Tetraselmis verrucosa	64.7	77.0	*Prorocentrum nanum*	61.2	67.2
Nannochloris atomus	91.7	84.4	*Scrippsiella trochoidea*	50.8	72.3
			Alexandrium tamarense	91.3	64.0
			Alexandrium lusitanicum	84.0	96.8
Average	**76.8**	**75.9**	**Average**	**70.6**	**72.2**

signatures of some same group/genus species, it also demonstrates that the relationship between the 7 optical characteristics is preserved in the generated 62 parameter data.

4.12.2 Rejection of Unknowns

Although the difference between the two highest output values produces high rejection of unknowns, it also rejects a high number of knowns making it an unsuitable criterion. Unlike the original multi-class network, rejection using hidden layer node thresholds is extremely poor for both the known and unknown species. This indicates the outputs of the Gaussian kernels for both known and unknown species are high. This would imply that the proximity of input signature vectors from the single species networks to kernel centres is, in relation, greater than the proximity of the seven-dimensional optical data to kernel centres in the original multi-class architecture.

The optimum rejection criterion for the 62 parameter data, is threshold imposition upon the highest valued output node. This highlights the importance, and influence, of the weight vectors between the hidden and output layer of the RBF decision network for these data. As the threshold values are increased the percentage of overlap and misidentity is reduced, therefore improving overall confidence of identification at the expense of number of species identified. The importance of this has already been discussed, and the gap between known and unknown species rejection may be increased by using a network trained on overlapping combinations of species, that are known to be hard to differentiate. This will increase confidence of identification with a lesser affect on overall performance.

4.12.3 Dynamic Selection of Species

The spatial distribution of phytoplankton flow cytometric signatures has already been discussed, and the advantages of the Mahalanobis distance metric over the Euclidean distance metric have been established. It is not surprising therefore, that the original multi-class architecture employing a Euclidean distance metric, produces the poorest identification as class numbers increase. Throughout the increase, the difference in performance of the two distance metrics, for the multiple network architecture, demonstrates the unnecessary requirement of the Mahalanobis distance metric in this approach. The Euclidean measure gives almost identical results to that of the Mahalanobis, in a fraction of the time. The comparable identification values between the

two architectures, indicates that the combination of dynamically selected single species networks, perform just as well as the original multi-class network, with the advantage of shorter training times, greater flexibility and less complexity.

It should also be mentioned that, the software used to train the RBF decision networks uses optimisation procedures that are not required for the generally spherical clusters produced from the single species network outputs, therefore training time may in fact be even shorter than recorded here.

4.12.4 Addition of Novel Species

As expected when new species are added, overlap increases, therefore reducing overall identification. Adding new classes to any identification system will have an affect on its performance. The overall identifications are slightly lowered by the increase, but the RBF decision network for combining outputs, still performs better than the maximum valued output method. The original multi-class network, trained on all 74 species, produced an overall identification comparable to that of the alternative multiple network, and exhibited similar areas of overlap between the novel and original species. However, this network took over 2.5 hours to train and optimise to an overall identification of approximately the same value as that of the multiple networks architecture, which took approximately 50 minutes.

The average identifications of the five taxonomic groups (Table 4.8), are relatively similar to their identification without additional species (Table 4.4). The Diatoms are an exception to this, where the average identification was lowered by the four additional Diatoms, due to high misidentification within the genus *Chaetoceros* and species from other groups. The identification of most species is of a similar magnitude when compared to the network trained without the additions. More apparent between these results, is the change in identification of species for whom mutual misidentification was already quite high. As new species are added the misidentities are still evident, but in different proportions. For example, *Hemiselmis brunnescens* misidentifies as *Hemiselmis rufescens* 34% in the network without unknowns (Section 4.7) and the mutual reverse to a value of 37%. Here, mutual misidentification is 69% and 9.5% respectively. The same change in proportion is also true for *Nephroselmis pyriformis* and *Nephroselmis rotunda,* amongst others. This is a consequence of boundary positions being different in the two networks,

where certain data points will now provoke a reaction from the node on the opposite side of its original placement, therefore affecting individual identification and proportion of misidentity, but not total misidentity. This shift is evident between species that constantly misidentify and may never be adequately separated.

The high identification of the novel species, and the consistency in identification and overlap of the original species, indicates the lack of importance of the content of the background class. This will allow single species networks to be trained and stored without the need for re-training when a novel species is encountered.

4.13 Conclusion

When comparing the identification of individual species by the two network architectures, similarities were evident across the groups (Table 4.9). Some differences arose where one performed better than the other for particular species. This is expected with any identification system, but what appears in this instance is that major discrepancies tend to be amongst those species for whom misidentity, mutual or otherwise, is already a factor, and the variance is a proportional balance. The misidentity of these overlapping species can be detrimental to their own individual identification, and the overall identification and confidence of identification of the network. As with the original multi-class approach, this affect can be lessened through combination of similar species and, together with the rejection of unknowns, can be used to improve performance.

The difference in overall performance of the two architectures is negligible. Although individual identifications vary between the two paradigms, the advantages of the alternative multiple networks approach are obvious. As the database of species increases, the original multi-class network will require far more optimisation than the alternative multiple network architecture. The original optical parameters are diverse, elliptical clusters requiring the more complicated Mahalanobis distance metric. The variance of these parameters, though stages of growth and multi-modality *etc.*, are eliminated in the range of values generated for input into the RBF decision network. These, therefore are adequately modelled by the simpler Euclidean distance metric, despite the possibility of a high number of dimensions if dynamic selection dictates it. The addition of a new species to the alternative multiple networks approach requires minimum training time and optimisation, producing a flexible system that can operate in real-time by non-computer

Table 4.9 Comparison of individual identification results for the optimum multi-class network from Chapter 3 (77.7% correct identification) and the optimum alternative multiple networks approach documented here (77.4% correct identification).

Taxonomic Group and Species name	Multiple Network		Original Multi-Class		Taxonomic Group and Species name	Multiple Network		Original Multi-Class	
	Corr	Conf	Corr	Conf		Corr	Conf	Corr	Conf
Cryptomonads					**Prymnesiomonads**				
Chroomonas sp.	95.0	94.5	95.2	97.9	*Chrysochromulina camella*	81.2	80.9	85.4	76.7
Chroomonas salina	90.0	94.7	92.4	96.9	*Chrysochromulina chiton*	58.5	59.0	61.2	60.6
Cryptomonas appendiculata	99.5	84.6	97.6	95.9	*Chrysochromulina cymbium*	35.3	54.3	46.2	56.1
Cryptomonas calceiformis	91.8	96.0	93.6	95.1	*Chrysochromulina polylepis*	63.8	59.6	60	56.8
Cryptomonas maculata	86.5	96.4	90.8	91.3	*Emiliania huxleyi 92*	80.3	73.1	80.8	73.5
Cryptomonas reticulata	91.6	95.8	95	97.5	*Emiliania huxleyi B11*	97.3	97.3	97.6	95.1
Cryptomonas rostrella	99.0	99.5	99.4	94.6	*Ochrosphaera neopolitana*	51.3	58.1	45.6	55.5
Hemiselmis brunnescens	54.6	58.2	65	67.4	*Pavlova lutheri*	76.0	69.8	77.6	70.6
Hemiselmis rufescens	52.3	60.4	64.4	68.8	*Phaeocystis pouchetii*	58.3	58.1	59.4	62.4
Hemiselmis virescens	97.0	89.7	95.8	95.8	*Pleurochrysis carterae*	86.2	91.0	92.2	83.8
Plagioselmis punctata	98.3	84.8	92	86.5	*Prymnesium parvum*	79.0	69.7	79.8	70.9
Rhodomonas sp.	93.1	98.7	93.4	94.2					
Average	**87.4**	**87.8**	**89.55**	**90.16**	**Average**	**69.8**	**70.1**	**71.44**	**69.2**
Flagellates					**Diatoms**				
Chlamydomonas reginae	87.0	93.7	91.8	76.9	*Amphora coffaeformis*	83.3	87.3	88	90.9
Chlorella salina	43.2	59.7	52.2	58.8	*Chaetoceros calcitrans*	83.2	85.3	87.6	83
Dunaliella minuta	71.7	81.7	67.4	75.1	*Phaeodactylum tricornutum*	91.5	88.0	93.4	90.8
Dunaliella primolecta	86.2	80.5	85.2	82.7	*Skeletonema costatum*	63.5	77.2	76.2	76
Dunaliella tertiolecta	78.3	80.1	82.4	75.7	*Thalassiosira weissflogii*	80.7	78.4	92.8	74.5
Micromonas pusilla	98.2	81.6	99.4	81.5					
Nephroselmis pyriformis	75.0	61.9	71	62.6	**Average**	**80.43**	**83.2**	**87.6**	**83**
Nephroselmis rotunda	42.0	67.9	54	61.8					
Ochromonas sp.	64.2	41.8	57.4	67.5	**Dinoflagellates**				
Pelagococcus subviridis	85.7	88.9	87	90.6	*Amphidinium carterae*	74.5	60.8	77.8	72.6
Porphyridium pupureum	95.8	97.3	95.2	97.7	*Aureodinium pigmentosum*	78.7	76.5	88.2	73.4
Pseudopedinella sp.	75.2	66.1	76	69.5	*Gymnodinium micrum*	58.2	59.5	71.2	62.6
Pyramimonas grossii	72.0	68.6	67.4	73.6	*Gymnodinium simplex*	62.2	68.6	69	66.9
Pyramimonas obovata	58.7	61.2	64	65.4	*Gymnodinium veneficum*	38.8	59.1	41.4	68.8
Rhodella maculata	98.5	84.4	93	94.5	*Gymnodinium vitiligo*	68.8	60.7	66.4	63.1
Stichococcus bacillaris	63.0	70.8	67.6	77.3	*Gyrodinium aureolum*	89.5	88.2	86	92.1
Tetraselmis impellucida	95.8	94.3	94.8	93.5	*Heterocapsa triquetra*	80.2	75.4	72.4	79.7
Tetraselmis striata	77.2	70.8	76	71.4	*Prorocentrum balticum*	77.0	82.4	70	73.5
Tetraselmis suecica	85.7	85.4	87	84	*Prorocentrum micans*	80.7	63.4	81.2	59.9
Tetraselmis tetrathele	95.5	92.1	94.6	89.2	*Prorocentrum minimum*	60.3	73.1	61.6	77.6
Tetraselmis verrucosa	66.8	74.1	60.2	71.7	*Prorocentrum nanum*	75.2	74.4	56.4	70.9
					Scrippsiella trochoidea	75.5	78.2	51.2	66.7
Average	**76.9**	**76.3**	**77.31**	**77.19**	**Average**	**70.73**	**70.7**	**68.68**	**71.3**

scientists. A library of constructed pre-trained single species networks, will allow dynamic selection of particular species as and when required, and a simple, rapid, low optimisation procedure for final identification is achieved using the RBF decision network.

5 Classification and Unsupervised Neural Networks

5.1 Introduction

Correct identification to species level by supervised networks is generally high, being poor in species only for whom overlap is such, that separation may never be possible. However, in some cases when the same species are identified to group or genus level the network's performance drops. This is due to the lack of correlation between flow cytometric signatures and the morphometric class labelling of some species (Morris & Boddy, 1995). It seems appropriate that in order to improve the identification of phytoplankton to group or genus level, an *'alternative structure'* needs to be introduced, that is more representative of flow cytometric signatures than phenetic similarities. This chapter investigates the closeness of taxa in terms of recorded optical parameters through the analysis of unsupervised neural networks.

Unsupervised neural networks differ from supervised in their ability to discover for themselves, the simultaneous relationships between variables of multi-dimensional data. They do not require the presence of any external or *a priori* knowledge and are used primarily for classification rather than identification. The Kohonen Self Organising Map (SOM) (Lippman, 1987; Kohonen 1988, 1990) is a topology preserving network, representing a simplified model of the feature mapping that takes place in the brain. It reduces multi-dimensional data to a more easily conceived dimensionality, whilst still attempting to retain the physical configuration of the input space. However, although similar patterns in the input space are mapped to adjacent areas in the output space, the distinction between probable clusters is less than obvious.

This chapter evaluates the network and introduces a number of methods for recognising probable boundaries and hence determining clusters on the SOM, as well as demonstrating the conflicts between morphometric groupings and flow cytometric data.

5.2 Data Classification

Classification is an instinctive process, which has been naturally carried out on animate and inanimate objects for centuries. Elements belonging to the same class do so because of some in-group similarity that is not apparent, or as pronounced, in objects belonging to a different class. These *distinct* sub-sets may subsequently belong to a larger group, for which the in-group similarity is coarser.

Statistical cluster analysis determines possible classes within a data set using algorithms that are either hierarchical or partitional. Both methods involve grouping similar objects into mutually exclusive clusters. Hierarchical methods iteratively form clusters by either merging the data; *i.e.* agglomerative methods, *e.g.* Single linkage, Wards method; or dividing the data; *i.e.* divisive methods *e.g.* Splinter-Average (Krzanowski, 1993); at various levels. Partitional methods differ from hierarchical by allowing a data point to be re-assigned to a different cluster if it becomes apparent that the initial choice was incorrect. They partition the data based on some pre-defined objective function, *e.g.* K-means (Tou & Gonzalez, 1974).

Scale space theory (Witkin, 1983) is an approach employed in a number of clustering fields. This requires scale-space representation of the data and a suitable scale parameter. Cluster number is determined as that number which persists over the greatest range of the defined scale interval (*e.g.* Wong, 1996; Kothari & Pitts, 1999; Pitts *et al.*, 1999).

Data clustering, however, is not an exact science. Despite some techniques possessing strengths over others, they very rarely yield a conclusive guaranteed result. There is no absolute definition of what constitutes a cluster, consequently, for many of the algorithms, the in-group similarity is regarded as a measure of closeness in hyper-dimensional space. Many algorithms are sensitive to initialisation conditions. There are problems with some of these methods when cluster shapes vary. For example, a set of two hypothetical elliptical clusters, discernible to the eye, can be problematic to a method for which cluster detection relies on distance (Fig. 5.1). The same elliptical distribution can cause ambiguous clusters, if an incorrect placement of probable centres within each are subsequently merged, due to the centres' proximity to each other. Many of the methods are *effective* only if the data produces spherical, distinct, evenly distributed clusters. Empirically, data rarely conform to this ideal condition, being sometimes noisy, contaminated with outliers and unevenly distributed. A large data set is likely to have areas of great distinction, as well as intense overlap, where the within-group similarity of one cluster may be very different to that of another. Thus, the number of clusters may vary greatly depending upon the scale the data is looked at, with no definite number being easily extracted.

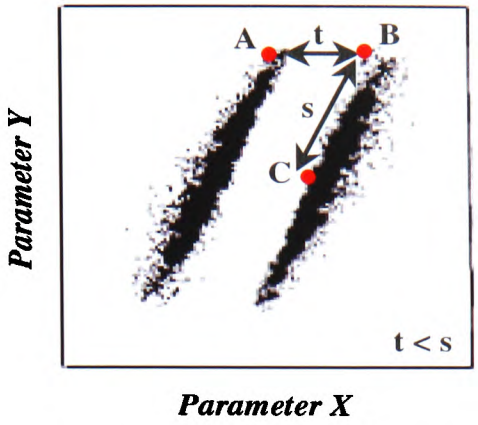


Figure 5.1 Two-dimensional plot of two data sets exhibiting elliptical clusters. Using a classical statistical method, employing a distance metric as a measure of similarity, may regard point B as being *closer* to point A than point C (*i.e.* t<s). Subsequent grouping may partition one or both clusters, giving an inaccurate conclusion of class membership.

5.3 Flow Cytometric Classification and Neural Networks

The existing taxonomies for organisms such as phytoplankton, have been largely developed through phenetic relationships. Some documented work relating to neural network analysis of morphometric data, has produced relatively good results. For example, Boddy *et al.* (1994b) identified fungal spores using morphometric measurements, such as length of basal appendage and maximum spore length, as input to both supervised and unsupervised networks. However, organisms that can be characterised solely on physical measurements, make creating a training file for a neural network a relatively easy option. Once scaled or normalised, the dimensional measurements can be translated into input parameters with little or no conflict. Organisms for which identification is both dimensional and feature based are more difficult. For example, the major distinguishing feature of the Flagellates is the presence and number of flagella. An interpretation of how to represent this physical characteristic would be required, related either to the number present or by implying its' presence (*i.e.* yes = 1, no = 0). In order to ensure the network places equal emphasis on each input parameter, the data must be rescaled (Boddy *et al.*, 2000). Thus, after linear rescaling this feature would add little if any discriminatory data, but may increase computational intensity. The use of flow cytometric signatures removes this problem. However, the characteristic flagella are not detected by flow cytometry, and in cases where this may be the specific morphological variance, the outcome of supervised training could be muddled. With identification, as opposed to classification, performed primarily by supervised neural networks, the presence of non-characteristic class targets makes identification to group, and sometimes genus level, not as successful as to species level. This was established from the results in Chapters 3 and 4 where overlap for some species was not necessarily in-group. Therefore, class labels, more characteristic of flow cytometric signatures for group and maybe genus division, discovered via unsupervised means, can only stand to improve identification by supervised networks.

5.4 Cluster Definition Problem

Cluster definition poses problems in a number of areas, the main two being: (1) determination of the number of clusters present in a data set; (2) to which cluster a data point should be assigned. Whether a hierarchical or partitional method is used, an analyst must pre-determine criteria for cluster definition. In many cases, knowledge of data

distribution, outliers, even the number of clusters is needed to select the best method to use. However, with *a priori* information generally not available, a user must make assumptions to decide upon method and procedure. When this is the case, two investigators using the same data and algorithm may produce very different results, due to a simple assumption made early on in the computations. This degree of user-control makes the analysis very susceptible to errors and inaccuracies.

Unsupervised networks offer an alternative approach to statistical methods which, unlike cluster analysis, are more able to cope with noisy data, outliers and non-uniform cluster distributions. They do not suffer from the cluster definition problem to the same extent as classical methods, not requiring any pre-determined knowledge to select them as an appropriate method. They are less sensitive to their starting conditions and do not need *a priori* information to approximate the data distribution. Without this data being imposed upon them by a user, they are able to determine for themselves underlying similarities. This removes any possible human inaccuracies at early stages and ensures that similar data are located within a close proximity (Section 5.5). However, some of the primary problems of cluster definition are still present. These issues are addressed (Section 5.6).

5.5 Kohonen's Self Organising Map

5.5.1 Architecture

Kohonen's SOM attempts to map events which are close to each other in the p-dimensional input space, to the same or adjacent areas on a two-dimensional (or sometimes one-dimensional or three-dimensional) feature map. The relatively simple architecture generally consists of two layers of nodes (Fig. 5.2). The input layer represents the p-dimensional feature vector, $\mathbf{x} = (x_1,x_2,.....,x_p)$, and is connected to every node in the second or Kohonen layer. The most common structure is a rectangular two-dimensional ordering, where every node within the body of the lattice is interconnected to its eight neighbours (except at the edge of the map).

5.5.2 Algorithm

The weight vectors (of the same dimension as the input data), $\mathbf{w_j} = (w_1,w_2,.....,w_p)$, representing the nodes within the Kohonen layer, are initially set to small random values. As a pattern, $\mathbf{x}$, is presented to the network a best match between it and the Kohonen nodes

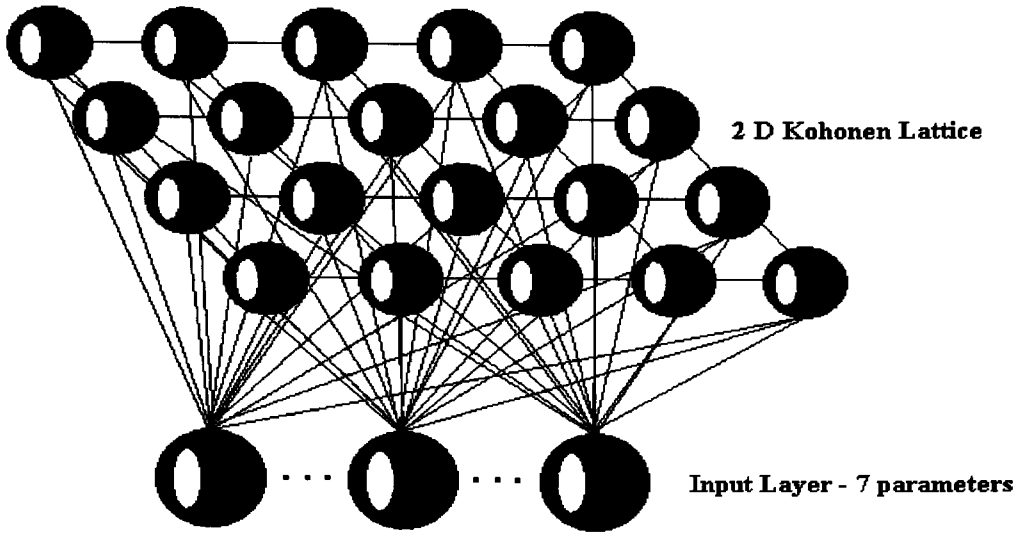


Figure 5.2 Kohonens Self Organising Map in two dimensions. In the rectangular lattice depicted, each internal node is laterally connected to its eight neighbours.

is determined. This is defined as that node, j, whose weight vector produces a minimum Euclidean distance, $D(x,w_j)$, to the pattern,

$$D(x,w_j)=\sqrt{\sum_{i=1}^{p}(x_i-w_{ji})^2}$$

where $j=1,2,\ldots,k$, and k is the user-defined number of nodes.

As the algorithm continues, competitive learning in the Kohonen layer produces a winning node for each pattern, that node is then shifted in the direction of the input vector. Prior to training a region of update is set around the winning node, which can be altered dynamically as training progresses. Nodes that fall within this region are also updated, either positively or negatively, depending on their distance from the winning vector. The excitatory or inhibitory effects of this neighbourhood resembles the *Mexican Hat* function (Fig. 5.3; Kohonen, 1997), where two distinct regions of lateral interaction can be observed. Nodes within region A are moved in the direction of the input vector, those closest to the origin being updated to a greater extent, and those that fall within region B are moved away. The winning node and its localised neighbours are updated according to:

$w_{ji}(t+1) = w_{ji}(t) + \alpha(t) f(d)\left[x_i(t) - w_{ji}(t)\right]$ if j falls within the update region

$w_j(t+1) = w_j(t)$ otherwise

where $0 < \alpha(t) < 1$

The degree of update of the nodes is dependant upon the learning parameter, $\alpha(t)$, which decreases as time progresses, and the neighbourhood function, $f(d)$, which is generally Gaussian:

$$f(d) = \exp\left(-\frac{d^2}{2\sigma^2(t)}\right)$$

Here, $\sigma(t)$ defines the neighbourhood radius and d is the distance between the winning node and the neighbouring units. The function maximum is centred at the winning node, for which $f(d)=1$, and decreases to zero as the distance between the winning node and neighbours increases. As preliminary placement of the nodes is random, there exists the possibility that nodes already representing the input data will continue to win more often than those placed in sparsely populated areas. To counter this problem, a conscience mechanism is initiated (DeSieno, 1988). If a node appears to be winning too often a negative bias is introduced, based on the nodes winning frequency, allowing other nodes

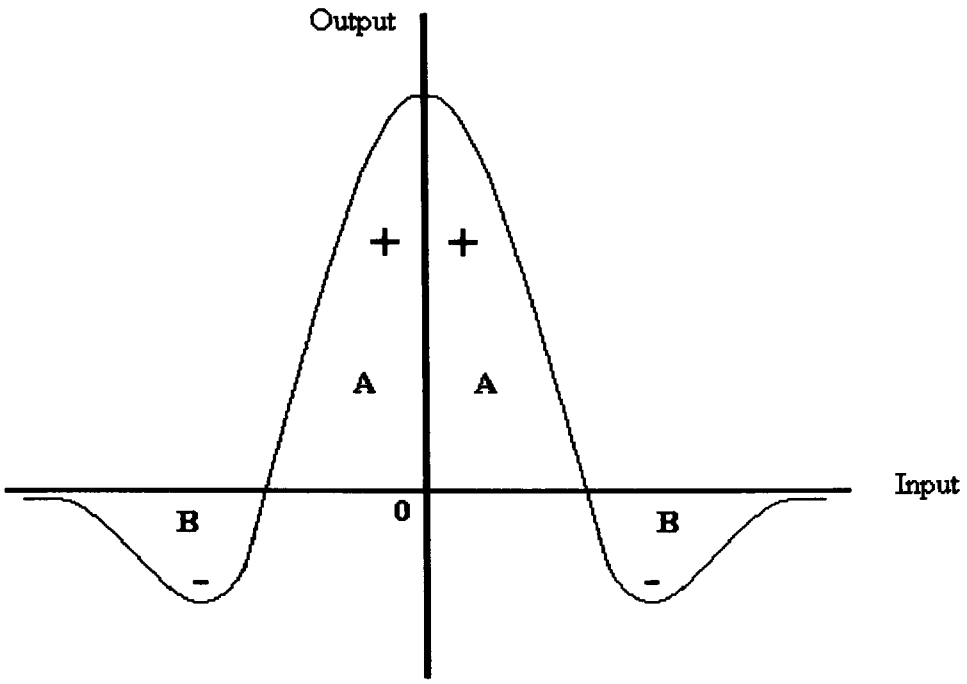


Figure 5.3 Update region within the Kohonen layer depicted by the Mexican hat function. Region A – excitatory, nodes are moved towards the input vector. Region B – inhibitory, nodes are moved away from the input vector.

the chance to be selected and updated (Appendix 2). This iterative process of pattern presentation and update continues until the parameters expire, if linear, or until there is little or no movement of the node positions. The Kohonen nodes then approximate the distribution of the input data, visually displayed in a topologically preserved two-dimensional map.

5.5.3 Initialisation

5.5.3.1 Update Region

The size of the update region is crucial to the final distribution of data in the output space. Too large an update region may not detect small discrete clusters in the input space, whereas too small an area may neglect to find any larger clusters. Kohonen (1990) suggests a compromise of initiating training with a large update region and high learning parameter, and then reducing both until just the winning node and its immediate neighbours benefit from learning (Appendix 2). This establishes the larger clusters and then goes on to detect any fine scale clustering, thereby improving discrimination and better representing the true distribution of the data.

5.5.3.2 Map Size and Dimension

The Kohonen self-organising map, however trained, will always produce a valid clustering outcome, *i.e.* proximity on the map indicates proximity in the input space, though not necessarily vice versa. This is particularly true when the input and output spaces are of the same dimension (Fig. 5.4a). However, if the input space has higher dimensionality than the output space, two points which are close in the input space may not be adjacent in the output space (Fig. 5.4b) (Kohonen, 1990).

The map size and dimension can also result in anomalies, depending on data structure. Too large a grid may separate out many events and possibly produce areas of little or no activity, whereas too small a grid may not partition clusters. The translation of the SOM from p dimensions to two dimensions allows only primary reactions of nodes to be displayed (*i.e.* the class for which a nodes reaction is maximum). Thus, too small a grid may leave some classes under-represented, or, if in a region dominated by one class, a subordinate class may only be apparent from the secondary reaction of a node, not readily

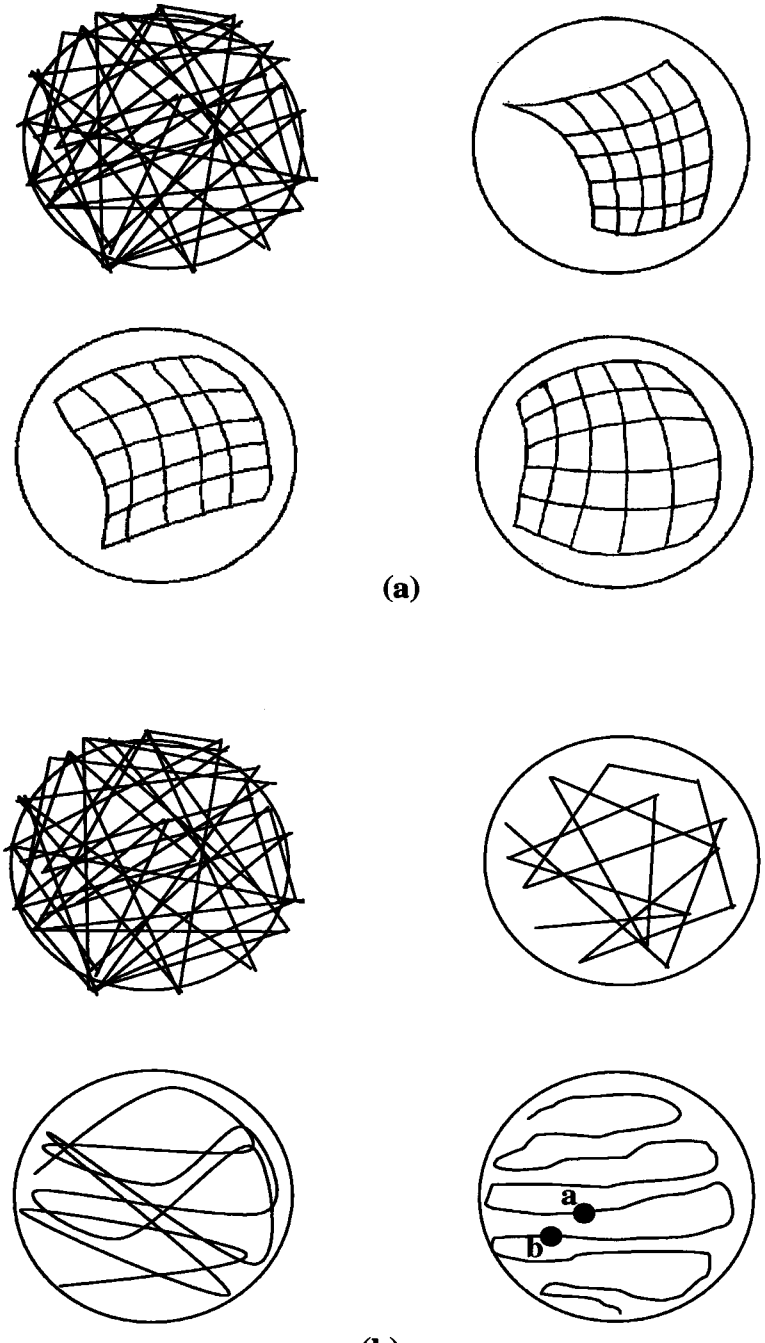


Figure 5.4 (a) Progression of a 6 by 6 Kohonen map, i.e. 36 nodes, learning to map 2 dimensional data arranged in a circular area. At each stage of update, the Kohonen nodes become more representative of the data in the input space mapping similar input patterns to adjacent clusters of nodes (b) Progression of a one-dimensional Kohonen network, with 36 nodes arranged in a line, learning to map the same data set. Despite the one-dimensional map approximating the circular data space, the dimensionality mismatch maps some points, which are close in the input space, to nodes which are a considerable distance from each other in the output space, i.e. **a** and **b**.

obvious as part of a possible cluster.

Analysis is thus required to detect these minor ambiguities, and determine which areas of the map hold similarities despite their positions.

5.6 Boundary and Cluster Detection on the Kohonen SOM

The nature of the SOM allows multiple operators, using the same data, to produce similar, but not identical results, regarding topology preservation. The algorithm alleviates some of the difficulties mentioned in the cluster definition problem (Section 5.4) by minimising the amount of user-defined criteria required. Once the network is trained, the topologically preserved two-dimensional map provides a visual display of the distribution of the data in the output space. Probable cluster determination is achieved via boundary recognition within this space. This is not an easy process, and an area in which research is diverse.

The Unified Matrix Method (UMM) (Ultsch & Siemon, 1989) is one approach which constructs a three-dimensional landscape from a distance matrix, allowing visualisation of the topology of the Kohonen feature map. The weights of the nodes in the Kohonen layer are analysed and the distance between them depicted as heights, with areas of considered closeness shown as valleys and greater distances as hills. Similarly the distance matrix can be represented in grey scale, with greater distances represented by the lower end of the spectrum (Kraaijveld *et al.,* 1992). It has been used successfully in a number of applications (*e.g.* Iivarinen *et al.,* 1994; Vesanto *et al.,* 1997). A similar approach has been employed for analysis of the phytoplankton data (Section 5.7.1.2).

Murtagh (1994) employed a different approach combining the SOM with contiguity-constrained clustering (CCC). Generally in agglomerative clustering techniques the two closest clusters, x and y, are replaced by their mean, w, and this process is continued until only one cluster remains. The CCC expands this method by clustering the two groups, or single vectors, only if there is some $q \in x$ and some $q' \in y$ such that q and q' are contiguous. Murtagh (1995) used a set of variables, derived from four flux values recorded from the Infrared Astronomical Satellite, comprising readings at a number of wavelengths for approximately 250,000 objects, including nebulae, quasars and stars of different types. The method requires some input from the user as to the number of clusters

or groups desired at the finish. Agglomerative clustering on the SOM is investigated further (Section 5.7.5).

Grid growing (Fritzke, 1991; Blackmore & Miikkulainen, 1993) adapts the basic two-dimensional grid produced by an SOM into a flexible map, where nodes can be added or subtracted, depending upon connect and disconnect thresholds. The threshold values are user defined and dependent upon an average measurement of closeness, introducing the possibility of loss of information due to the nature of mean value calculations. Other methods such as deterministic annealing, require *a priori* information as to the number of clusters present and selection of cluster centres.

A number of methods for boundary and cluster detection on the SOM have been investigated. The methodology and results of these approaches are discussed and compared, in relation to the difficulties with classical phytoplankton taxonomy when considering flow cytometric signatures.

5.7 Boundary Detection Methods

5.7.1 Visualisation of Hyper-Dimensional Euclidean Distances

5.7.1.1 Borders between nodes

Once training is complete the Euclidean distance between nodes of the SOM can be visualised as borders of varying thickness. Using purpose written software (Wilkins *et al.*, 1994b), Euclidean distances can be shown either between nodes representing a particular class (if known), or between every node on the grid. In the case of a labelled data set, this allows not only the similarity between classes to be visualised, but also the within-class variation of flow cytometric signatures. Additionally, a threshold, t, can be imposed upon the Euclidean distances between nodes, allowing only those borders representing a distance greater than t to be plotted, thereby allowing the user control over the cluster membership and hence the number of clusters.

5.7.1.2 Grey Scale representation of Euclideans

Once the network has approximated the data distribution, the Euclidean distances between the position vectors of the Kohonen nodes can be arranged in a distance matrix. The average Euclidean distance between a non-edge node and its 8 surrounding neighbours can be calculated (Kraaijveld *et al.*, 1992). Areas where the Euclidean average is high, are

shown as grey scale values at the darker end of the range, and those of lower values at the lighter end. Because of the SOM's competitive learning, those nodes representative of the inter-object similarity within a cluster are moved towards that cluster, and those not representative pushed away, perhaps towards their own clusters. This partitioning creates a region of high Euclidean distances in an area that offers the possibility of representing a boundary. However, the problem with this method is that of information loss through an average calculation, where for example, high Euclidean distances on one side of a node may be lost if the values on the opposite side are low. This subsequently produces a grey scale value that is not actually representative of the particular area.

To counter this, a second grey scale image is produced here that illustrates the four Euclidean distances between a node and its four corner neighbours. This second image shows four grey scale blocks for each node, thus preserving the information around the central kernel. Both methods are employed using an exaggerated scale to emphasise possible boundaries. Interpretation of the grey scale as a boundary is left to the user.

5.7.1.3 Edge Detection

Differentiation of the grey scale images can be furthered by the employment of image processing techniques. The different areas of grey scale representing high Euclidean distances and average Euclidean distances (Section 5.7.1.2), can be considered as borders between homogenous image regions. These possible borders can be enhanced by the application of an edge detection algorithm. The Sobel edge detection algorithm (Pitas, 1993), the mathematics of which will not be discussed here, has been utilised for this purpose.

5.7.2 Redundant Nodes

Natural clusters are described as continuous regions appearing in a z-dimensional space, where each of the z variables represent the axes of a z co-ordinate system, defined in the space by the range of values for each of the parameters representative of the input data. These probable clusters are considered to hold a high density of data points, separated from other clusters by regions of low density (Dillon and Goldstein, 1984). With the employment of competitive learning and the gradual reduction of the update region, nodes within the SOM are moved away from sparsely populated regions and closer to denser

areas. However, if the nodes fall within an area between two probable clusters, where the density of points is extremely low, it is possible that some nodes will not be assigned to either of the groups, producing redundant nodes which are not representative of any data. Using a large Kohonen map allows a greater number of nodes to be available for modelling the training data. This will leave more low density areas on the map, forcing some of the nodes into these relatively empty zones. Using the purpose-written software (Wilkins *et al.*, 1994b), these regions of redundant nodes can be visualised on a two-dimensional grid, showing possible areas for the more distinctive cluster boundaries.

5.7.3 Proportional Node Responses

Although competitive learning allows all nodes to model the data, there are still those that respond more frequently than others. These nodes are naturally found in areas of high density, where similar patterns are repeatedly mapped to the same area or within its proximity. As the update region decreases during training, the node central to the region is the last and most frequently updated, consequently resulting in a higher proportional node response than its neighbours. By imposing a threshold upon proportional node response, a number of these nodes can be considered as possible cluster centres. From the number produced it is then possible to assess whether certain nodes selected may, in fact, be representative of the same cluster, and therefore only one of which should be chosen as the probable centre. Clusters can then be built up around the nodes by considering their proximity to the cluster centre through Euclidean distances.

5.7.4 Visual Population Density

Although competitive learning gives distant nodes a chance to model the data, it does not ensure that every class within the data set is represented by one of the most frequently updated nodes. As node adjustment would be towards the more tightly packed regions, where greater mappings of patterns occur, the majority of nodes above the chosen threshold will inevitably exist in densely populated areas. For this reason, it is more beneficial to select a threshold value giving a greater number of nodes than one per expected number of clusters. This initiation provides an idea of cluster density and can be coupled with a visual selection process to give a variation to the proportional node response method.

Once training is complete, the two-dimensional map approximating the distribution of the input data, will provide a visual representation of region densities. It is then possible to assess whether certain high responding nodes, may in fact be representative of the same cluster, and therefore only one of which should be chosen as the probable centre. Nodes can be visually selected at or near the centre of the regions, and clusters built up around them as before (Section 5.7.3). There may of course exist areas of lower density, with no high responding nodes located within their proximity but empty regions surrounding them, indicating a possible cluster. Centres can be visually selected within these areas, thus attempting to minimise the number of nodes chosen from the same cluster to represent different classes.

5.7.5 Agglomerative Clustering

As already discussed (Chapter 1, Section 1.6.1), there are a number of statistical methods of hierarchical clustering that use a minimum distance rule. The particular algorithm used here, takes the actual position vectors of the trained Kohonen nodes and uses them to produce a matrix of Euclidean distances. The minimum distance between any two nodes is discovered, and the respective nodes are fused to establish the first cluster. Nodes are then either added to the existing cluster, or another containing a further two nodes is formed. The criteria for membership to a cluster depend on whether the distance between the unclustered node and a node in the cluster, is less than the distance between the two unclustered nodes (*i.e.* single linkage; Krzanowski, 1993). Remaining nodes are either added to an existing cluster or new clusters are formed. This procedure continues until a particular number are discovered or all objects belong to a single group.

5.7.6 Decomposition

Decomposition involves partitioning the data set into a number of individual sub-sets (*e.g.* Raghavan *et al.*, 1991; Yang *et. al.*, 1996). Once the SOM is trained, each node, providing it is not redundant, represents a certain number of patterns. As the network is initiated with a user-defined number of nodes, the data represented by an individual node can be filtered into the same number of sub-sets. The method has been investigated for two reasons. Firstly, if a data set comprises a number of distinct groups, it serves as an initial primary partitioning of the data space, indicating coarse clusters. The resulting N

sub-sets can then be used as N separate input files for a further N SOMs. These additional SOMs can then be decomposed further if required, or used to detect finer clusters using one of the above methods. Secondly, to illustrate empirically, the lack of correlation between the flow cytometric data of some species and their classical taxonomic groupings.

5.7.7 Notes on Boundary Detection Methods

The data sets used for this research are labelled with class membership. However, this has no influence on network learning and are present only to demonstrate both the methods, and the relationships between flow cytometric signatures and morphology of phytoplankton. The boundary detection methods presented, develop clusters of Kohonen nodes from the two-dimensional map and not the data directly. As an individual data point can influence cluster formation, the clustering of the nodes makes the approach less sensitive to anomalies or outliers that may be present in the data. Once boundaries have been finalised, actual data classifications are constructed from those data points represented by a particular cluster of nodes.

It was considered that methods 5.7.3 (Proportional Node Response), 5.7.4 (Visual Population Density) and 5.7.5 (Agglomerative Clustering), whose final groupings are based on Euclidean distances, may incorporate ambiguous groupings due to the obscure placement of redundant nodes. The position vectors of these nodes are located in sparsely populated regions and would experience little if any update. As all nodes are clustered by the methods, those that are redundant will inevitably be grouped into the nearest cluster. This information is extraneous and therefore discarded in the analysis. In the above mentioned methods the redundant nodes are highlighted, and not included in cluster formation.

By similar means, the agglomerative method will attempt to cluster the redundant nodes into an existing, or new group. With some redundant nodes a considerable distance from any cluster and possibly close to each other, they may be merged into their own group, again forming ambiguous clusters. To remedy this sensitivity the agglomerative clustering method is initiated with a high number of groups, and the redundant nodes again discarded.

5.8 Experimental procedure

5.8.1 Training Files

Preparation of the flow cytometric data was as described in Chapter 2 (Section 2.6). To investigate the boundary detection methods, four separate data sets were generated. The initial set consisted of eleven separable species (Table 5.1), selected from the results of Chapter 3 (Section 3.8.2, Table 3.7). The eleven species were chosen based on their high identifications, and were rigorously gated to exclude any possible outliers or noise events, producing tight distinct distributions with 300 events per species (Fig. 5.5). To clarify their high separability they were analysed using a 1 node per class RBF network, employing a Mahalanobis distance metric and no optimisation procedure, producing an overall identification success of 99%. Although the eleven species were separable there were areas of minor overlap, so a fictitious seven-dimensional data set was generated, to produce two completely differentiable classes (Fig. 5.6). Each class contained 300 events.

The third data set comprised thirty species of phytoplankton with 300 events per species (Table 5.2; Fig. 5.7). The species were selected at random from the optimum network in Chapter 3 (Section 3.8.2; Table 3.7), exhibiting a range of identification success, from 99.4% for *Micromonas pusilla*, to 45.6% for *Ochrosphaera neopolitana*.

Finally, the fourth data set comprised sixty of the sixty-two species database, with 300 events per species (Table 5.3; Fig. 5.8). Software restrictions allowed a maximum of sixty classes, therefore, *Gymodinium micrum* and *Gymodinium veneficum* were excluded, as this overlapping genus was already well represented.

5.8.2 Kohonen Network Training

To investigate the performance of the boundary detection methods, independent of map size and initialisation, two Kohonen SOMs were produced for analysis of the generated two class data: an 8 by 8 map, containing 64 nodes, and a 24 by 24 map, containing 576 nodes. For the eleven species data set three Kohonen maps were trained, an 18 by 18, 22 by 22 and a 24 by 24. Square grids were chosen for simplicity, although the methods are not restricted by map configuration. All networks were trained three times to ensure reproducibility.

To investigate the methods for larger data sets four networks were trained, two for the 30 species data set and two for the 60 species data set. A 24 by 24 (576 nodes) and a

Table 5.1 Eleven species data set. Class numbers indicate labelling shown on the Kohonen two dimensional grids. Individual species identification by a RBF network trained on the eleven species data set with 300 events and 1 node per class.

Taxonomic Group	Species	Class Number	Size μm	Individual i.d. (%)
Cryptomonads	*Chroomonas sp.*	1	8-10	99.7
	Cryptomonas appendiculata	2	15-25	98
	Cryptomonas calceiformis	3	10-15	97.3
	Cryptomonas reticulata	4	18-25	99
	Cryptomonas rostrella	5	16-25	99.3
	Hemiselmis virescens	6	5-8	99.7
Flagellates	*Micromonas pusilla*	7	1-3	100
	Porphyridium pupureum	8	4-6	100
	Tetraselmis tetrathele	9	10-16	99.3
Prymnesiomonads	*Emiliania huxleyi B11*	10	5-7	99.3
Diatom	*Phaeodactylum tricornutum*	11	8-35	100

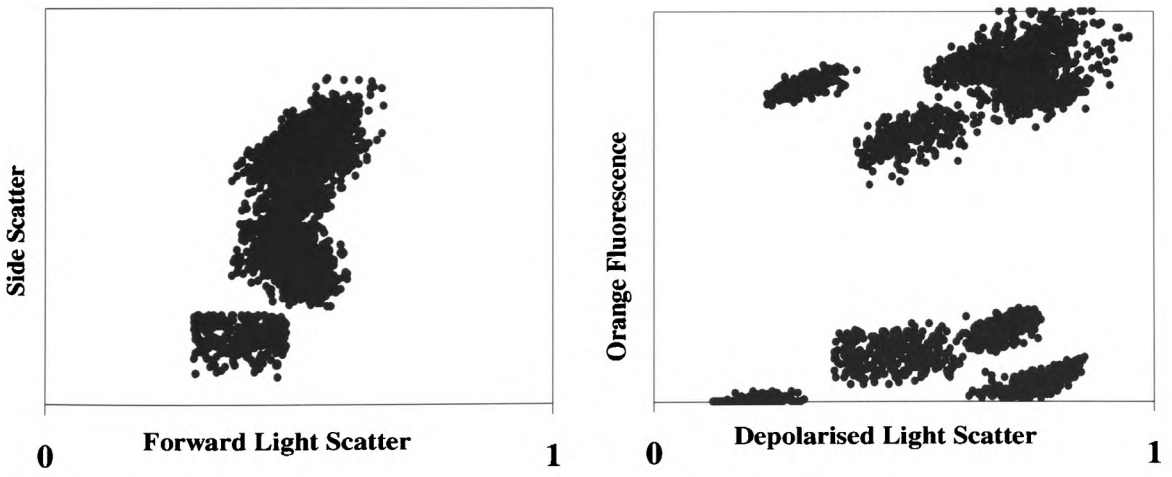


Figure 5.5 Scatter plot depicting two-dimensional views of two sets of optical parameters for the eleven species data set (Table 5.1).

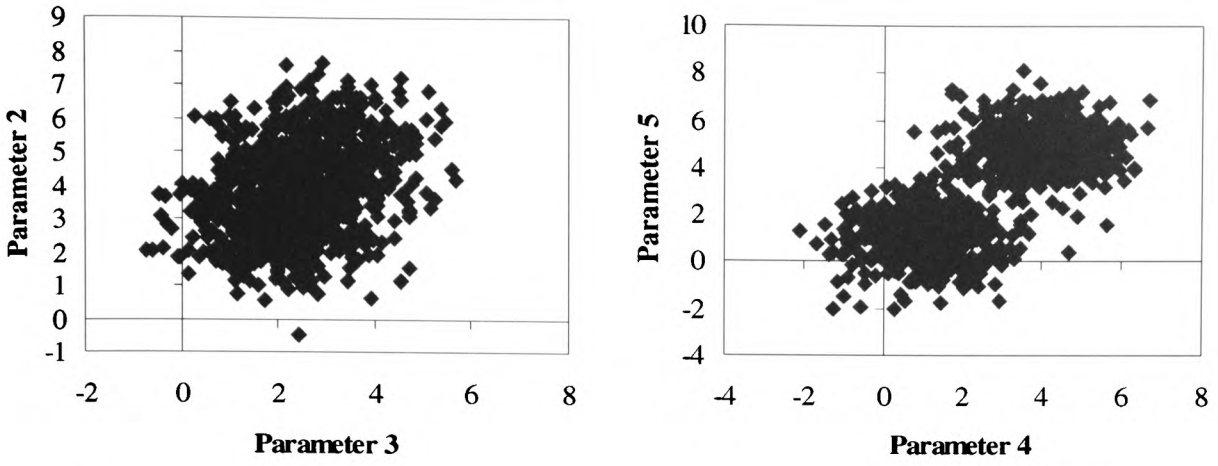


Figure 5.6 Scatter plot showing two-dimensional views of two sets of parameters for the generated two class data set. n.b. Although there appears to be some areas of overlap, it is attributed to the two-dimensional images of the seven-dimensional data.

Table 5.2 Thirty species data set. Class numbers indicate labelling shown on the Kohonen two dimensional grids. Individual identification by the optimum RBF network from Chapter 3 (Overall identification - 77.6%).

Taxonomic Group	Species	Size μm	Class No.	Individual i.d. (%)
Cryptomonads	*Cryptomonas calceiformis*	10-15	1	93.6
	Cryptomonas maculata	12-20	2	90.8
	Hemiselmis brunnescens	5-8	3	65
	Hemiselmis rufescens	4-9	4	64.4
	Hemiselmis virescens	5-8	5	95.8
	Plagioselmis punctata	6-9	6	92
Flagellates	*Chlamydomonas reginae*	11-20	7	91.8
	Dunaliella primolecta	5-12	8	85.2
	Micromonas pusilla	1-3	9	99.4
	Nephroselmis pyriformis	4-7	10	71
	Nephroselmis rotunda	6-8	11	54
	Porphyridium pupureum	4-6	12	95.2
	Tetraselmis tetrathele	10-16	13	94.6
	Tetraselmis verrucosa	3-11	14	60.2
Prymnesiomonads	*Chrysochromulina camella*	6-12	15	85.4
	Chrysochromulina cymbium	6-10	16	46.2
	Chrysochromulina polylepis	6-8	17	60
	Emiliania huxleyi 92	5-6	18	80.8
	Ochrosphaera neopolitana	8-10	19	45.6
	Pleurochrysis carterae	10-18	20	92.2
	Prymnesium parvum	8-10	21	79.8
Diatoms	*Amphora coffaeformis*	10-20	22	88
	Chaetoceros calcitrans	4-6	23	87.6
	Phaeodactylum tricornutum	8-35	24	93.4
	Thalassiosira weissflogii	12-20	25	92.8
Dinoflagellates	*Amphidinium carterae*	15-20	26	77.8
	Aureodinium pigmentosum	7-12	27	88.2
	Gymodinium simplex	6-10	28	69
	Prorocentrum micans	30-40	29	81.2
	Scrippsiella trochoidea	30-42	30	51.2

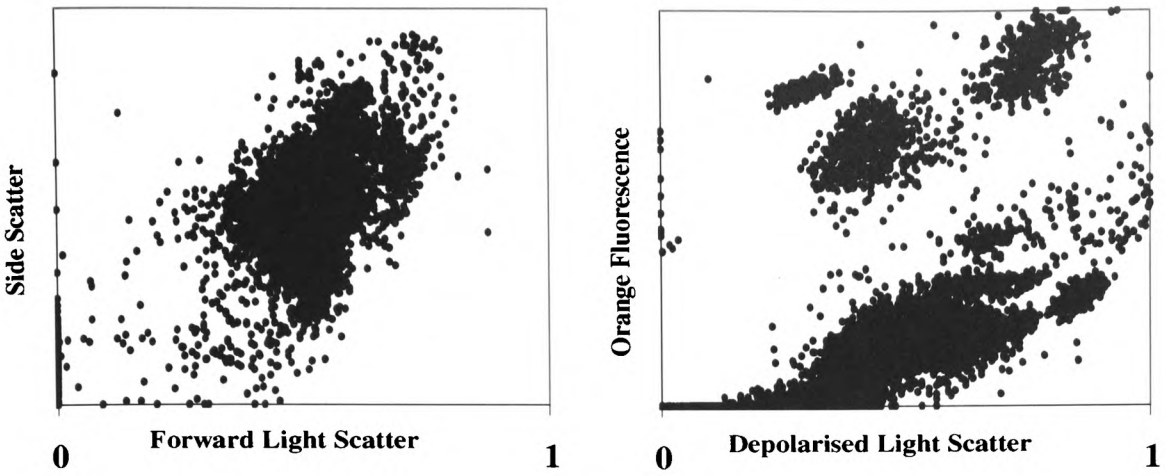


Figure 5.7 Scatter plot showing two-dimensional views of two sets of optical parameters for the thirty species data set (Table 5.2).

Table 5.3 Sixty species data set. Class numbers indicate labelling shown on the Kohonen two-dimensional grids. Individual identification by the optimum RBF network from chapter 3 (Overall identification – 77.6%).

Taxonomic Group	Species Name	Size μm	Class No.	Individual i.d. (%)
Cryptomonads	*Chroomonas sp.*	8-10	1	95.2
	Chroomonas salina	5-12	2	92.4
	Cryptomonas appendiculata	15-25	3	97.6
	Cryptomonas calceiformis	10-15	4	93.6
	Cryptomonas maculata	12-20	5	90.8
	Cryptomonas reticulata	18-25	6	95
	Cryptomonas rostrella	16-25	7	99.4
	Hemiselmis brunnescens	5-8	8	65
	Hemiselmis rufescens	4-9	9	64.4
	Hemiselmis virescens	5-8	10	95.8
	Plagioselmis punctata	6-9	11	92
	Rhodomonas sp.	8-13	12	93.4
Flagellates	*Chlorella salina*	4-8	13	52.2
	Chlamydomonas reginae	11-20	14	91.8
	Dunaliella minuta	3-12	15	67.4
	Dunaliella primolecta	5-12	16	85.2
	Dunaliella tertiolecta	6-12	17	82.4
	Micromonas pusilla	1-3	18	99.4
	Nephroselmis pyriformis	4-7	19	71
	Nephroselmis rotunda	6-8	20	54
	Ochromonas sp.	3-12	21	57.4
	Pseudopedinella sp.	8-10	22	76
	Pelagococcus subviridis	2-3	23	87
	Porphyridium pupureum	4-6	24	95.2
	Pyramimonas grossii	5-10	25	67.4
	Pyramimonas obovata	4-8	26	64
	Rhodella maculata	7-24	27	93
	Stichococcus bacillaris	5-8	28	67.6
	Tetraselmis impellucida	11-19	29	94.8
	Tetraselmis striata	6-8	30	76
	Tetraselmis suecica	6-15	31	87
	Tetraselmis tetrathele	10-16	32	94.6
	Tetraselmis verrucosa	3-11	33	60.2
Prymnesiomonads	*Chrysochromulina camella*	6-12	34	85.4
	Chrysochromulina chiton	5-9	35	61.2
	Chrysochromulina cymbium	6-10	36	46.2
	Chrysochromulina polylepis	6-8	37	60
	Emiliania huxleyi 92	5-6	38	80.8
	Emiliania huxleyi B11	5-7	39	97.6
	Ochrosphaera neopolitana	8-10	40	45.6
	Pavlova lutheri	4-6	41	77.6
	Phaeocystis pouchetii	3-6	42	59.4
	Pleurochrysis carterae	10-18	43	92.2
	Prymnesium parvum	8-10	44	79.8
Diatoms	*Amphora coffaeformis*	10-20	45	88
	Chaetoceros calcitrans	4-6	46	87.6
	Phaeodactylum tricornutum	8-35	47	93.4
	Skeletonema costatum	3-5	48	76.2
	Thalassiosira weissflogii	12-20	49	92.8

Table 5.3 continued

Taxonomic Group	Species Name	Size μm	Class No.	Individual i.d.
Dinoflagellates	*Amphidinium carterae*	15-20	50	77.8
	Aureodinium pigmentosum	7-12	51	88.2
	Gymodinium simplex	6-10	52	69
	Gymnodinium vitiligo	7-22	53	66.4
	Heterocapsa triquetra	15-27	54	72.4
	Prorocentrum balticum	9-15	55	70
	Prorocentrum micans	30-40	56	81.2
	Prorocentrum minimum	16-18	57	61.6
	Prorocentrum nanum	8-10	58	56.4
	Scrippsiella trochoidea	30-42	59	51.2
	Gyrodinium aureolum	35-45	60	86

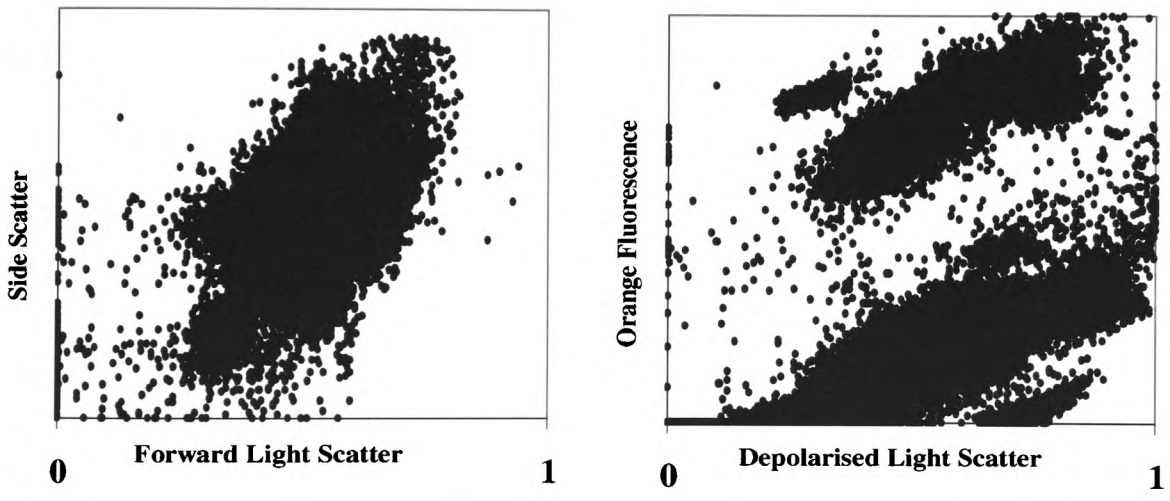


Figure 5.8 Scatter plot showing two-dimensional views of two sets of optical parameters for the sixty species data set (Table 5.3).

30 by 30 (900 nodes) for the 30 species data sets and a 22 by 22 (484 nodes) and a 30 by 30 (900 nodes - the largest the software would allow) for the 60 species data set.

To investigate the decomposition method a 1 by 11 network was trained for the eleven species data set, and a 1 by 5 network for the 60 species data set. The size of the latter network was chosen to establish the lack of definitive separability of the flow cytometric signatures for the 5 taxonomic groups.

All maps were generated using the purpose written software (Wilkins *et al.*, 1994b) which employed Kohonen's recommendations for training an unsupervised network with competitive learning (Appendix 2).

For all boundary detection methods, any parameters that require stating are noted in the results section. The following section presents the results for the particular boundary detection method, followed by the subsequent images relating to it. On figures depicting superimposition of the data sets on the two-dimensional maps (*i.e.* Fig's 5.9, 5.10, 5.14, 5.15, 5.18 5.19, 5.24 & 5.25), positions marked with •, indicate nodes chosen as centres for the visual population density approach only and are not related to other methods.

5.9 Results

5.9.1 Visualisation of Hyper-Dimensional Euclidean Distances

5.9.1.1 Borders between nodes

The plot of the two group data set superimposed on the two-dimensional 8 by 8 map (Fig. 5.9), shows a very obvious separation of 2 distinct clusters. However, the 24 by 24, map of the same data set, does not infer the same results (Fig. 5.10). Without any prior knowledge, there appear to be two relatively sparse areas that could represent possible boundaries, implying three clusters. Figures 5.11 and 5.12 show the Kohonen grids for the two maps. Euclidean distances between allocated nodes are depicted as borders with varying thickness, whether they represent different or same class members. The class membership numbers indicate primary class allocation to a particular node. Once a threshold is imposed upon the Euclidean distances (Table 5.4), the distinction between them is more obvious (Fig. 5.13 & 5.14).

Despite the variation in physical outcome, the topology of the eleven species data was preserved, *i.e.* for the three networks trained the general layout of classes and their

neighbours remained similar, if not always in identical physical areas. From these results the 22 by 22 map was chosen for analysis of the boundary detection methods. The superimposition of data points, displays 11 distinct probable clusters, separated by areas of little or no data points with a few areas of minor overlap (Fig. 5.15). This is supported by the node allocation grid (Fig. 5.16) and the threshold grid (Fig. 5.17), where boundaries below the threshold coincide with areas of slight overlap on the two-dimensional map. For example *Cryptomonas calceiformis* (label 3) and *Cryptomonas reticulata* (label 4).

Clusters on both the 24 by 24 (Fig. 5.18) and 30 by 30 (Fig. 5.19) two-dimensional maps for the 30 species data set are less visually discernible. Although the grids with and without threshold for the 24 by 24 (Fig. 5.20 & 5.21) and 30 by 30 (Fig. 5.22 & 5.23) show some areas of distinction, it appears to represent coarser clustering. However, in comparing the two grid sizes (Fig. 5.20 & 5.22), the consistency of physical placement between similar classes is evident despite map size. A number of probable coarse clusters appear when thresholds are imposed on both maps (Fig. 5.21 & 5.23). For example species *Hemiselmis brunnescens, Hemiselmis rufescens* and *Plagioselmis punctata* (3, 4 and 6 respectively) cluster together on both maps as do *Cryptomonas calceiformis* and *Cryptomonas maculata* (1 and 2), as well as others.

As the class numbers are increased to 60, overlap becomes considerable and real cluster definition is unclear on both the 22 by 22 and 30 by 30 maps (Fig. 5.24, 5.25). Topology preservation is again evident between the map sizes, without and with threshold (Fig. 5.26, 5.27, 5.28 & 5.29). However, as the class number increase some species are not allocated a node's primary reaction and are therefore not present on the map. This is more apparent with the 22 by 22 map, where less nodes are available to represent the data. Threshold imposition depicts only areas of low density when compared to the Kohonen maps (Figs 5.24 & 5.25), but reinforces the coarse allocation between both maps.

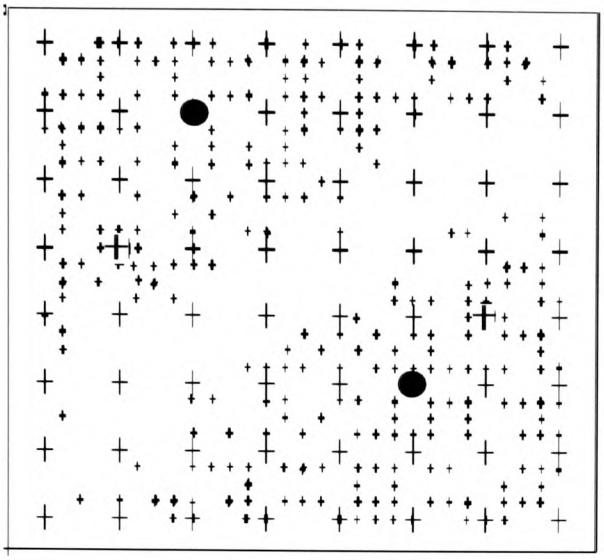


Figure 5.9 8 by 8 Kohonen map produced for the two group data set after training.
+ Dimension of map, + data points superimposed in two-dimensional. An obvious area of low density offers the possibility of 2 distinct clusters (• shows nodes chosen as centres for visual population density method).

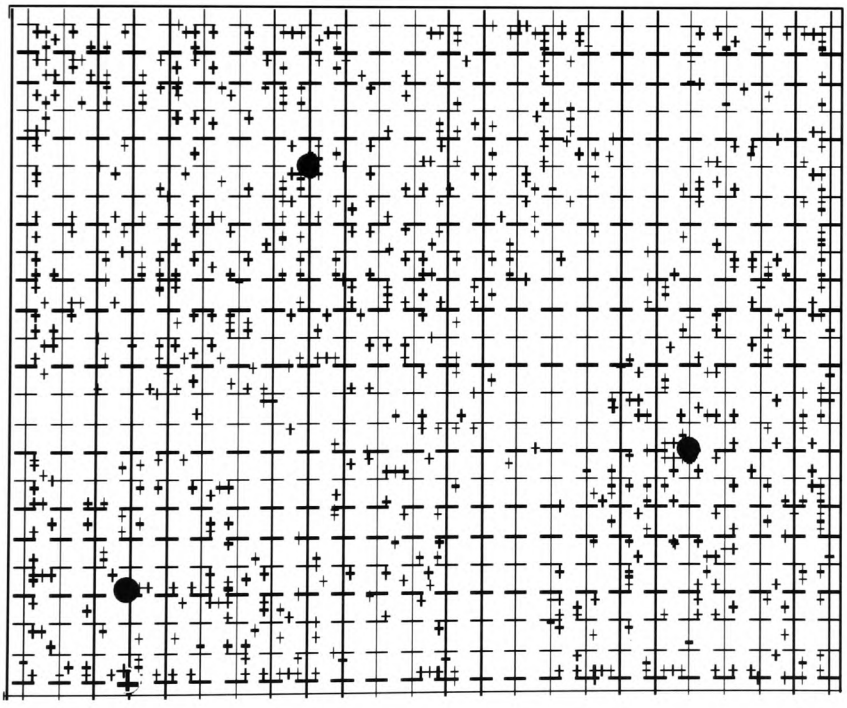


Figure 5.10 24 by 24 Kohonen map produced for the two group data set after training.
+ Dimension of map, + data points superimposed in two-dimensional. Areas of low density imply a presence of three possible clusters. (• shows centres chosen for visual population density method).

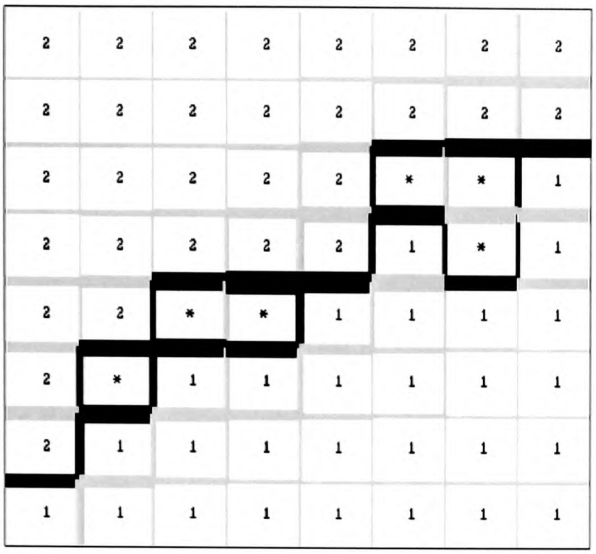


Figure 5.11 8 by 8 Kohonen grid produced for the two class data set.

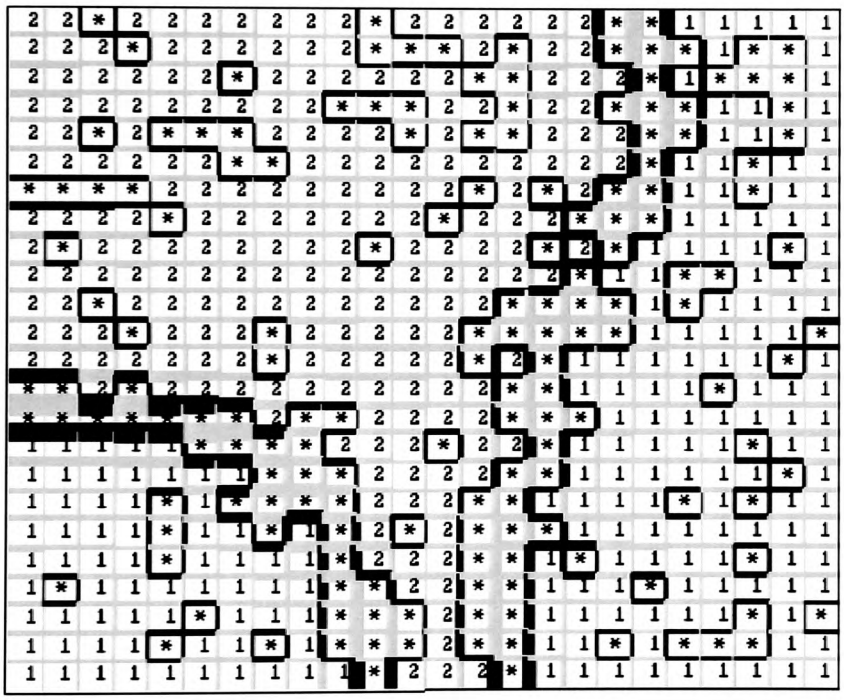


Figure 5.12 24 by 24 Kohonen grid produced for the two class data set.

Note : The numbers depicted on the above two diagrams are for visual purposes only and depict primary allocation, by each node, to a particular class. Euclidean distances between nodes are represented as borders with varying thickness. Black lines are shown between nodes allocated to different classes, and grey between nodes allocated to the same class. Redundant nodes, i.e. nodes that have no events allocated to them are shown as * . This applies to all images depicting Kohonen grids, with and without thresholds.

Table 5.4 Range of Euclidean distances between nodes on each map(s) trained for each particular data set. Threshold values imposed to illustrate the *'borders between nodes'* detection method.

Data Set	Map size	Threshold range	Threshold value
Two Group	8 by 8	0.038667 - 0.216181	0.13
Two Group	24 by 24	0.020960 - 0.245549	0.14
Eleven Species	22 by 22	0.008449 - 0.335903	0.06
Thirty Species	24 by 24	0.008007 - 0.436095	0.06
Thirty Species	30 by 30	0.013200 - 0.510127	0.05
Sixty Species	22 by 22	0.015437 - 0.293154	0.07
Sixty Species	30 by 30	0.008976 - 0.384128	0.06

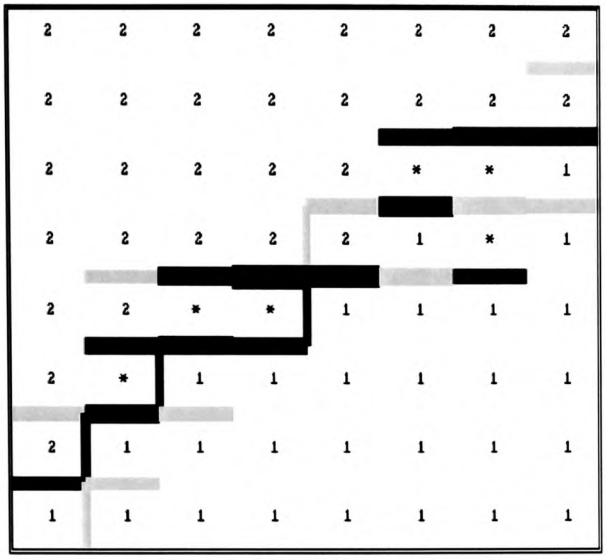


Figure 5.13 8 by 8 Kohonen grid produced for the two class data set with threshold imposed on Euclidean distances between nodes as recorded in Table 5.4.

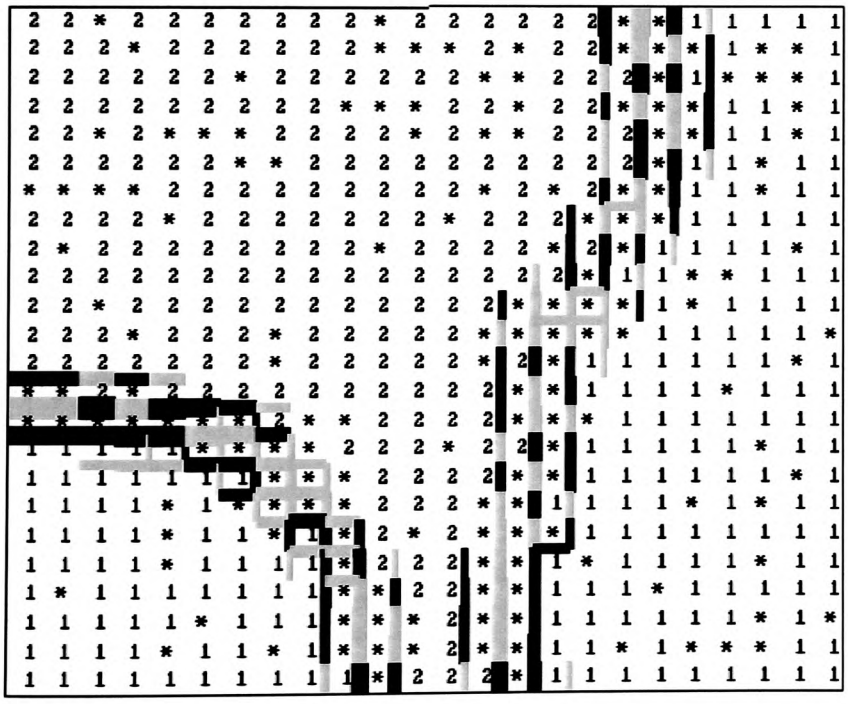


Figure 5.14 24 by 24 Kohonen grid produced for the two class data set with threshold imposed on Euclidean distances between nodes as recorded in Table 5.4

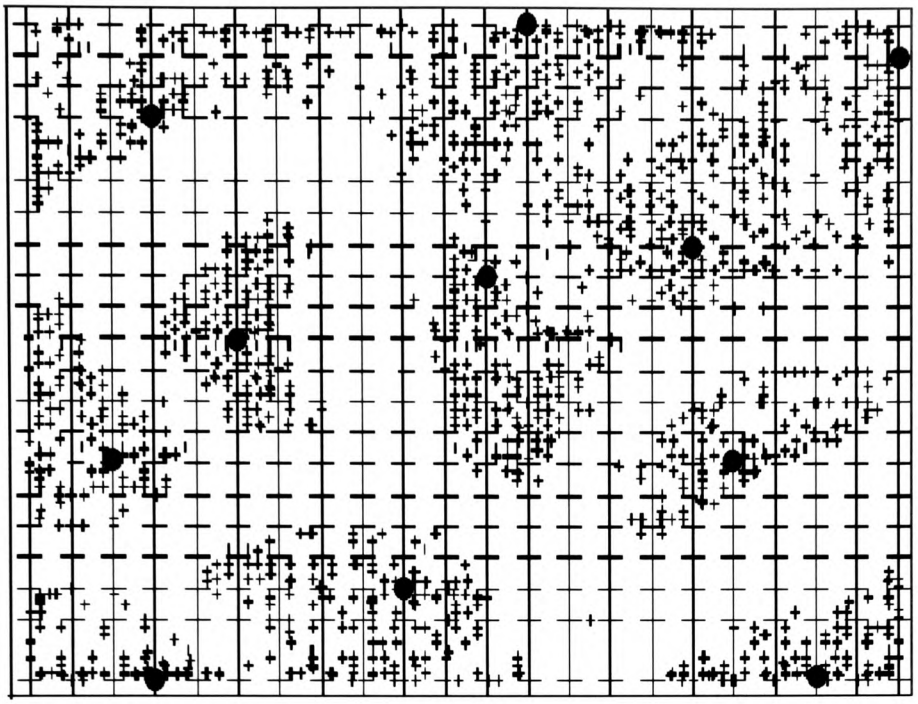


Figure 5.15 22 by 22 Kohonen map produced for the eleven species data set after training. + Dimension of map, + data points superimposed in two-dimensional. High areas of density coupled with empty regions indicates clearly the presence of 11 possible clusters. (• shows centres chosen for visual population density method).

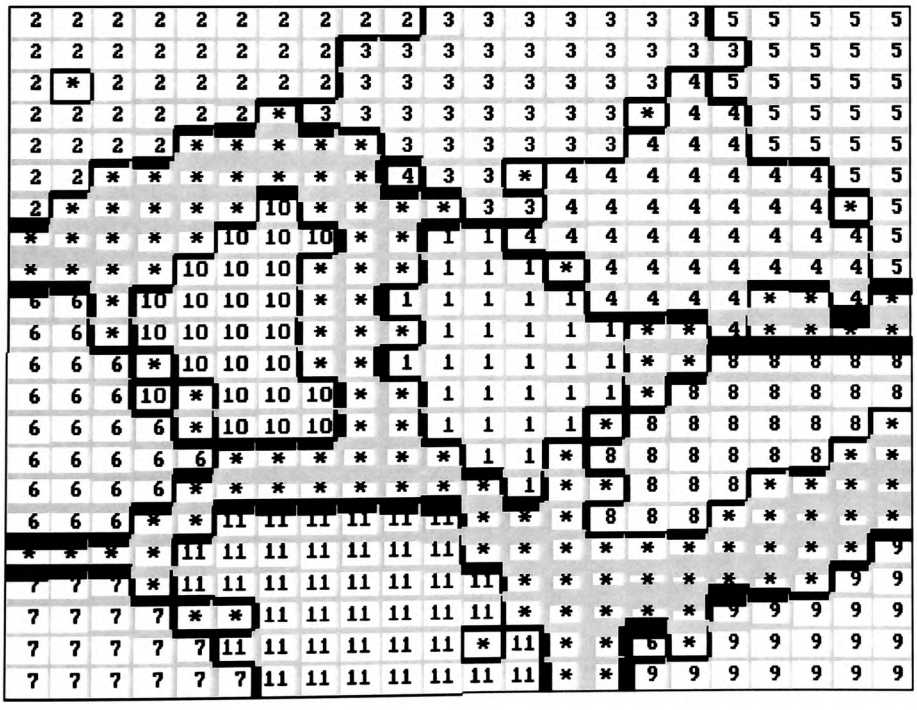


Figure 5.16 22 by 22 Kohonen grid produced for the eleven species data set.

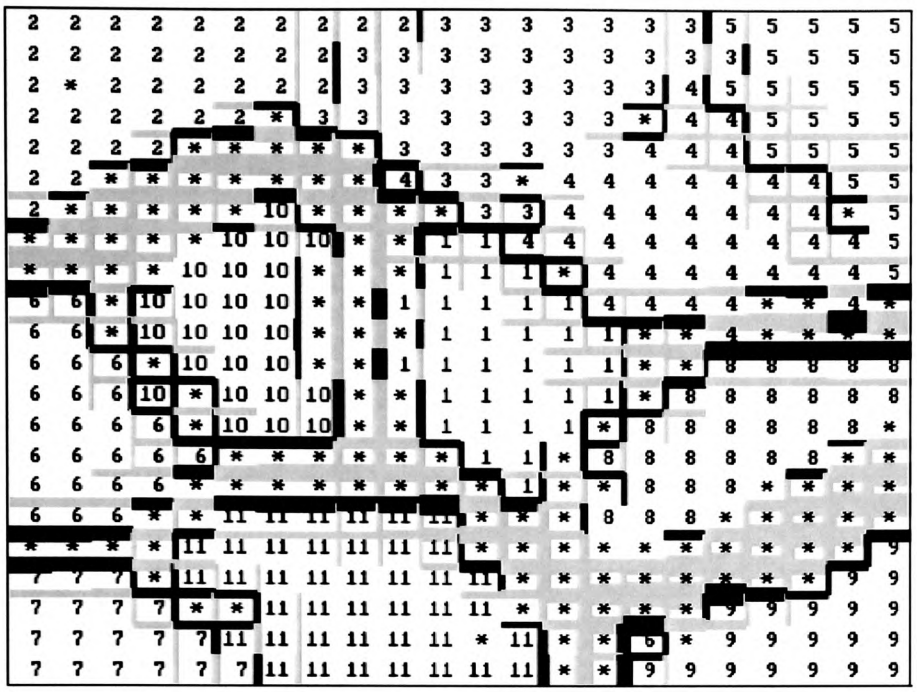


Figure 5.17 22 by 22 Kohonen grid produced for the eleven species data set with threshold imposed on Euclidean distances between nodes as recorded in Table 5.4

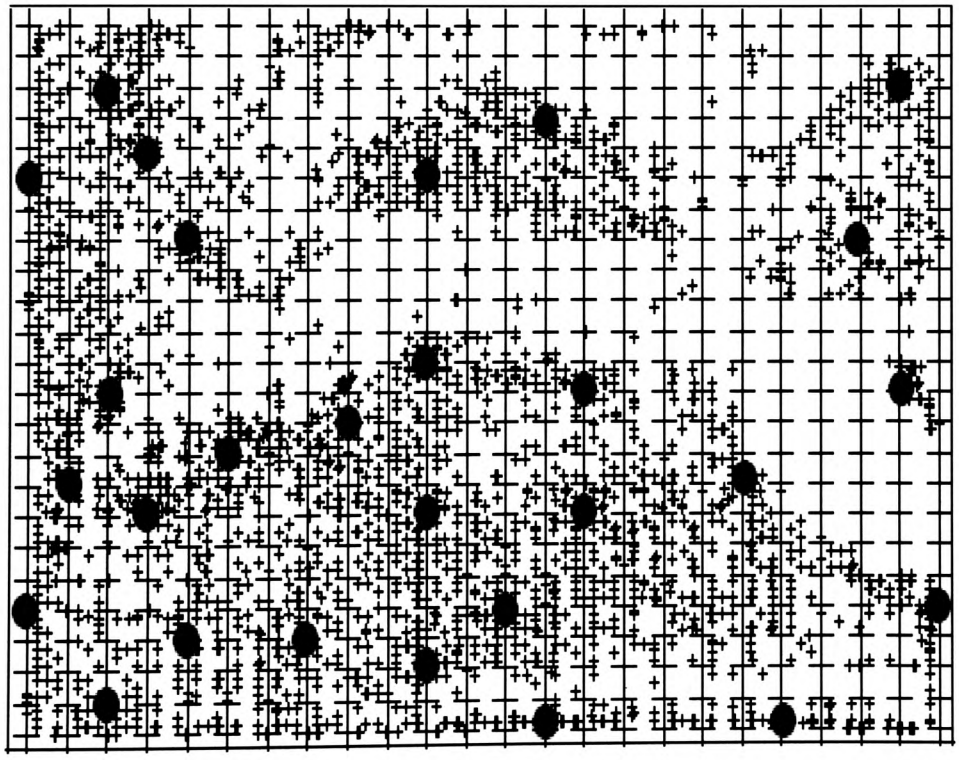


Figure 5.18 24 by 24 Kohonen plot produced for the thirty species data set after training. Some wide unpopulated regions are evident between areas of high density, obvious cluster are less visually discernible. (• shows centres chosen for visual population density method).

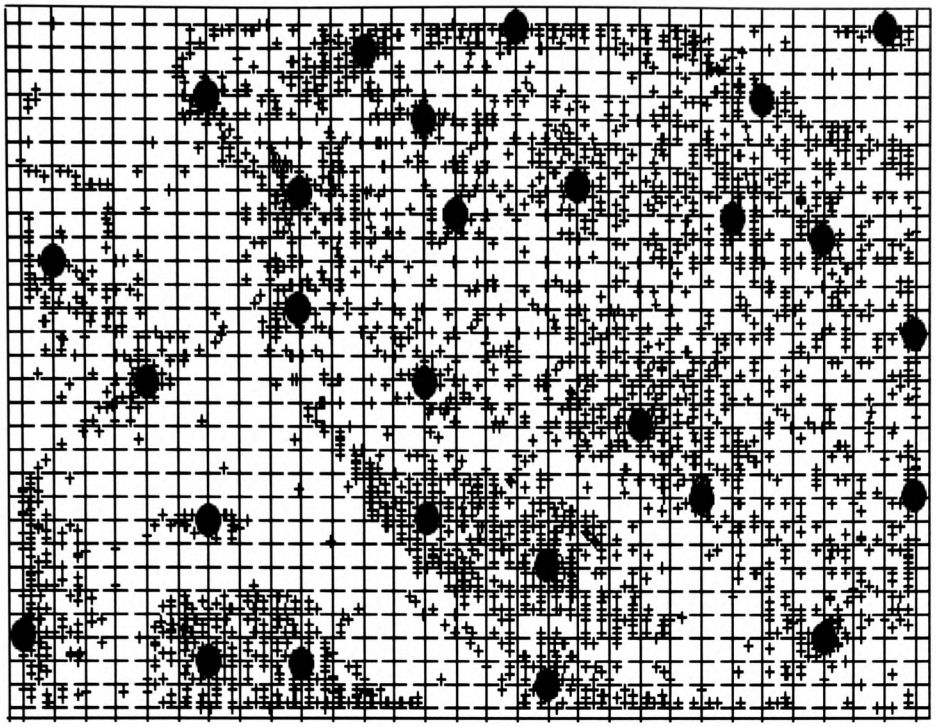


Figure 5.19 30 by 30 Kohonen plot produced for the thirty species data set after training. Greater areas of low density are apparent as the map size increases (• shows centres chosen for visual population density method).

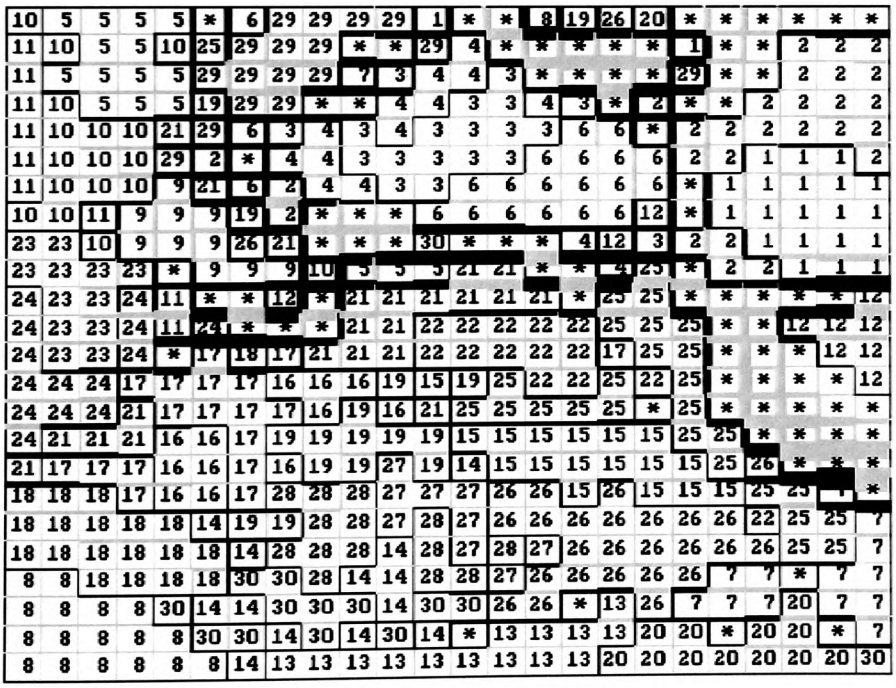


Figure 5.20 24 by 24 Kohonen grid produced for the thirty species data set.

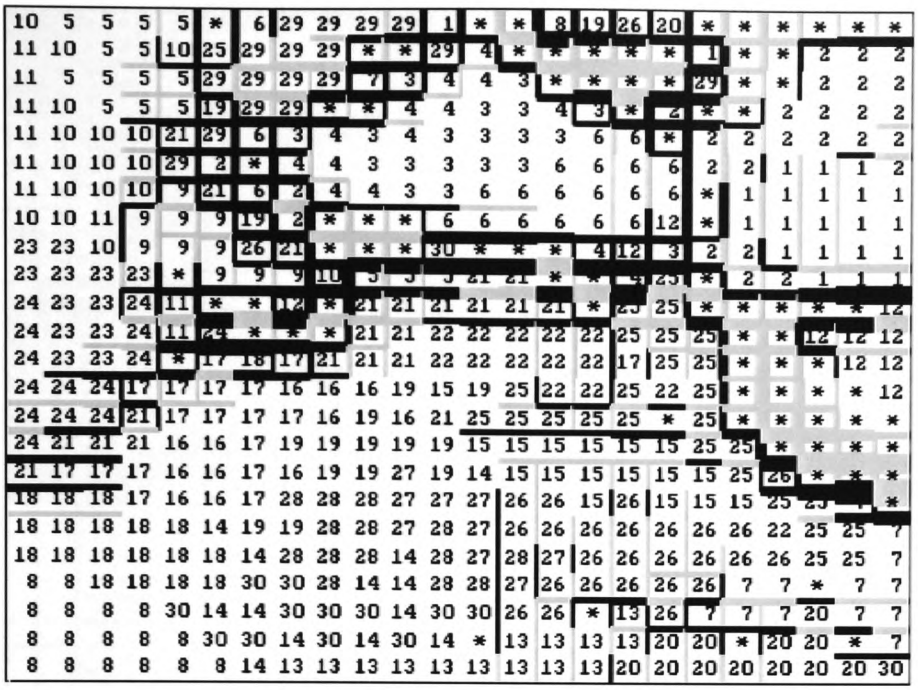


Figure 5.21 24 by 24 Kohonen grid produced for the thirty species data set with threshold imposed on Euclidean distances between nodes as recorded in Table 5.4. Distinct cluster allocation is seen for *Micromonas pusilla* (label 9) and *Porphyridium pupureum* (label 12)

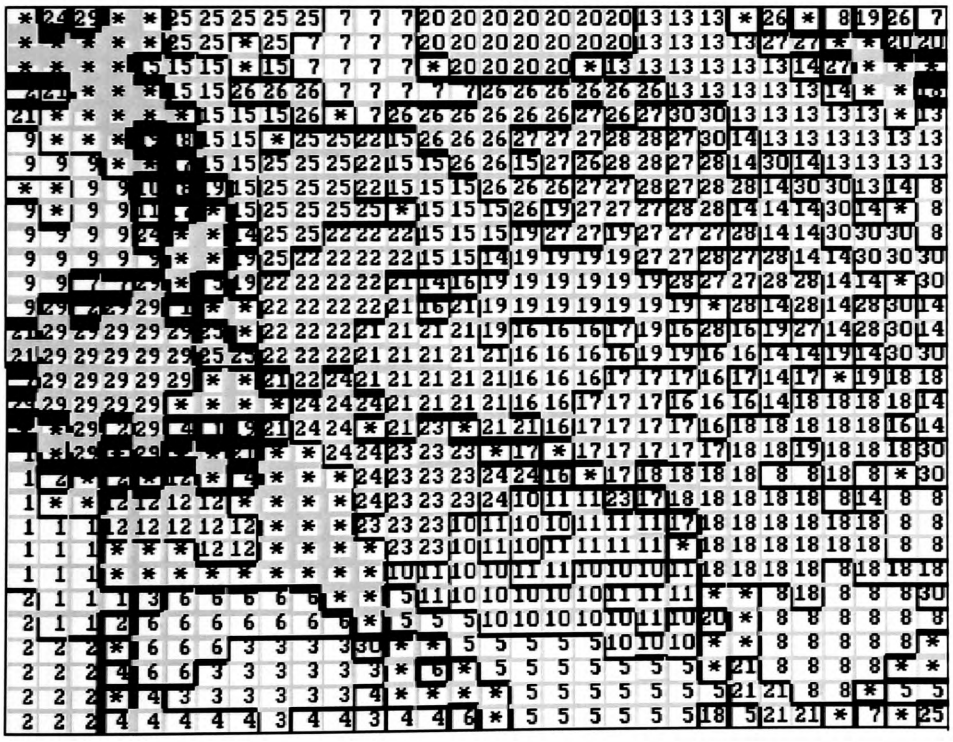


Figure 5.22 30 by 30 Kohonen grid produced for the thirty species data set.

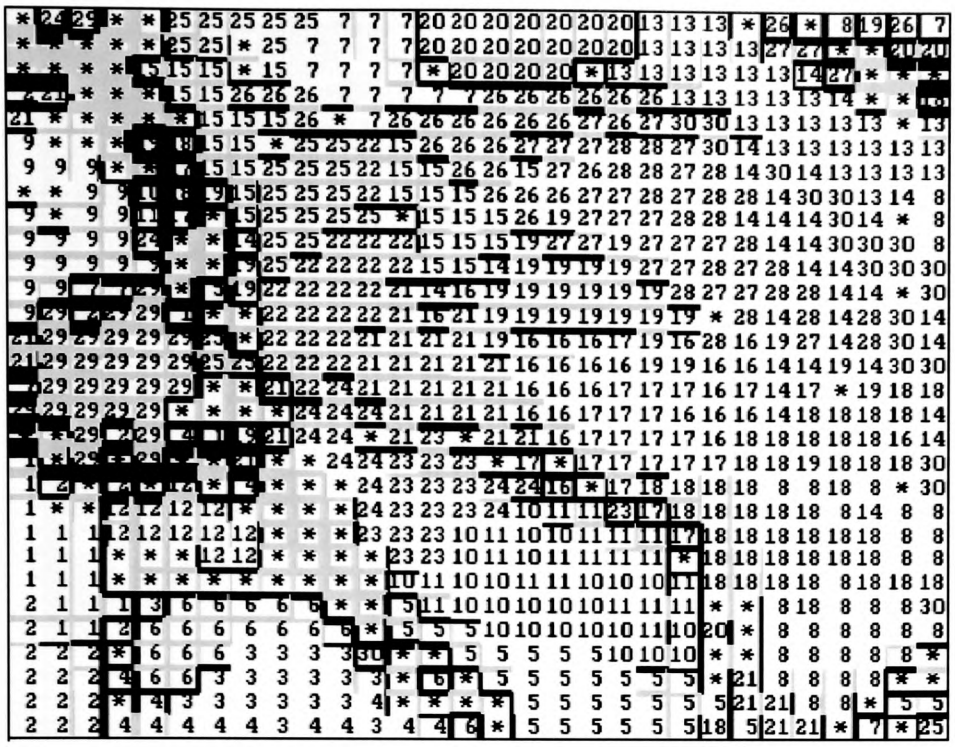


Figure 5.23 30 by 30 grid produced for the thirty species data set with a threshold imposed on Euclidean distances as recorded in Table 5.4. Distinct cluster allocation is seen for *Micromonas pusilla* (label 9) and *Porphyridium pupureum* (label 12).

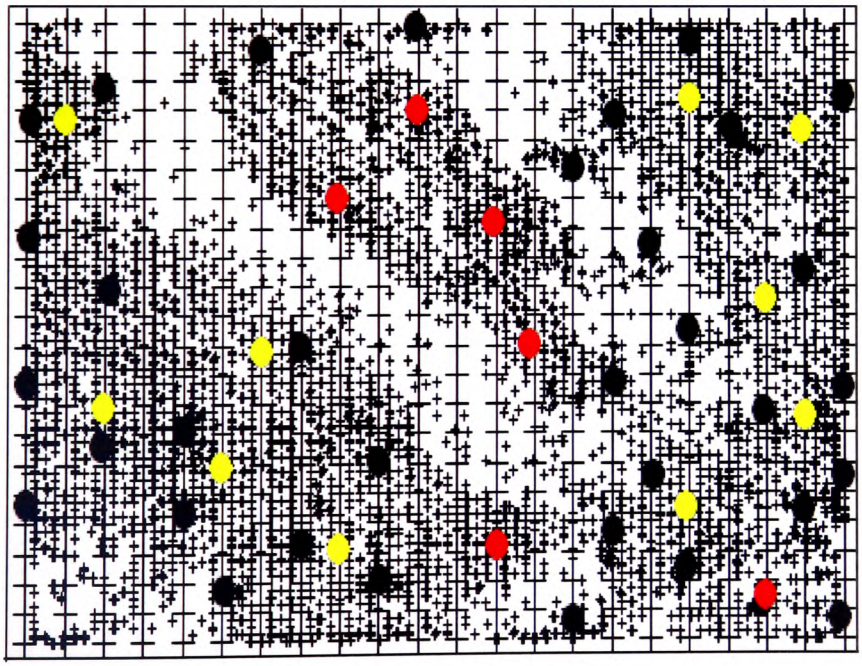


Figure 5.24 Two-dimensional plot showing 22 by 22 Kohonen plot produced for the sixty species data set after training. (• shows centres chosen for visual population density method – black and red points represent the positions of the 40 and red and yellow the 16 centres).

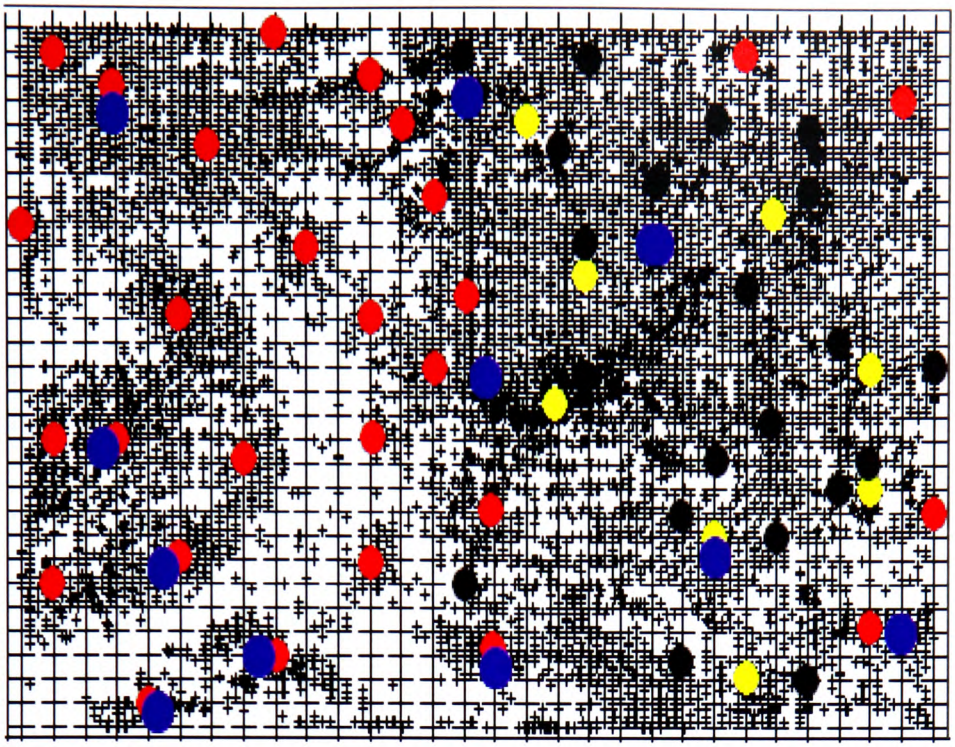


Figure 5.25 Two-dimensional plot showing 30 by 30 Kohonen plot produced for the sixty species data set after training. (• shows centres chosen for visual population density method – black and red points represent the positions of the 49, red and yellow the 36 centres and blue the 11 – where overlap is seen between red and blue the position of both is red).

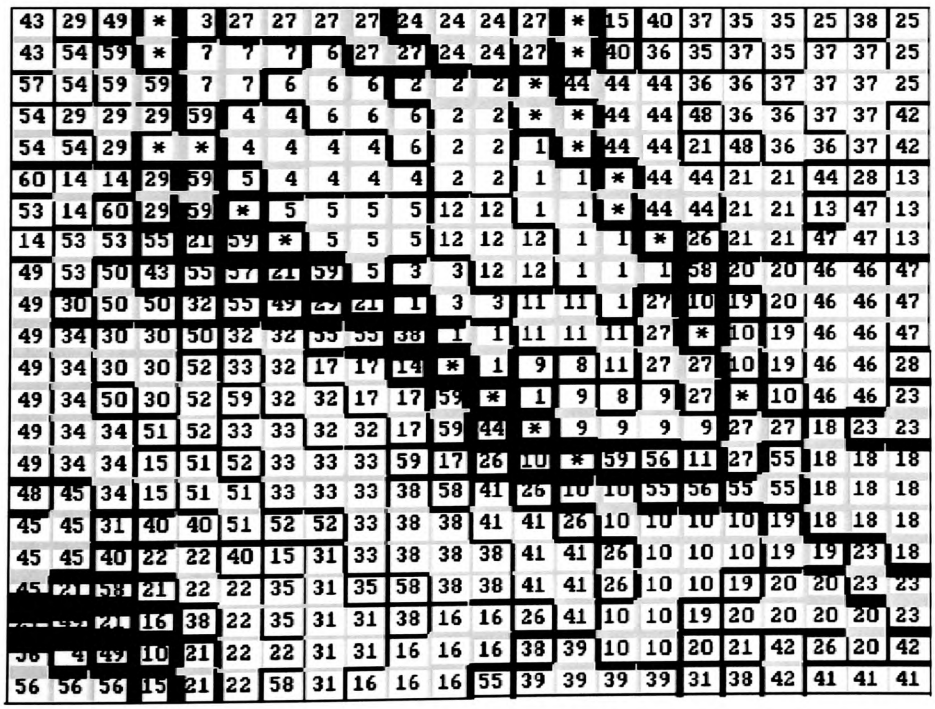


Figure 5.26 22 by 22 Kohonen grid produced for the sixty species data set.

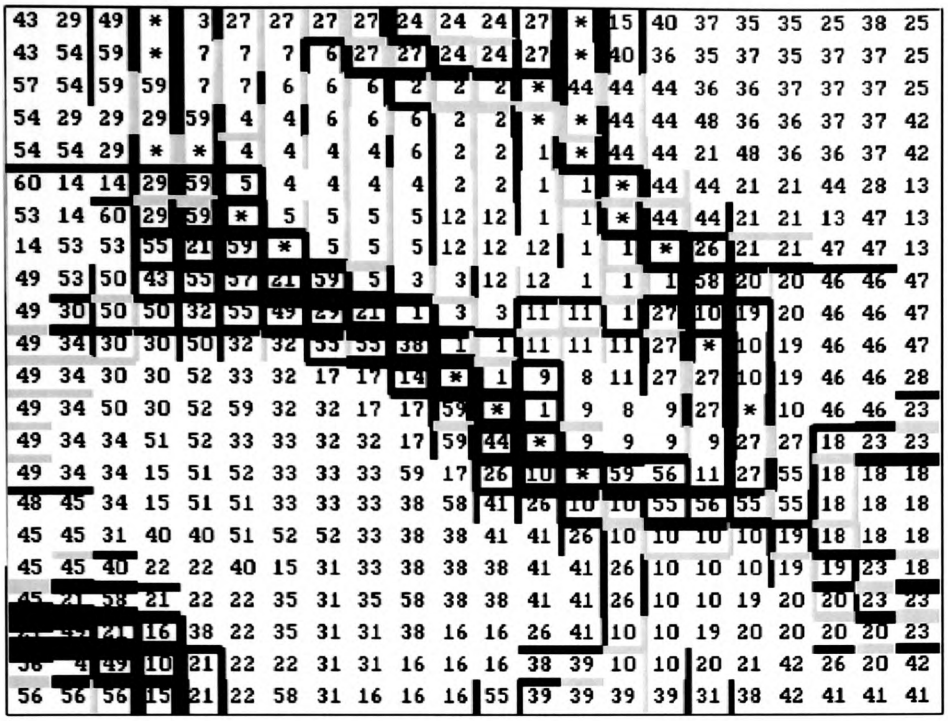


Figure 5.27 22 by 22 grid produced for the sixty species data set with a threshold imposed on Euclidean distances between nodes as recorded in Table 5.4. Distinct cluster allocation is still evident for *Micromonas pusilla* (label 18) and *Porphyridium pupureum* (label 24) with an increased number of classes.

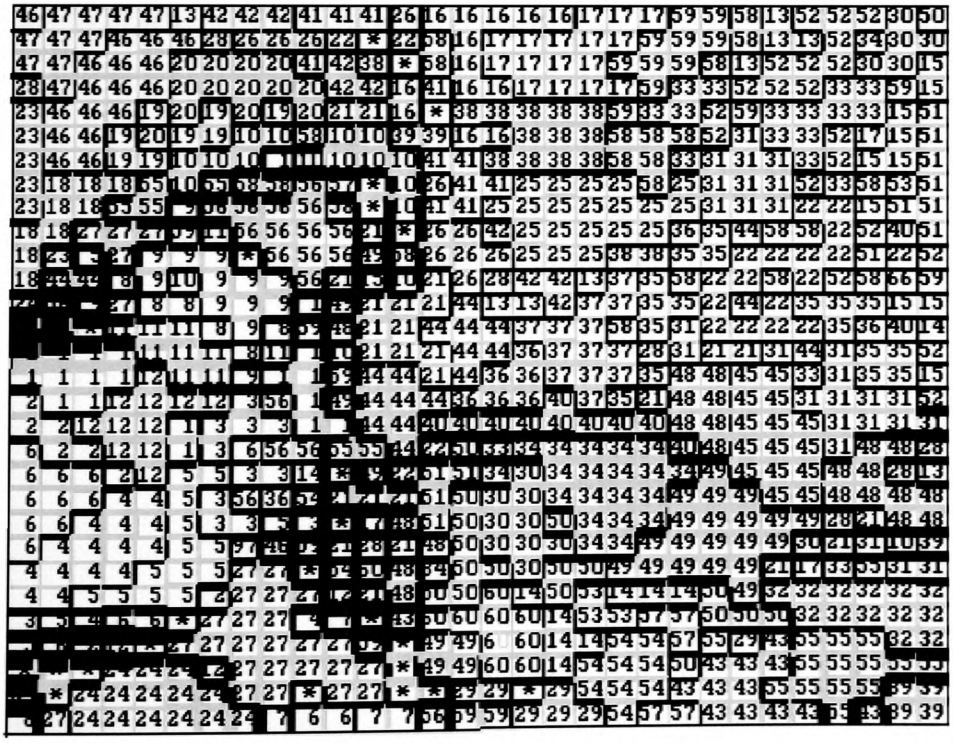


Figure 5.28 30 by 30 Kohonen grid produced for the sixty species data set.

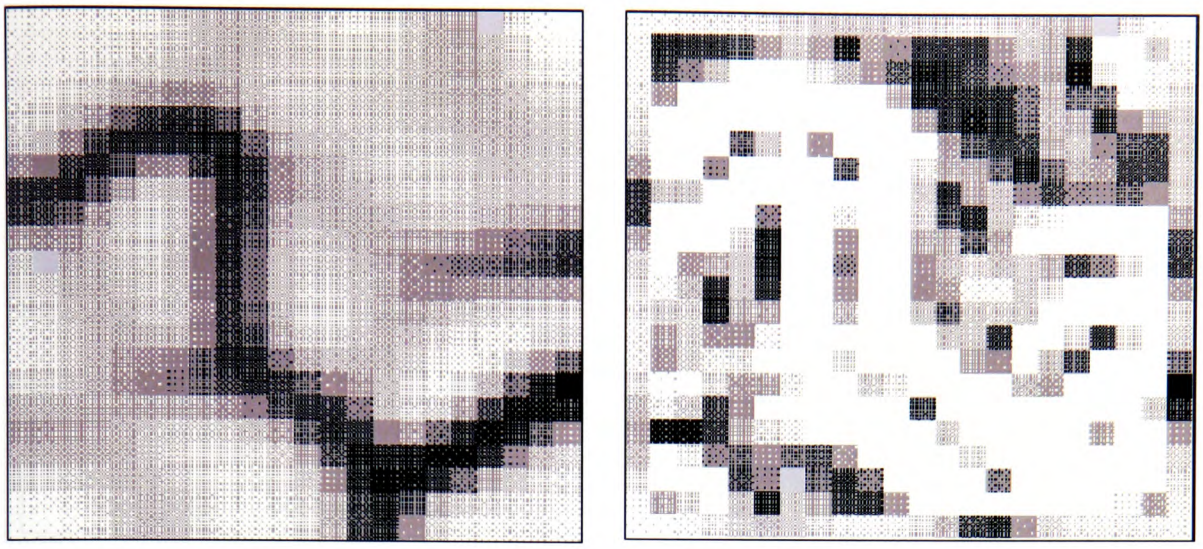
Figure 5.29 30 by 30 grid produced for the sixty species data set with a threshold imposed on the Euclidean distances between nodes as recorded in Table 5.4.

5.9.1.2 Grey Scale Representation of Euclidean Distances.

As the separation of the two class data was so distinct, a grey scale plot of Euclidean distances and Euclidean averages was not generated. Figures 5.30a, 5.32a, 5.34a, 5.36a and 5.38a depict Euclidean distances as grey scale images for all maps representing the 11, 30 and 60 species data sets. The darker end of the scale illustrates areas of higher Euclidean averages. While Figures 5.31a, 5.33a, 5.35a, 5.37a and 5.39a show the images produced using four Euclidean distances to represent each node. The images illustrating four values for each node, show the same areas of high Euclidean distance as that of the images using the average value only, but also exhibit some less pronounced boundaries not evident on the image of averages. When compared to the Kohonen threshold grids for each respective set (Figs. 5.17, 5.21, 5.23, 5.27 & 5.29), areas of similarity are evident with the primary correlation existing in those areas of obvious low density.

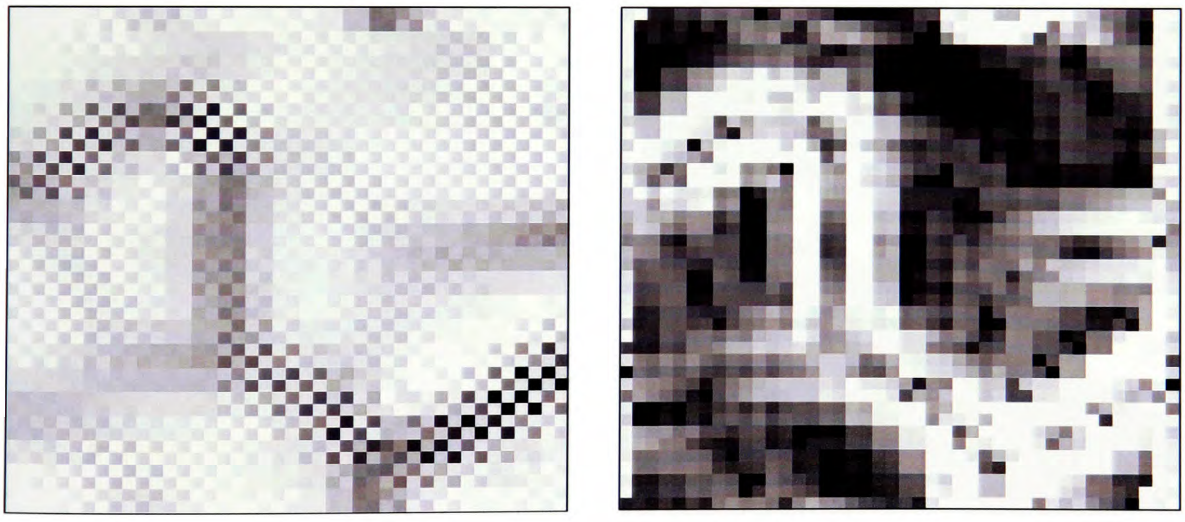
5.9.1.3 Edge Detection

Image b in Figures 5.30 – 5.39 show the affects of applying the Sobel edge detection algorithm to all grey scale images. The lighter end of the scale depicts possible boundaries. The construction of any definitive clusters from these images is, however, very much at the discretion of the reader. The method serves as a verification, when compared to the Kohonen grids, that less populated areas are distanced from populated regions.



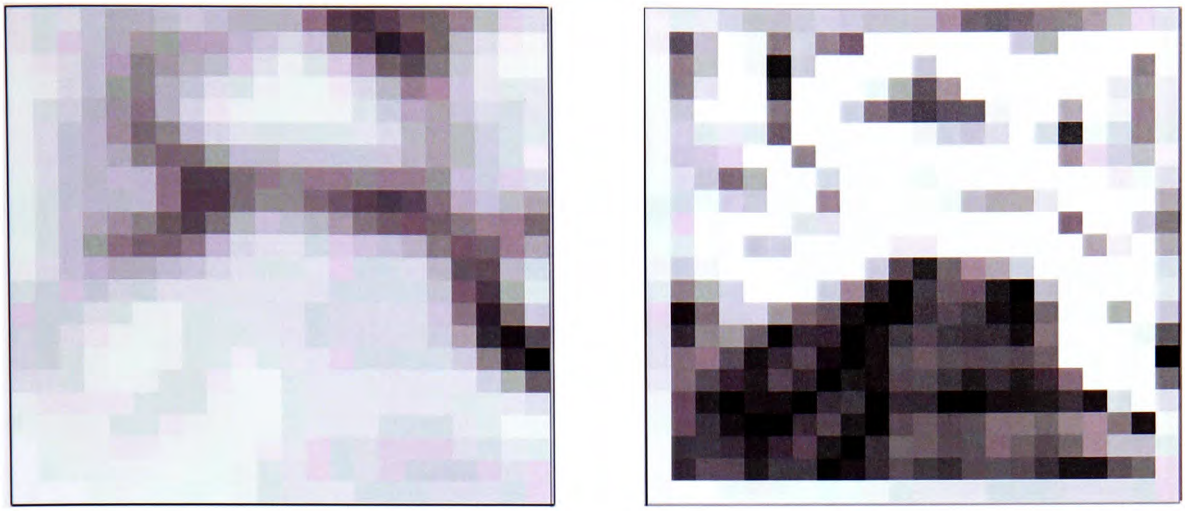
(a) **(b)**

Figure 5.30 (a) Grey scale image of each node represented as the average Euclidean distance between it and its eight neighbours on the 22 by 22 map for the eleven species data set. (b) Image produced after applying the Sobel edge detection algorithm.



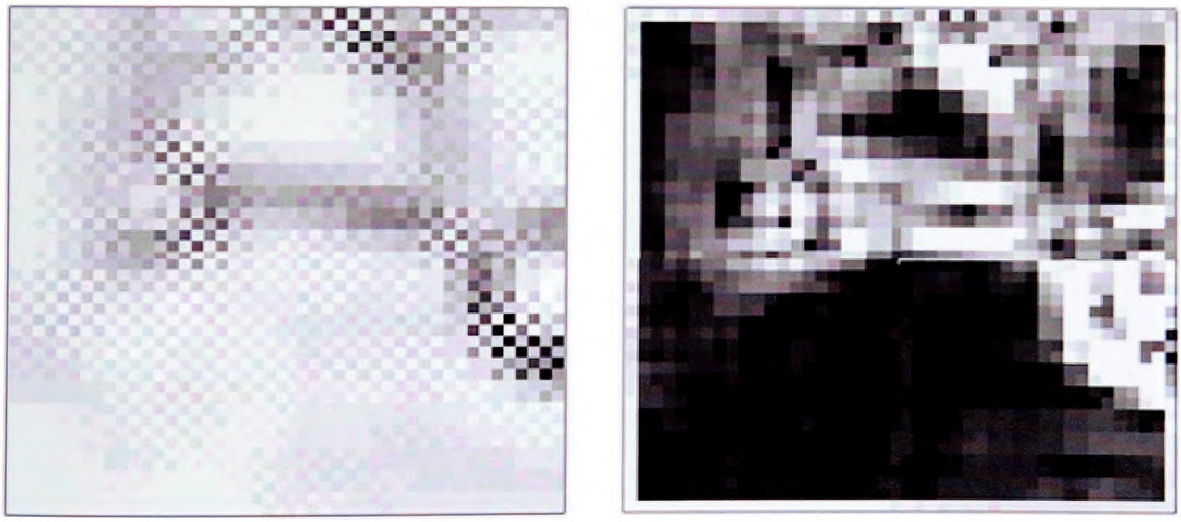
(a) **(b)**

Figure 5.31 (a) Grey scale image of each node as four blocks representing the Euclidean distance between a node and its four corner neighbours on the 22 by 22 map for the eleven species data set. (b) Image produced after applying the Sobel edge detection algorithm.



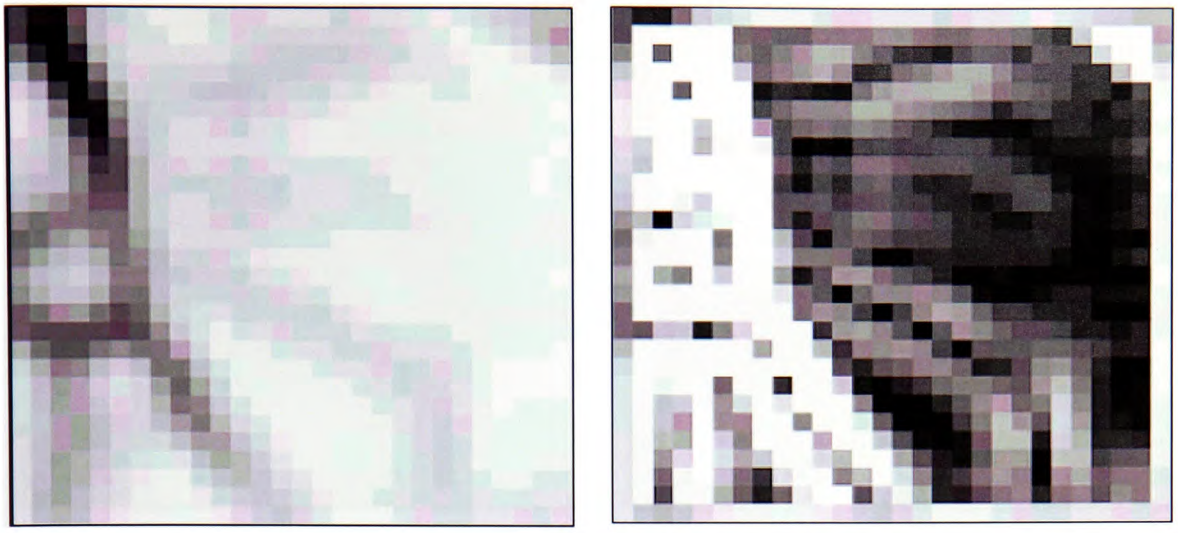
(a) (b)

Figure 5.32 (a) Grey scale image of each node represented as the average Euclidean distance between it and its eight neighbours on the 24 by 24 map for the thirty species data set. (b) Image produced after applying the Sobel edge detection algorithm.



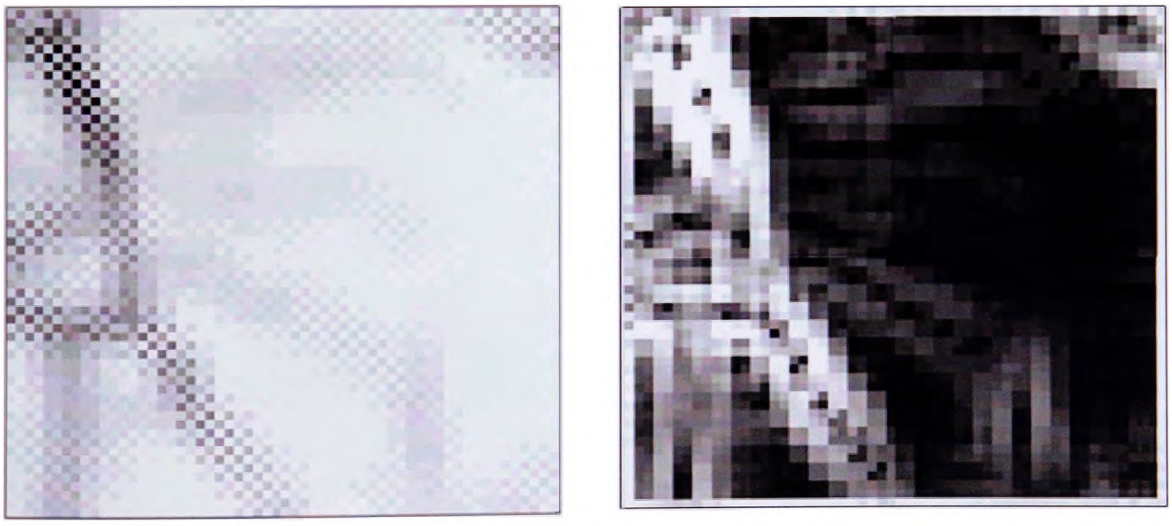
(a) (b)

Figure 5.33 (a) Grey scale image of each node as four blocks representing the Euclidean distance between a node and its four corner neighbours on the 24 by 24 map for the thirty species data set. (b) Image produced after applying the Sobel edge detection algorithm.



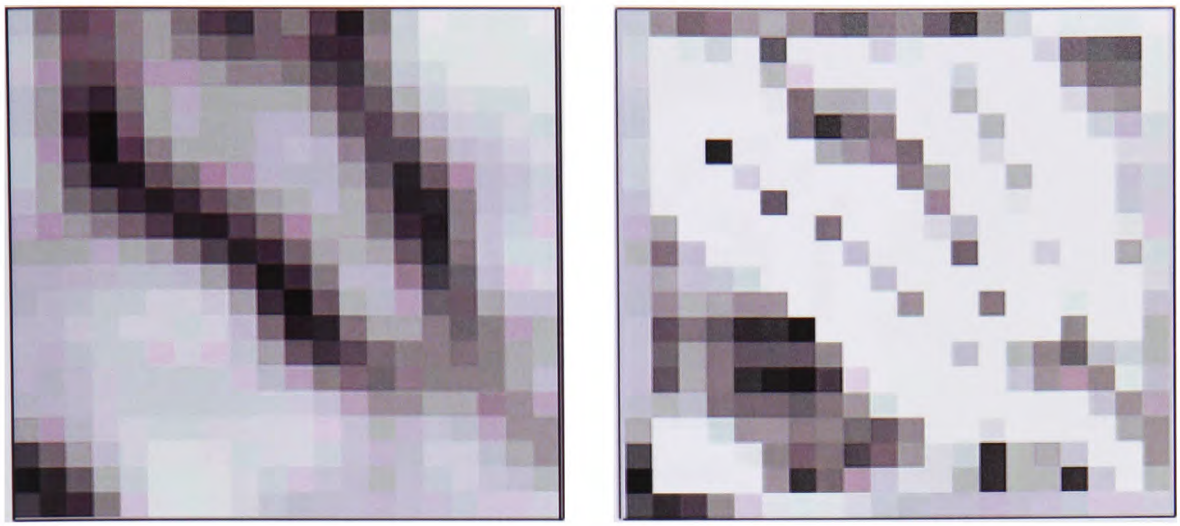
(a) (b)

Figure 5.34 (a) Grey scale image of each node represented as the average Euclidean distance between it and its eight neighbours on the 30 by 30 map for the thirty species data set. (b) Image produced after applying the Sobel edge detection algorithm.



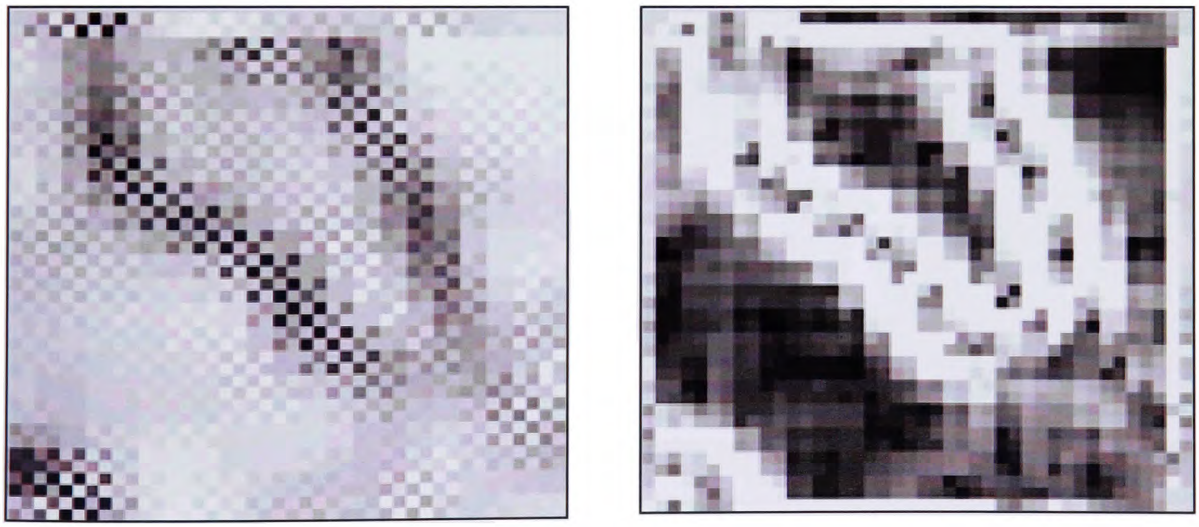
(a) (b)

Figure 5.35 (a) Grey scale image of each node as four blocks representing the Euclidean distance between a node and its four corner neighbours on the 30 by 30 map for the thirty species data set. (b) Image produced after applying the Sobel edge detection algorithm.



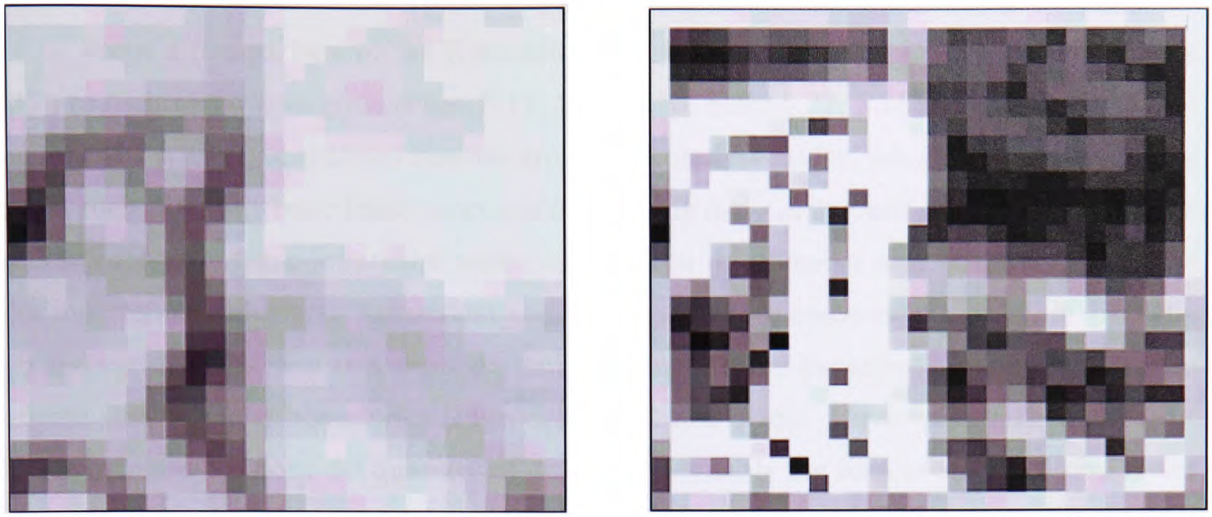
(a) **(b)**

Figure 5.36 (a) Grey scale image of each node represented as the average Euclidean distance between it and its eight neighbours on the 22 by 22 map for the sixty species data set. (b) Image produced after applying the Sobel edge detection algorithm.



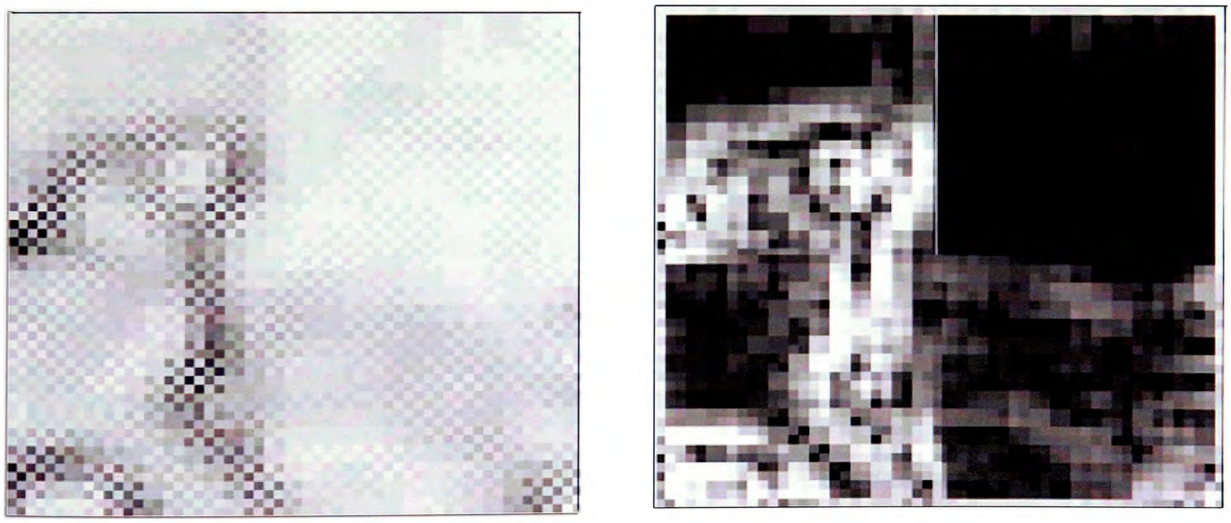
(a) **(b)**

Figure 5.37 (a) Grey scale image of each node as four blocks representing the Euclidean distance between a node and its four corner neighbours on the 22 by 22 map for the sixty species data set. (b) Image produced after applying the Sobel edge detection algorithm.



(a) (b)

Figure 5.38 (a) Grey scale image of each node represented as the average Euclidean distance between it and its eight neighbours on the 30 by 30 map for the sixty species data set. (b) Image produced after applying the Sobel edge detection algorithm.



(a) (b)

Figure 5.39 (a) Grey scale image of each node as four blocks representing the Euclidean distance between a node and its four corner neighbours on the 30 by 30 map for the sixty species data set. (b) Image produced after applying the Sobel edge detection algorithm.

5.9.2 Redundant Nodes

From a comparison of the Kohonen plots (Figs. 5.9, 5.10, 5.15, 5.18, 5.19, 5.24 & 5.26) to their respective grids (Figs. 5.11, 5.12, 5.16, 5.20, 5.22, 5.26 & 5.28), it is evident that areas of sparsely populated regions are represented by nodes, whose distance from any data is such that they have little or no reaction during the training and update process. The areas occupied by these redundant nodes can be seen to represent very definite boundaries between probable clusters. Redundant nodes are evident, to some extent, on all maps. The two and eleven class data sets show definitive results, where the redundant nodes appear to indicate very obvious boundaries (Figs. 5.11, 5.12 and 5.16). However, as the number of classes increases so does the quantity of data and inevitably the data overlap, leaving fewer sparsely populated areas and subsequently less possibility of redundant nodes.

5.9.3 Proportional Node Responses

In order to assess this method threshold values were chosen that imposed the same number, or as close to the same number, of classes present in the respective data sets. The thresholds imposed for this method, and as a starting point for over-selection of centers in the visual population density method, are shown in Table 5.5 Node allocation to the chosen centres are shown in Figures 5.40-5.48. The 8 by 8 map produced almost ideal clustering of the two group data set (Fig. 5.40), while the 24 by 24 indicated some areas of discrepancy (Fig. 5.41). Comparing the results to the threshold grids (Figs 5.13, 5.14, 5.17, 5.21, 5.23, 5.25 & 5.27) indicates only slight similarity. Even where cluster distribution is distinct (*e.g.* 11 species data set) this method heavily partitions many obvious groups, while others are merged together.

Table 5.5 Threshold values chosen for the four data sets on each respective map size for the Proportional Node Response Method (PNR) and as a starting point for the Visual Population Density Method (VPD). Due to software limitations only the 22 by 22 map for the 60 species data set was considered for the proportional node response method.

Data set	Map size	Threshold range	Threshold	Centers	Method
Two group	8 by 8	0 - 0.068	0.045	2	PNR
"	"	"	0.02	21	VPD
"	24 by 24	0 – 0.01	0.01	2	PNR
"	"	"	0.006	20	VPD
Eleven species	22 by 22	0 - 0.021	0.009	14	PNR
"	"	"	0.005	64	VPD
Thirty species	24 by 24	0 - 0.021	0.006	17	PNR
"	"	"	0.005	34	PNR
"	"	"	0.004	84	VPD
"	30 by 30	0 - 0.0165	0.007	7	PNR
"	"	"	0.004	30	PNR
"	"	"	0.003	54	VPD
Sixty species	22 by 22	0 - 0.01777	0.01	5	PNR
"	"	"	0.005	29	PNR

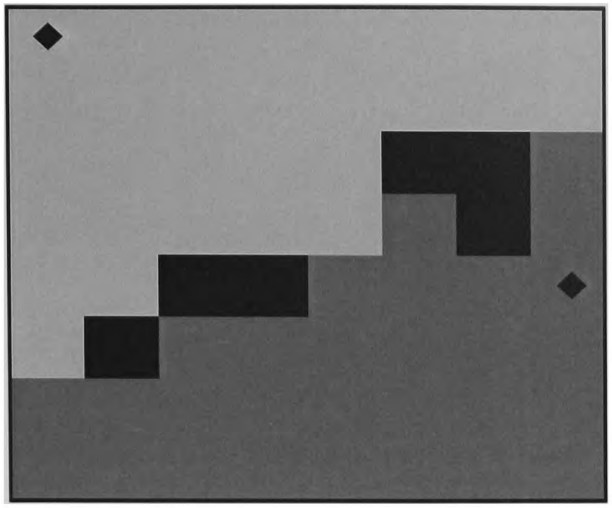


Figure 5.40 Groupings produced by the proportional node response method on the 8 by 8 Kohonen map for the two class data set. A threshold value of 0.045 selected 2 centres. Each group generated by the method are shown in different colours with the centres selected by the method depicted by ◆.

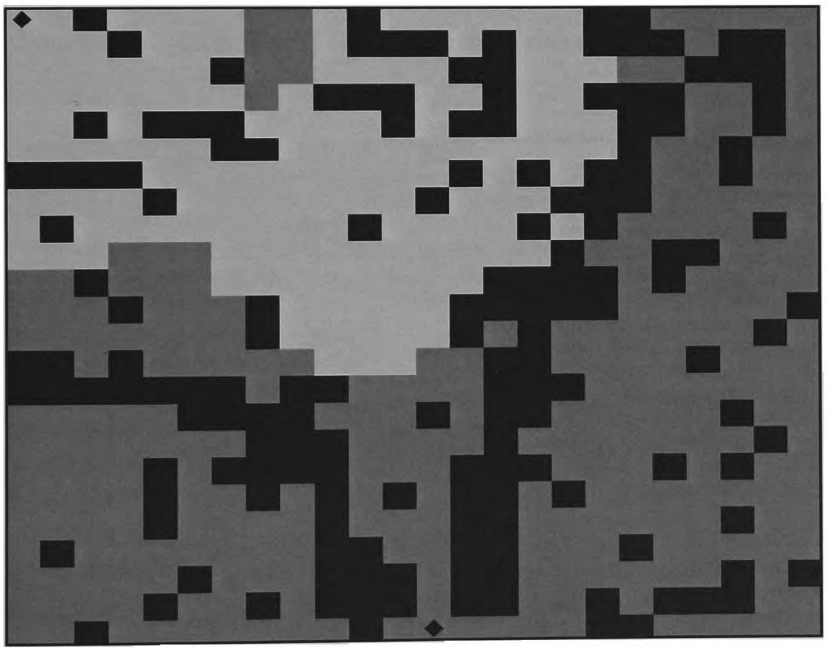


Figure 5.41 Groupings produced by the proportional node response method on the 24 by 24 Kohonen map for the two class data set. A threshold value of 0.01 selected 2 centres. Each group generated by the method are shown in different colours with the centres selected by the method depicted by ◆.

Note : On the above and following figures (*i.e.* Fig's 5.40-5.75) nodes allocated to the same class, by the method being discussed, are indicated as the same colour. These colours are not indicating particular species or taxonomic groupings.

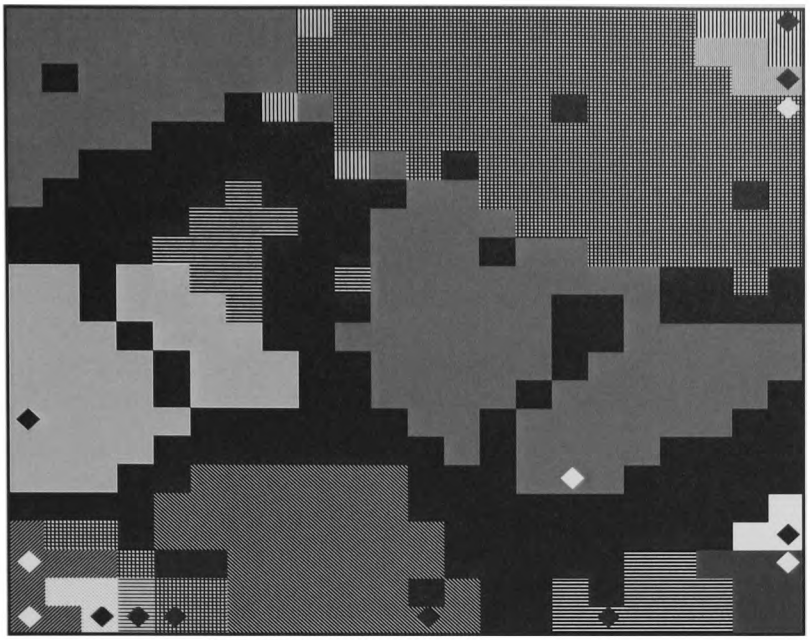


Figure 5.42 Groupings produced by the proportional node response method on the 22 by 22 Kohonen map for the eleven species data set. A threshold value of 0.009 selected 14 centres. Groups generated by the method are shown in different patterns with the centres selected by the method depicted by ♦ (black or white dependent upon background colour).

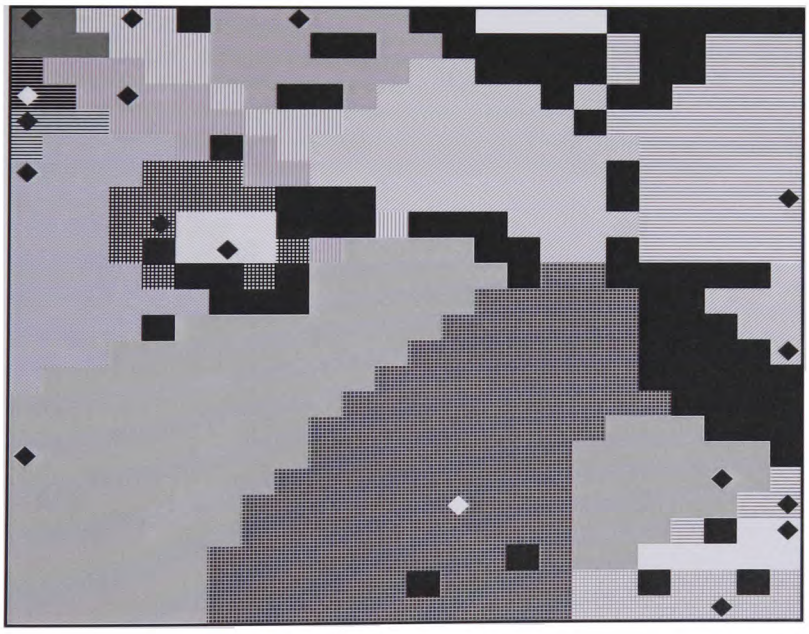


Figure 5.43 Groupings produced by the proportional node response method on the 24 by 24 Kohonen map for the thirty species data set. A threshold value of 0.006 selected 17 centres. Groups generated by the method are shown in different patterns with the centres selected by the method depicted by ♦ (black or white dependent upon background colour).

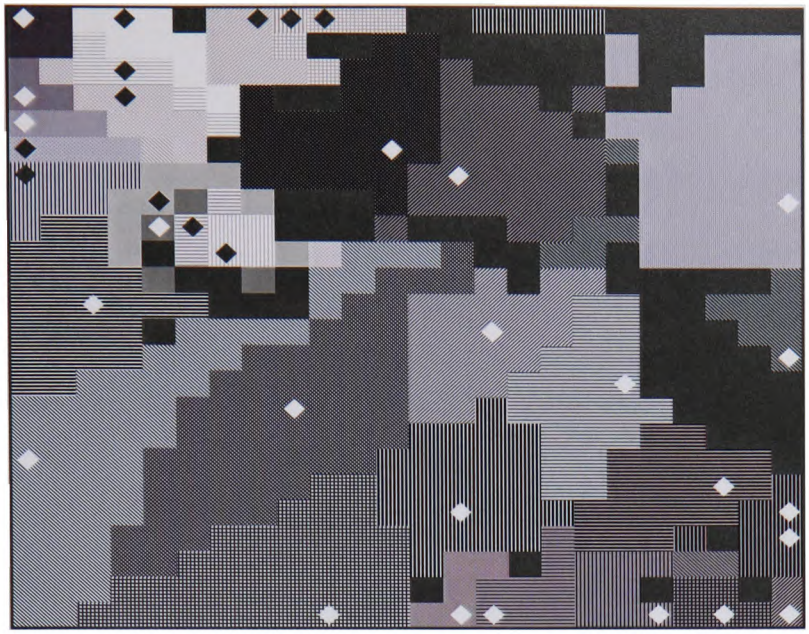


Figure 5.44 Groupings produced by the proportional node response method on the 24 by 24 Kohonen map for the thirty species data set. A threshold value of 0.005 selected 34 centres. Groups generated by the method are shown in different patterns with the centres selected by the method depicted by ♦ (black or white dependent upon background colour).

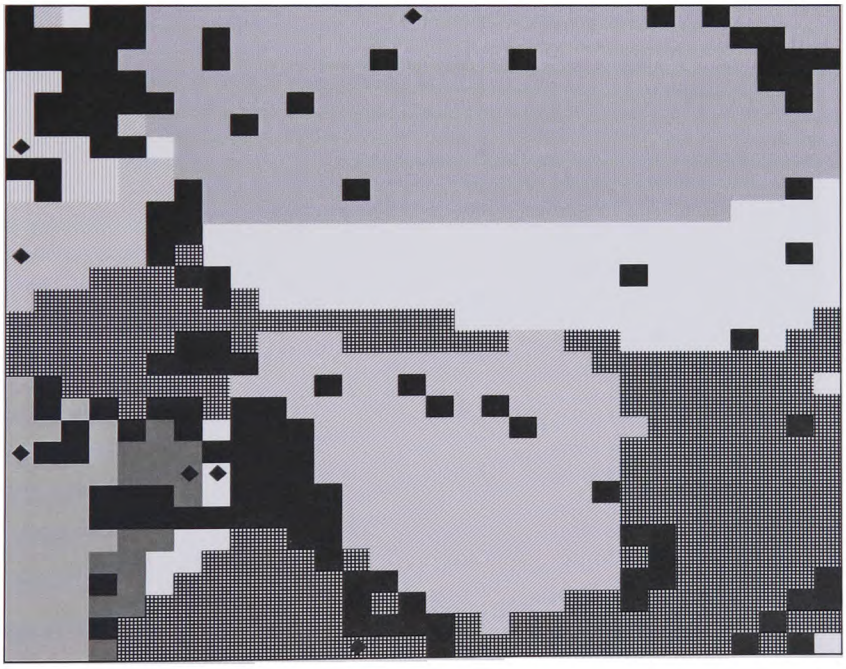


Figure 5.45 Groupings produced by the proportional node response method on the 30 by 30 Kohonen map for the thirty species data set. A threshold value of 0.007 selected 7 centres. Groups generated by the method are shown in different patterns with the centres selected by the method depicted by ♦ (black or white dependent upon background colour).

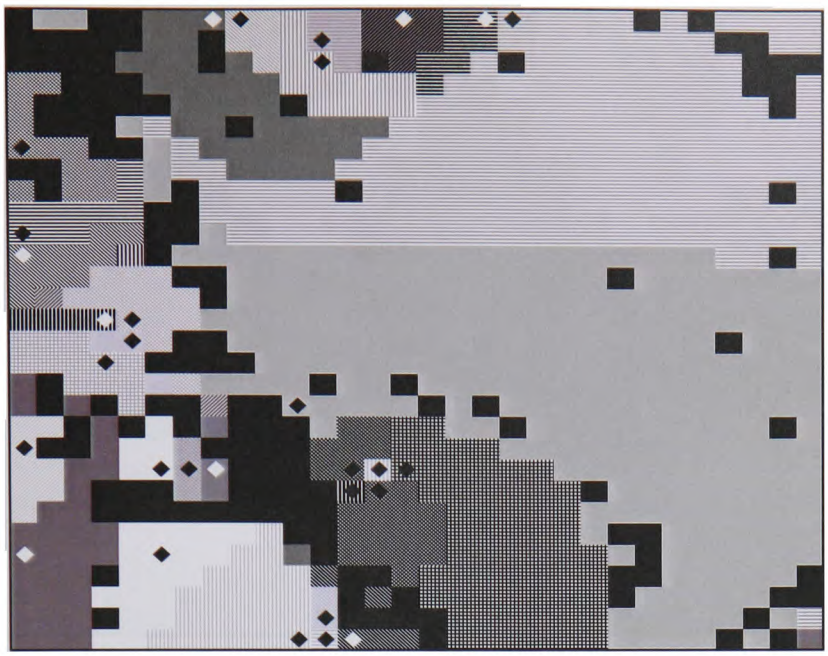


Figure 5.46 Groupings produced by the proportional node response method on the 30 by 30 Kohonen map for the thirty species data set. A threshold value of 0.004 selected 30 centres. Groups generated by the method are shown in different patterns with the centres selected by the method depicted by ♦ (black or white dependent upon background colour).

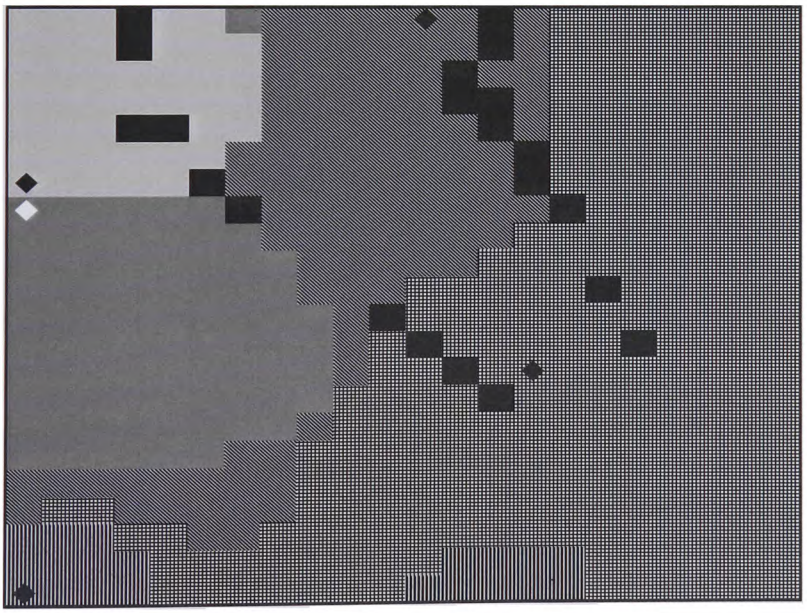


Figure 5.47 Groupings produced by the proportional node response method on the 22 by 22 Kohonen map for the sixty species data set. A threshold value of 0.01 selected 5 centres. Groups generated by the method are shown in different patterns with the centres selected by the method depicted by ♦ (black or white dependent upon background colour).

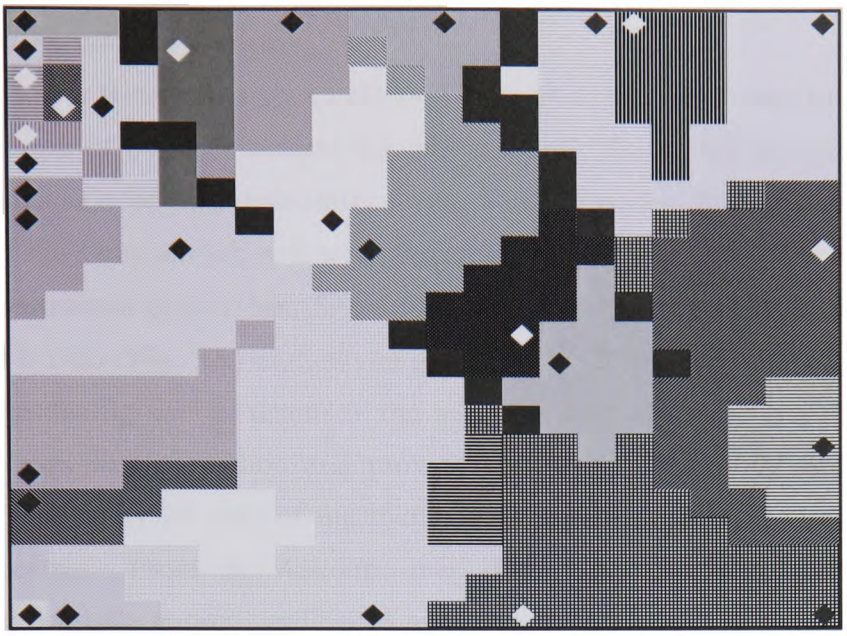


Figure 5.48 Groupings produced by the proportional node response method on the 22 by 22 Kohonen map for the sixty species data set. A threshold value of 0.0055 selected 29 centres. Groups generated by the method are shown in different patterns with the centres selected by the method depicted by ◆ (black or white dependent upon background colour).

5.9.4 Visual Population Density

The lower threshold values indicated in Table 5.5, were chosen to over select nodes as an indication of possible centres for this method. Those that are assumed to be high responders because of their proximity to the winning nodes during training, are possibly members of the same cluster, and were therefore ignored and others selected visually from the two-dimensional plots (Figs. 5.9, 5.10, 5.14, 5.18, 5.19, 5.24 & 5.26). With the two group 8 by 8 map (Fig. 5.9) and eleven group 24 by 24 map (Fig. 5.15), visual selection was easy, producing similar results to the networks own allocated groupings (Fig. 5.49 & 5.50). To support the method and provide clarification regarding node proximity, the analysis of the 24 by 24 map of the two class data was performed twice, using different nodes as centres. From the Kohonen map (Fig. 5.10) implying 3 possible clusters, 2 centres were initially chosen, and nodes assigned accordingly (Fig. 5.51a). This displayed groupings around probable centres identical to that of the Kohonen grid (Fig. 5.12). Two further positions were then chosen for the second analysis, selecting a different centre for the cluster which appears to have been separated. Again the groupings produced were identical to the network's allocation, supporting the theory that nodes on opposite sides may still belong to the same class (Fig. 5.51b).

The overlap of data on the 24 by 24 and 30 by 30 map, of the 30 species data set, makes visual identification of 30 centres very difficult. Despite the fact that generally class number would not be known, centres are chosen to compare the method to the groupings produced by the SOM. Therefore, 28 and 27 various centres were chosen respectively (Fig. 5.18 & 5.19) at the author's interpretation as 30 could not visually be detected. The surrounding nodes were clustered accordingly (Fig. 5.52 & 5.53).

Node selection for the 60 species data set on both the 22 by 22 and 30 by 30 maps was also difficult. The spread and overlap of data makes it hard to visually infer boundaries. Sets of centres were chosen for each map indicated on the two-dimensional plots (Figs.5.24 & 5.26). For the 22 by 22 map, 40 and 16 centres were selected, and for the 30 by 30 map, 49, 36 and 11 centres were chosen, with groupings shown accordingly (Fig. 5.54, 5.55, 5.56, 5.57 & 5.58). The selection of 16 and 11 centres on each map respectively, allowed closer analysis of the variation between flow cytometric signatures and the taxonomic divisions of phytoplankton, originating from morphometric similarities.

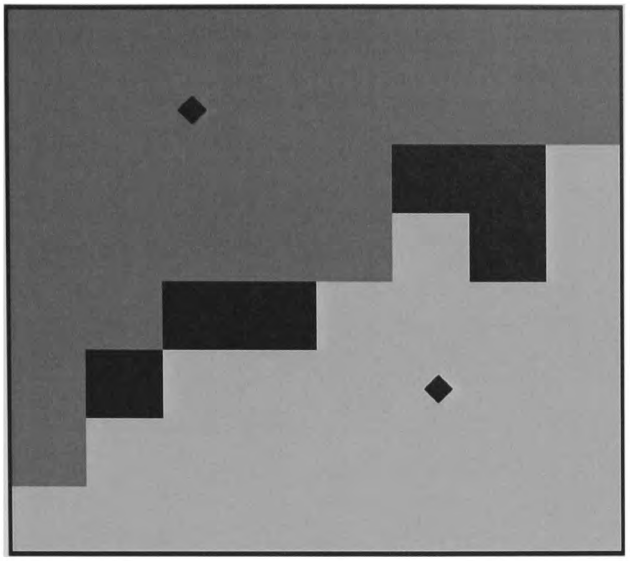


Figure 5.49 Groupings produced by the visual population density method on the 8 by 8 Kohonen map for the two class data set. Two centres were chosen. Groups generated by the method are shown in different colours with the centres selected depicted by ♦.

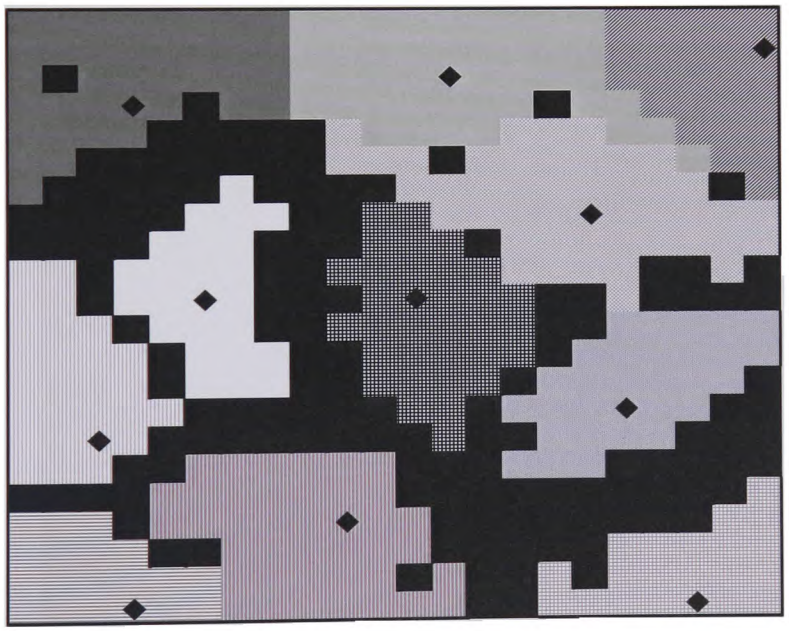


Figure 5.50 Groupings produced by the visual population density method on the 22 by 22 Kohonen map for the eleven species data set. 11 centres were chosen. Groups generated by the method are shown in different patterns with the centres selected depicted by ♦ (black or white dependent upon background colour).

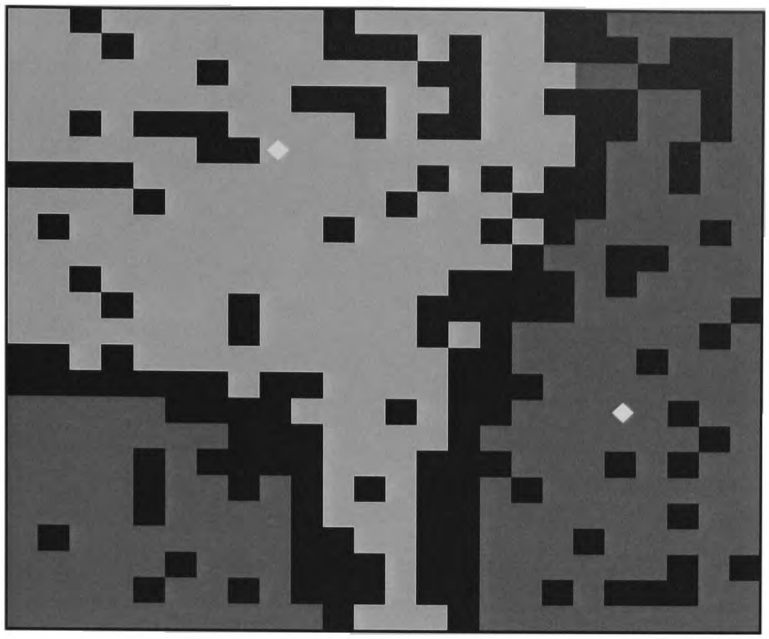


Figure 5.51a Groupings produced by the visual population density method on the 24 by 24 Kohonen map for the two class data set. Two centres were chosen. Groups generated by the method are shown in different colours with the centres selected depicted by ♦ (black or white dependent upon background colour).

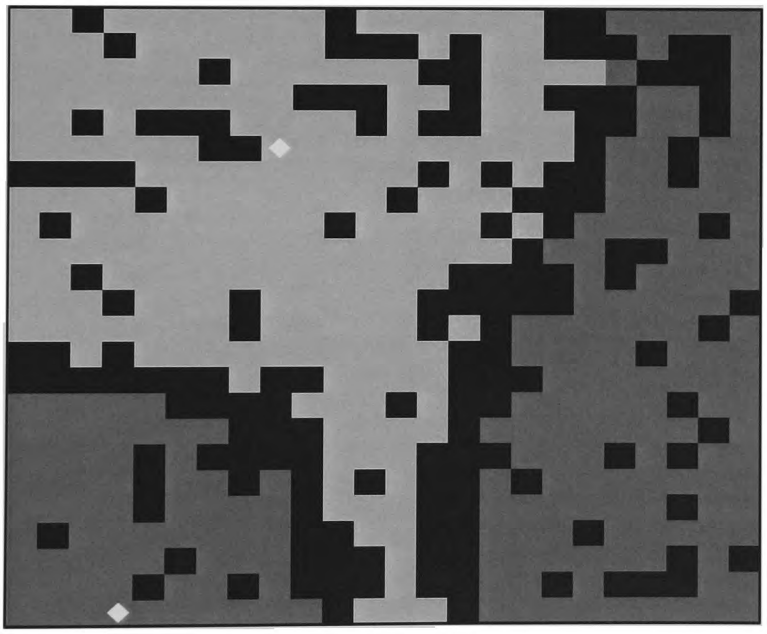


Figure 5.51b Groupings produced by the visual population density method on the 24 by 24 Kohonen map for the two class data set. Two centres were chosen. Groups generated by the method are shown in different patterns with the centres selected depicted by ♦ (black or white dependent upon background colour).

Figure 5.52 Groupings produced by the visual population density method on the 24 by 24 Kohonen map for the thirty species data set. 28 centres were chosen. Groups generated by the method are shown in different colours with the centres selected depicted by ♦ (black or white dependent upon background colour).

Figure 5.53 Groupings produced by the visual population density method on the 30 by 30 Kohonen map for the thirty species data set. 27 centres were chosen. Groups generated by the method are shown in different colours with the centres selected depicted by ♦ (black or white dependent upon background colour).

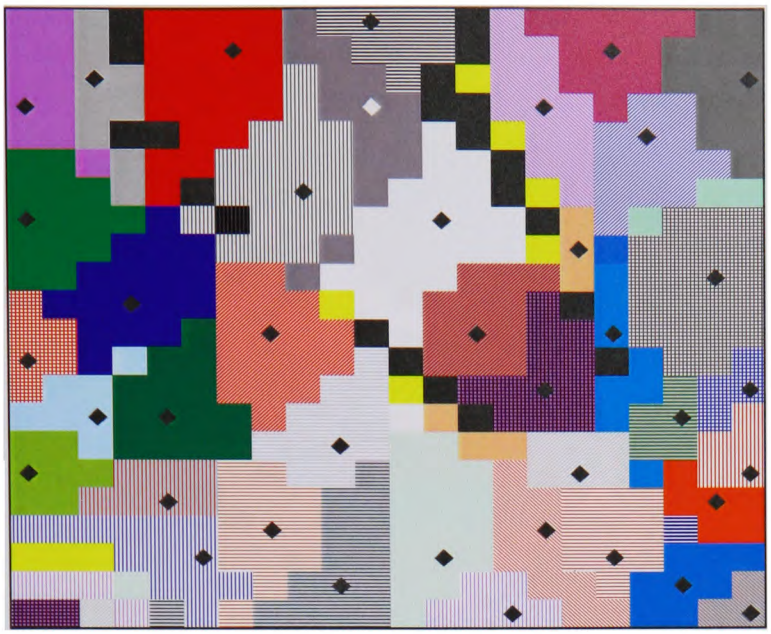


Figure 5.54 Groupings produced by the visual population density method on the 22 by 22 Kohonen map for the sixty species data set. 40 centres were chosen. Groups generated by the method are shown in different colours with the centres selected depicted by ♦ (black or white dependent upon background colour).

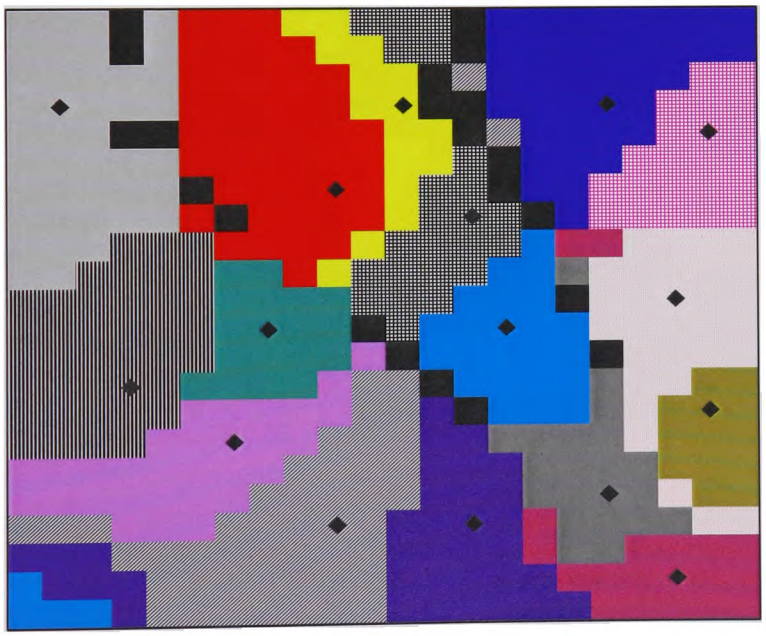


Figure 5.55 Groupings produced by the visual population density method on the 22 by 22 Kohonen map for the sixty species data set. 16 centres were chosen. Groups generated by the method are shown in different colours with the centres selected depicted by ♦ (black or white dependent upon background colour).

Figure 5.56 Groupings produced by the visual population density method on the 30 by 30 Kohonen map for the sixty species data set. 49 centres were chosen. Groups generated by the method are shown in different colours with the centres selected depicted by ♦ (black or white dependent upon background colour).

Figure 5.57 Groupings produced by the visual population density method on the 30 by 30 Kohonen map for the sixty species data set. 36 centres were chosen. Groups generated by the method are shown in different colours with the centres selected depicted by ♦ (black or white dependent upon background colour).

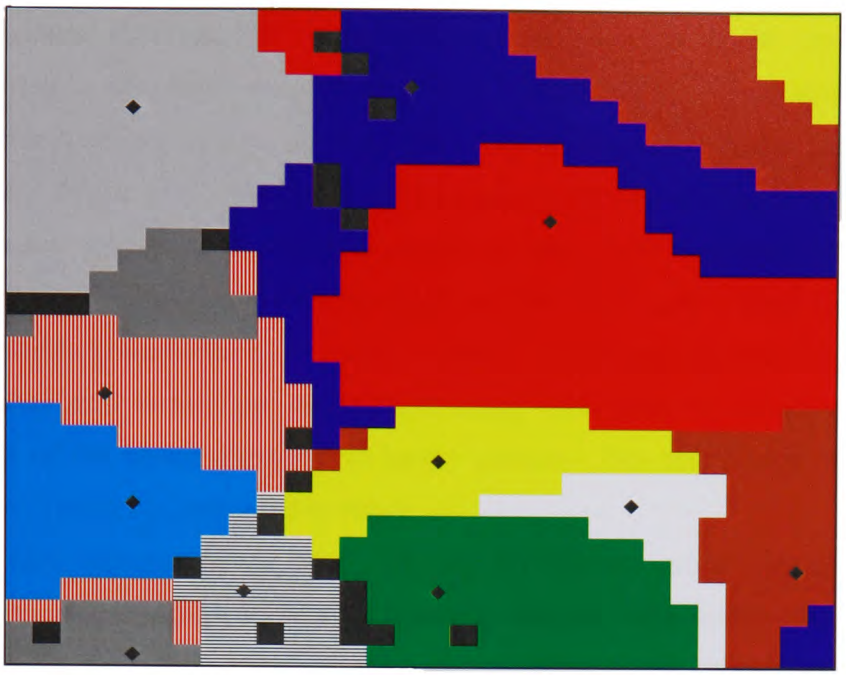


Figure 5.58 Groupings produced by the visual population density method on the 30 by 30 Kohonen map for the sixty species data set. 11 centres were chosen. Each group generated by the method are shown in different colours with the centres selected depicted by ◆ (black or white dependent upon background colour).

5.9.5 Agglomerative Clustering

As mentioned (Section 5.8.3), the inclusion of redundant nodes into this method requires the initiation of a large number of groups. Figures 5.59 – 5.75 show the images produced by the method, where greatest partitioning appears mainly in the sparsely populated regions of the map. Using 20 agglomerative clusters for the 8 by 8 map of the two group data set, gives almost identical results to that of the network's allocation (Fig. 5.59). Clusters containing individual or small numbers of nodes, tend to be on areas indicated as possible boundaries by the other methods. They appear to lie on both sparsely populated regions and redundant nodes, indicating their lack of association with the position vectors of the nodes allocated to larger clusters. For this reason, any group not allocated at least 4 nodes is indicated in white.

This is also apparent in the 24 by 24 map for the two group data set (Fig. 5.60), where 20 agglomerative groups indicate 3 possible clusters, with small groups of 3 nodes or less lying next to or within the proximity of redundant nodes or areas of low density. Using 100 agglomerative groups to initiate clustering of the 11 species data set, implies less than the number of classes known to be present (Fig. 5.61). As this number increases to 150, merged areas separate and more resemblance to the networks allocations are evident (Fig. 5.62). Further than this, the method begins to generate greater partitioning of data, and a greater number of clusters containing 3 nodes or less appear (Fig. 5.63).

The clusterings apparent for the 30 and 60 species data sets (Figs. 5.64-5.75), imply probable clusters that cover large areas of the map. Increasing the number of required groups causes the areas of low density, which have been allocated a smaller number of nodes to partition further, while the denser areas lose members only at their edges.

5.9.6 Decomposition

The 1 by 5 map trained on all 60 species as five taxonomic groups, exhibits a complete break up of the classes. The maximum representation by a node is 47% to the Dinoflagellates (Fig. 5.76). For the 1 by 11 map, not all 11 species are represented as primary reactions of the nodes (Fig. 5.77). Approximately 100% of *Micromonas pusilla* and *Cryptomonas rostrella* are allocated to individual nodes. The remaining classes are spread between either two or three nodes, with at least 50% of each species allocated to any one node (Fig. 5.78).

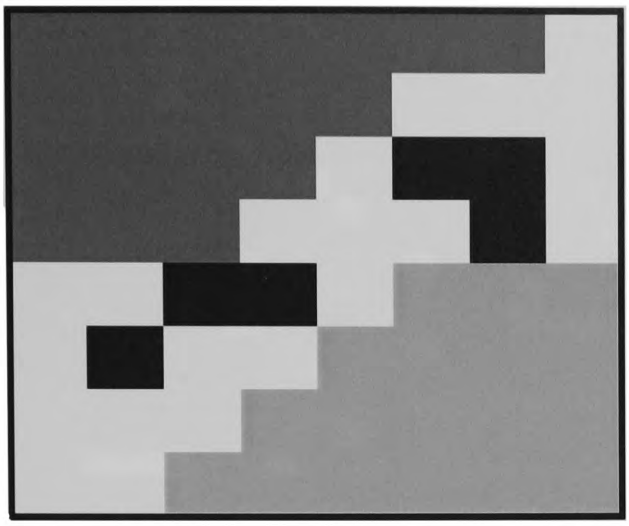


Figure 5.59 Groupings produced for the two class data set on the 8 by 8 map using the agglomerative method to cluster node position vectors. 20 groups were used for initialisation. Redundant nodes are shown in black and possible agglomerative boundaries in white.

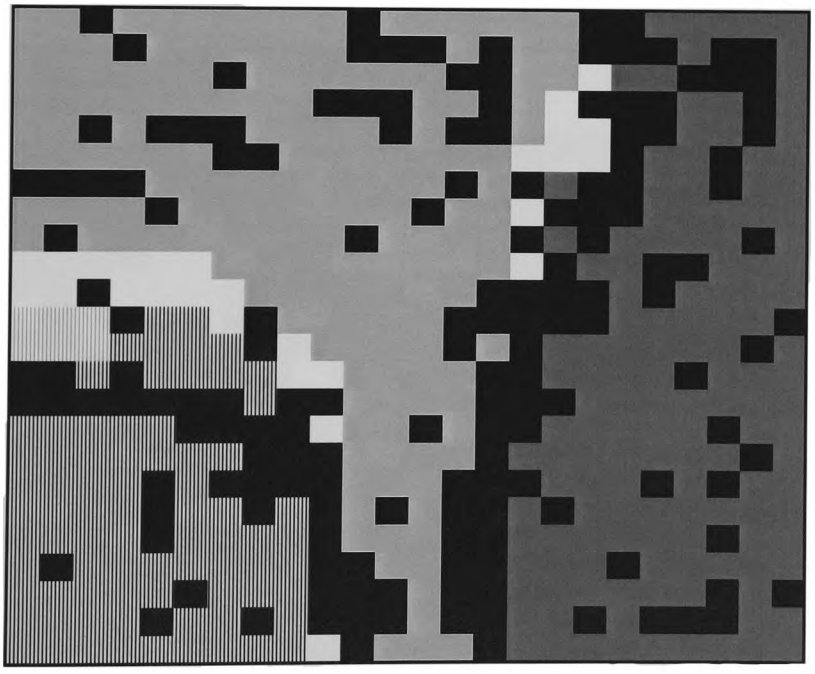


Figure 5.60 Groupings produced for the two class data set on the 24 by 24 map using the agglomerative method to cluster node position vectors. 20 groups were used for initialisation. Redundant nodes are shown in black and possible agglomerative boundaries in white.

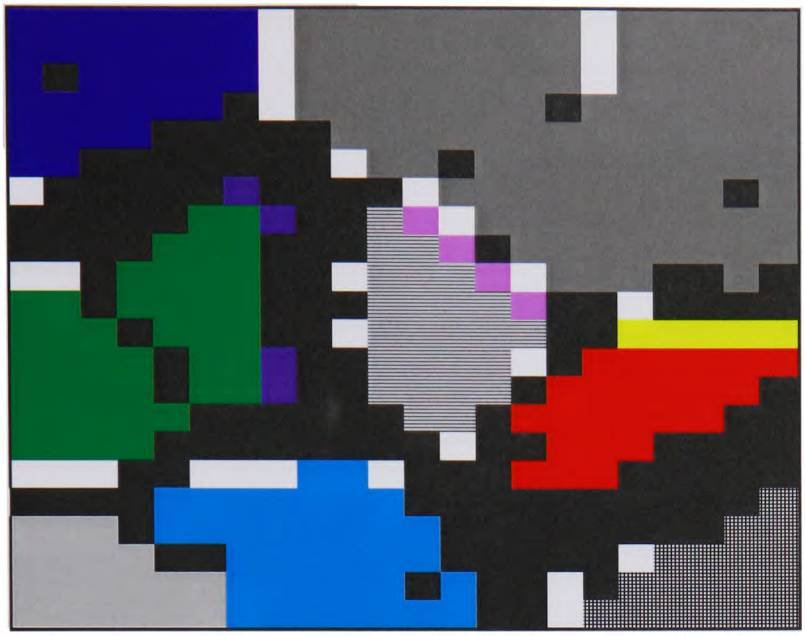


Figure 5.61 Groupings produced for the 11 species data set on the 22 by 22 map using the agglomerative method to cluster node position vectors. 100 groups were used for initialisation. Redundant nodes are shown in black and possible agglomerative boundaries in white.

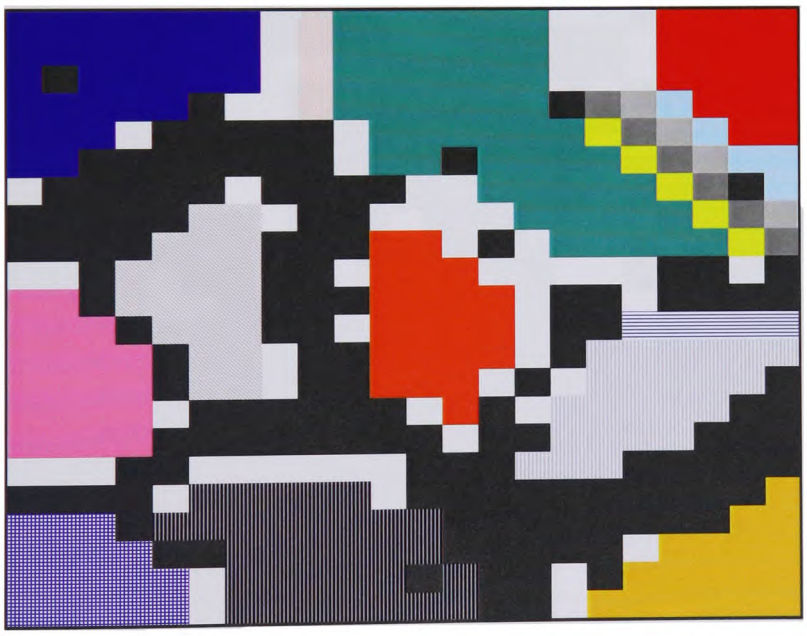


Figure 5.62 Groupings produced for the 11 species data set on the 22 by 22 map using the agglomerative method to cluster node position vectors. 150 groups were used for initialisation. Redundant nodes are shown in black and possible agglomerative boundaries in white.

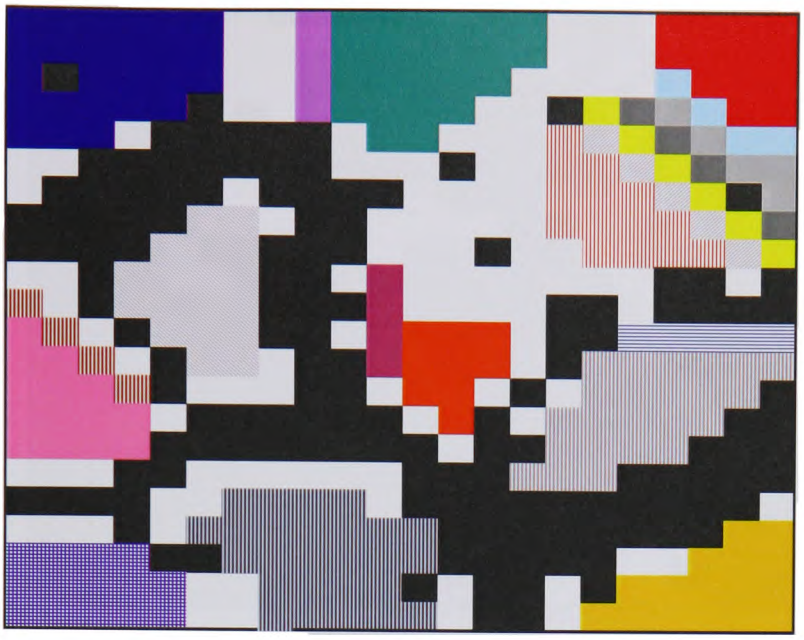


Figure 5.63 Groupings produced for the 11 species data set on the 22 by 22 map using the agglomerative method to cluster node position vectors. 200 groups were used for initialisation. Redundant nodes are shown in black and possible agglomerative boundaries in white.

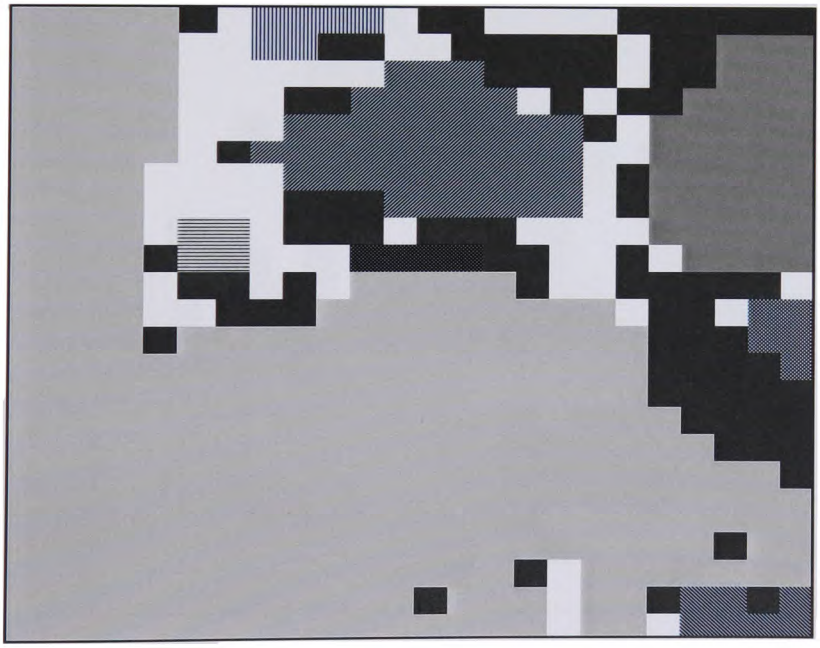


Figure 5.64 Groupings produced for the 30 species data set on the 24 by 24 map using the agglomerative method to cluster node position vectors. 100 groups were used for initialisation. Redundant nodes are shown in black and possible agglomerative boundaries in white.

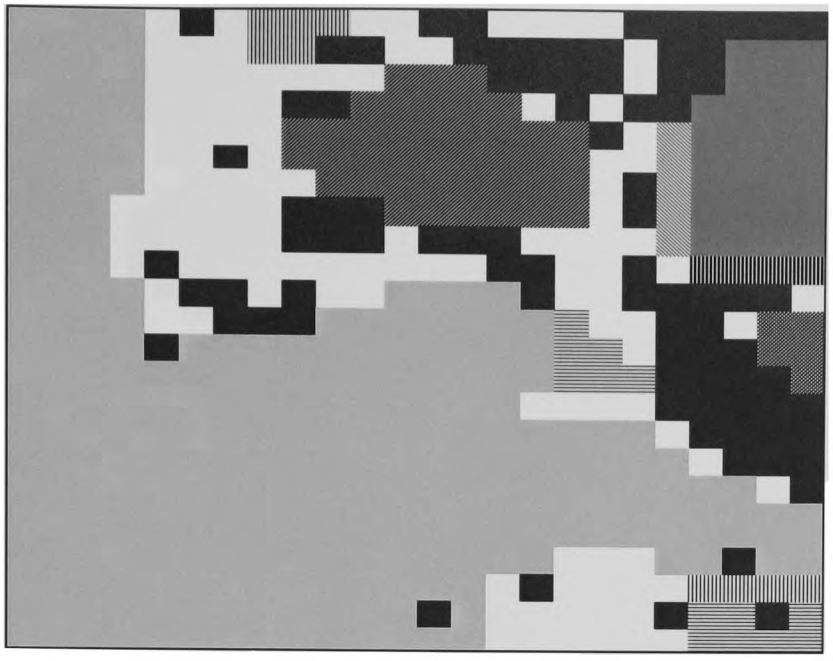


Figure 5.65 Groupings produced for the 30 species data set on the 24 by 24 map using the agglomerative method to cluster node position vectors. 150 groups were used for initialisation. Redundant nodes are shown in black and possible agglomerative boundaries in white.

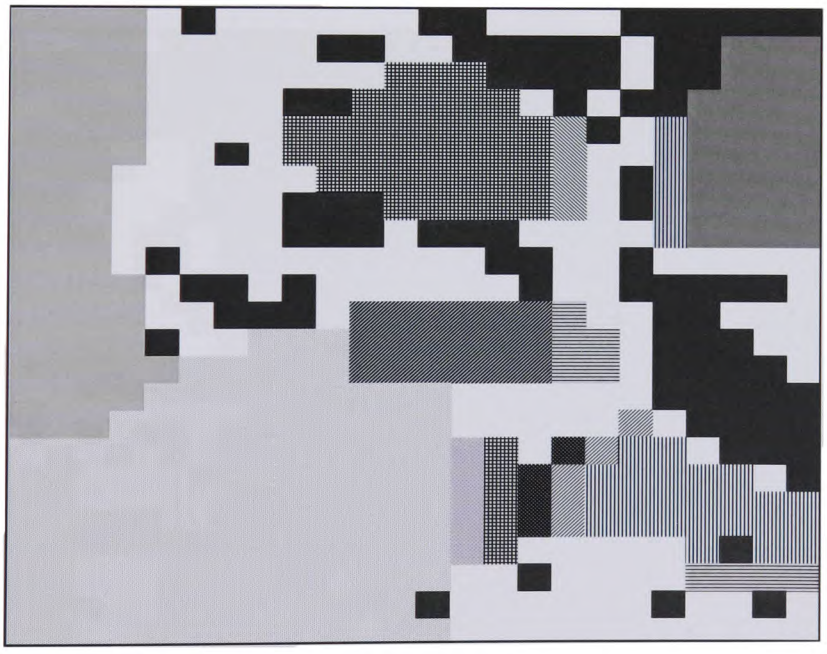


Figure 5.66 Groupings produced for the 30 species data set on the 24 by 24 map using the agglomerative method to cluster node position vectors. 200 groups were used for initialisation. Redundant nodes are shown in black and possible agglomerative boundaries in white.

Figure 5.67 Groupings produced for the 30 species data set on the 30 by 30 map using the agglomerative method to cluster node position vectors. 100 groups were used for initialisation. Redundant nodes are shown in black and possible agglomerative boundaries in white.

Figure 5.68 Groupings produced for the 30 species data set on the 30 by 30 map using the agglomerative method to cluster node position vectors. 150 groups were used for initialisation. Redundant nodes are shown in black and possible agglomerative boundaries in white.

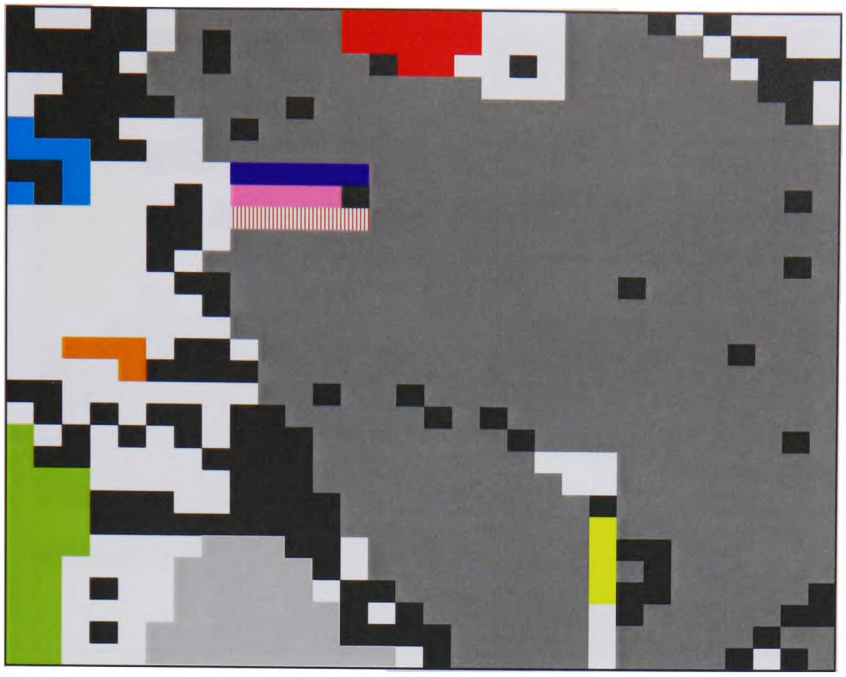


Figure 5.69 Groupings produced for the 30 species data set on the 30 by 30 map using the agglomerative method to cluster node position vectors. 200 groups were used for initialisation. Redundant nodes are shown in black and possible agglomerative boundaries in white.

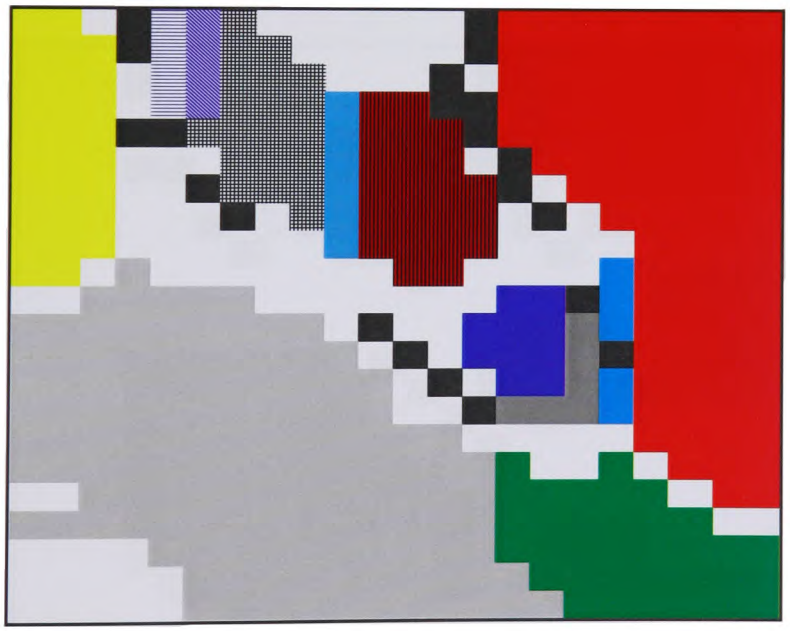


Figure 5.70 Groupings produced for the 60 species data set on the 22 by 22 map using the agglomerative method to cluster node position vectors. 100 groups were used for initialisation. Redundant nodes are shown in black and possible agglomerative boundaries in white.

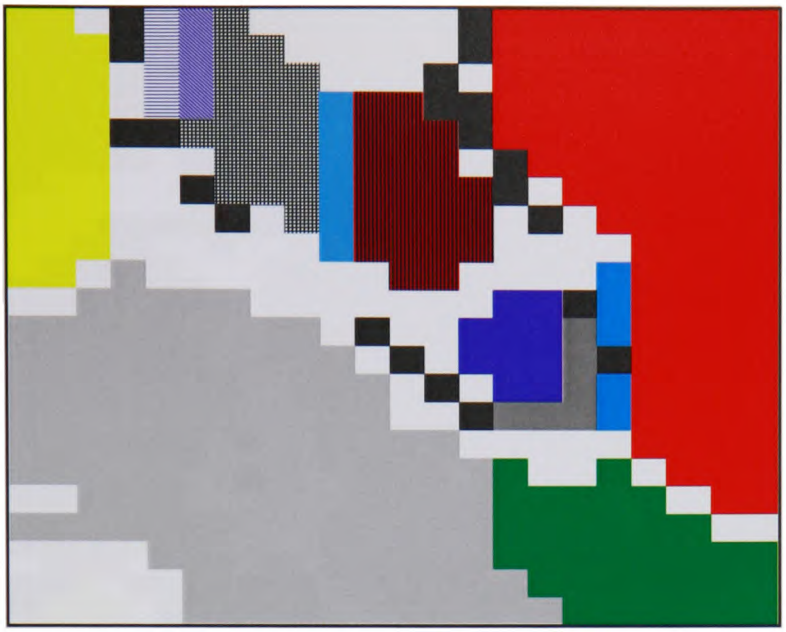


Figure 5.71 Groupings produced for the 60 species data set on the 22 by 22 map using the agglomerative method to cluster node position vectors. 150 groups were used for initialisation.

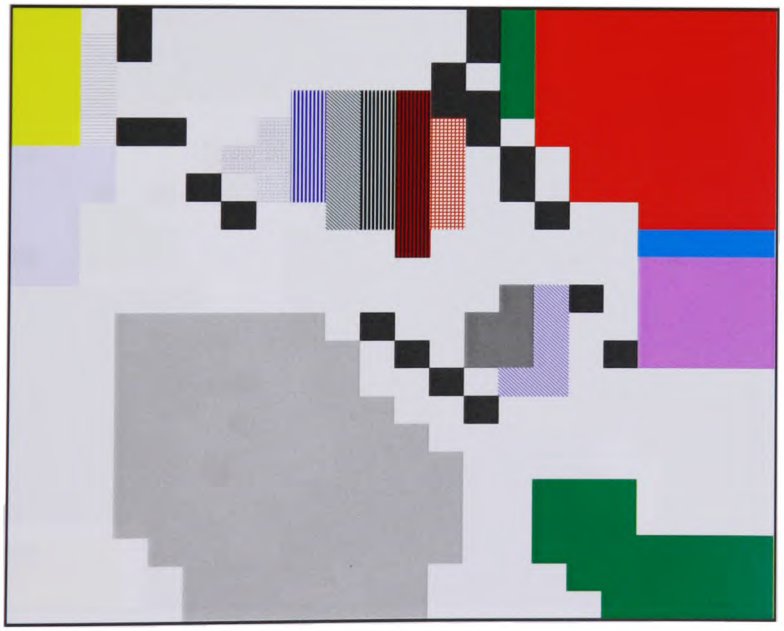


Figure 5.72 Groupings produced for the 60 species data set on the 22 by 22 map using the agglomerative method to cluster node position vectors. 200 groups were used for initialisation. Redundant nodes are shown in black and possible agglomerative boundaries in white.

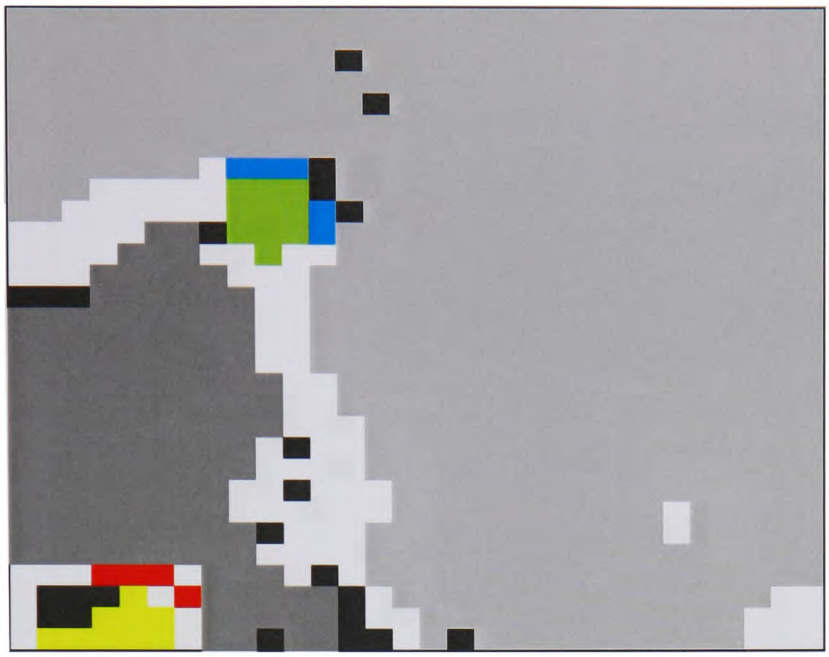


Figure 5.73 Groupings produced for the 60 species data set on the 30 by 30 map using the agglomerative method to cluster node position vectors. 100 groups were used for initialisation. Redundant nodes are shown in black and possible agglomerative boundaries in white.

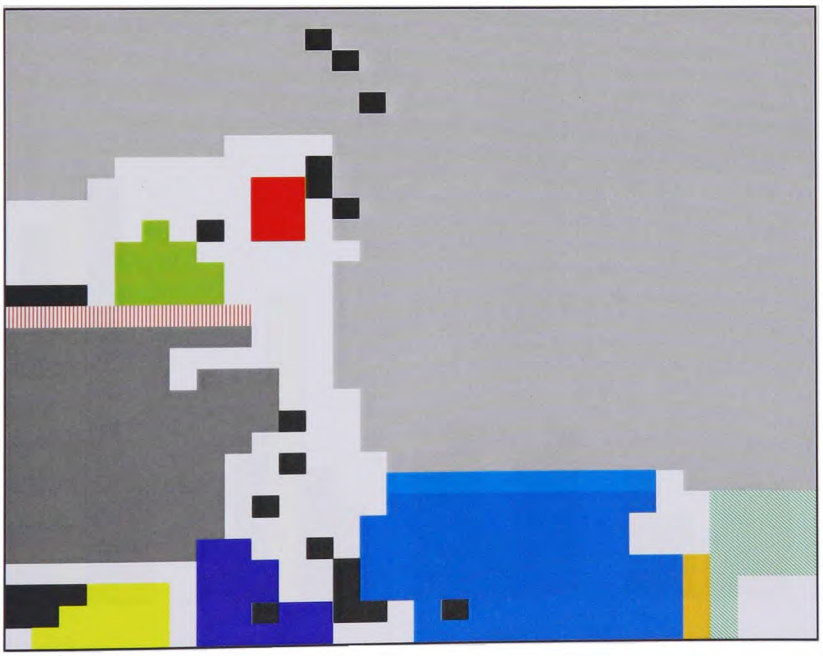


Figure 5.74 Groupings produced for the 60 species data set on the 30 by 30 map using the agglomerative method to cluster node position vectors. 150 groups were used for initialisation. Redundant nodes are shown in black and possible agglomerative boundaries in white.

Figure 5.75 Groupings produced for the 60 species data set on the 30 by 30 map using the agglomerative method to cluster node position vectors. 200 groups were used for initialisation. Redundant nodes are shown in black and possible agglomerative boundaries in white.

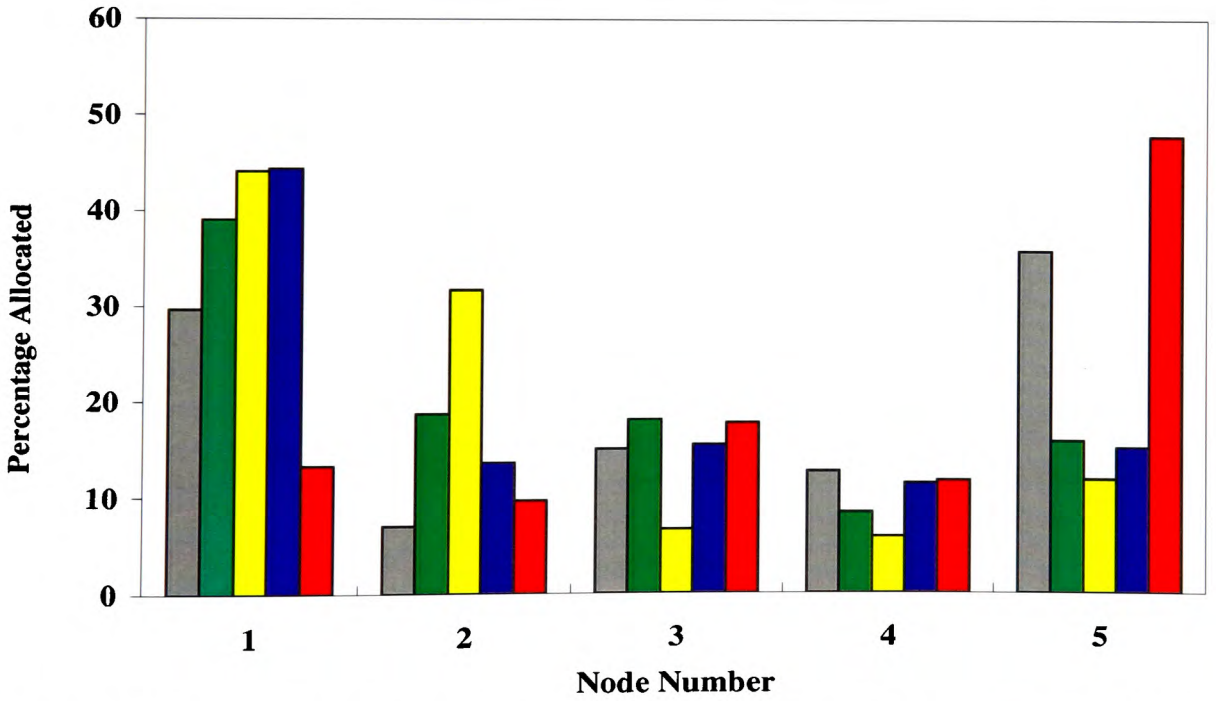


Figure 5.76 Percentage of each of the five taxonomic groups allocated, after training, to the 5 nodes in the Kohonen 1 by 5 map. ■ Cryptomonads ■ Dinoflagellates ■ Diatoms ■ Prymnesiomonads ■ Flagellates.

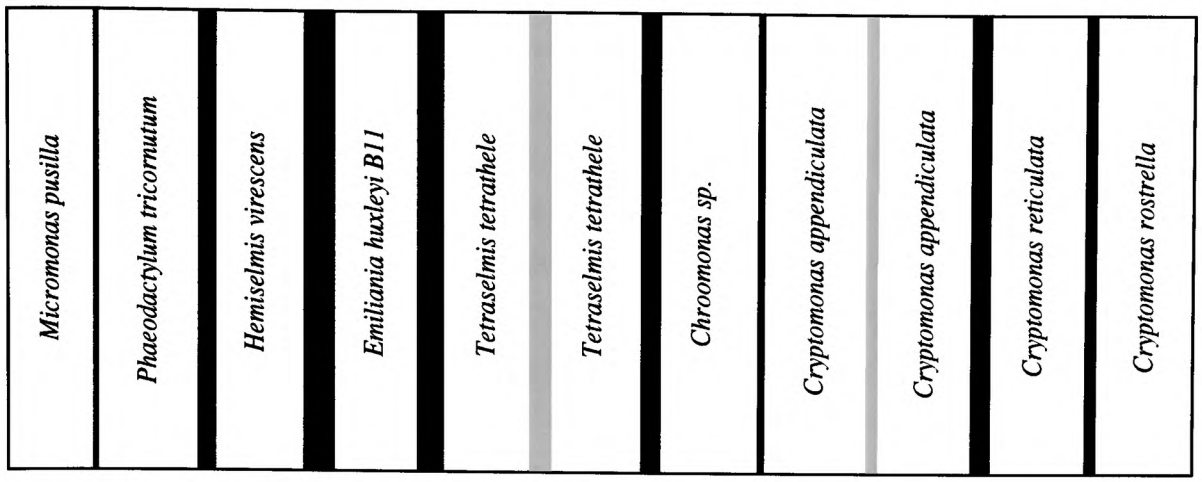


Figure 5.77 Primary node allocation to each of the 11 classes present in the eleven species data set by the 1 by 11 map. Euclidean distances are shown as black borders between nodes allocated to different classes and grey borders between nodes allocated to the same class.

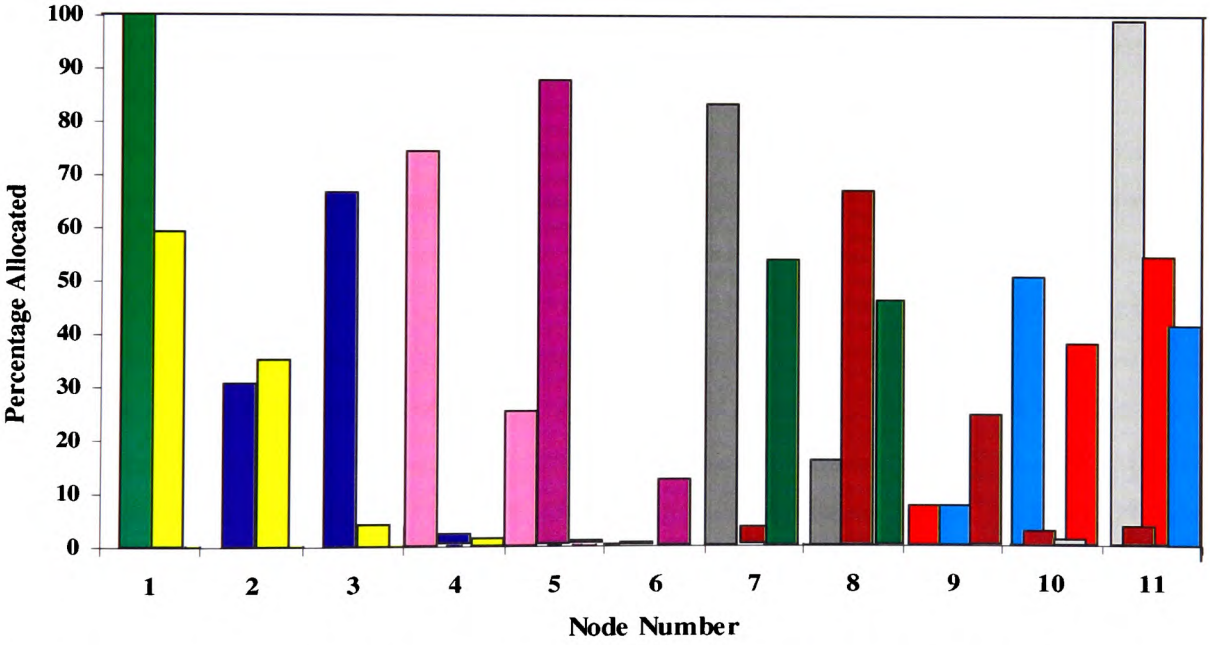


Figure 5.78 Percentage of each of the eleven species allocated, after training, to the 11 nodes in the Kohonen 1 by 11 map. *Chroomonas sp.* *Cryptomonas appendiculata* *Cryptomonas calceiformis* *Cryptomonas reticulata* *Cryptomonas rostrella* *Hemiselmis virescens* *Micromonas pusilla* *Porphyridium pupureum* *Tetraselmis tetrathele* *Emiliania huxleyi B11* *Phaeodactylum tricornutum.*

5.10 Discussion

5.10.1 Visualisation of Euclidean Hyper-Dimensional distances

5.10.1.1 Borders between nodes

When the data set contains separable classes, the imposition of a threshold upon Euclidean distances offers a very obvious cluster analysis method. The between class borders tend to be greater than the within-class borders, for the two group and eleven species data sets, illustrated by the threshold grids (Fig. 5.13, 5.14 & 5.17). As expected, the two group data set produces very definite borders on both maps, with no overlap of classes. The eleven species network also offers definitive boundaries around class allocated nodes, there are however some areas of discrepancy. A number of boundaries appear between nodes allocated to the same class, possibly due to either insufficient gating (*i.e.* removal of noise or outliers) or more probably the diversity between some same species cells. Minor overlap is evident between *Cryptomonas reticulata* and *Cryptomonas rostrella* (labels 4 & 5), *Cryptomonas calceiformis* and *Cryptomonas reticulata* (labels 3 & 4) and *Cryptomonas appendiculata* and *Cryptomonas calceiformis* (labels 2 & 3). Despite these species being correctly identified to at least 90% by supervised networks (Table 5.1), they are from the same genus and the small amount of overlap indicates some underlying optical similarities.

The overlap between species becomes more evident as class numbers increase, forcing the networks to generalise. Some of the placements are common to genus or taxonomic group, but others seem more obscure and threshold imposition demonstrates how realistic physical proximity actually is. Distinct species such as *Micromonas pusilla* (label 9) and *Porphyridium pupureum* (label 12), are quite obviously separated from all other classes when members of the 30 species data sets (Fig. 5.21 & 5.23). As class numbers are increased to 60 and the network is further forced to generalise, both species still exhibit their seperability (Fig. 5.27 & 5.29. *Micromonas pusilla* - label 18 and *Porphyridium pupureum* - label 24).

The method works well for a small number of separable classes and a network large enough to allow adequate representation of all data. As the class numbers increase so does data overlap, and imposition of high threshold values will show only coarse scale clustering. If the threshold value is decreased, fine scale clustering should become more apparent, but will increase the number of borders making interpretation of coarse

clustering more difficult. It would therefore be wise to employ an iterative process, beginning with a small threshold and increasing, thereby producing a hierarchy of probable clusters.

Although map size will not affect physical placement, if too small it may be necessary to consider a node's secondary reaction (*i.e.* the class for which its overall proportional response is second to highest) when discovering probable clusters of under-represented classes. This seems appropriate, as the secondary reaction of a node for a small map, is found to be the primary reaction of its neighbouring nodes when the map size is increased.

5.10.1.2 Grey scale representation

Using grey scale to represent the average Euclidean distance between a node and its eight neighbours, illustrates clearly that areas of greater Euclidean distances generally follow the path of sparsely populated regions. Using the four corner Euclidean distances for each node, gives insight into areas of finer borders, not apparent from the plot of average Euclidean distances. As the number of classes increases, overlap becomes more apparent and differences in grey scale are harder to detect. Variation in shading can be very slight and other than the obvious boundaries, interpretation is difficult. Employment of the Sobel algorithm offers little improvement in identifying possible *edges* and, in fact, appears to lose some information on finer boundaries, that was evident prior to application.

5.10.2 Redundant Nodes

The redundant nodes offer not only areas of probable boundaries, but information regarding extent of class similarity. As the grey scale representation indicates, redundant nodes are distanced from allocated nodes, making it therefore safe to assume that species separated by redundant nodes are relatively different. The eleven species data set exhibits this through a lack of redundant nodes between the species of the genus *Cryptomonas* (labels 2,3,4 and 5) (Fig. 5.16). Although species *Porphyridium pupureum* (label 8) and *Tetraselmis tetrathele* (label 9) are both Flagellates, they are separated by a very definite number of redundant nodes. This is explained in their optical parameters where *Porphyridium pupureum* has a much higher orange fluorescence, indicating its high phycoerythrin content. *Hemiselmis virescens* (label 6), which is non-typical of the

Cryptomonads in this database due to its low red fluorescence (Chapter 1, Fig. 1.3), migrates away from its group (labels 1-5), with fewer redundant nodes between it and the Prymnesiomonad, *Emiliania huxleyi B11* (10), than there are between it and the surrounding Cryptomonads. The difference in the flow cytometric signature of this species and its taxonomic group is again evident when present in the 30 (label 5; Fig. 5.21 & 5.23) and 60 group data sets (label 10; Fig. 5.27 & 5.29), where the nodes closest to it are allocated to Flagellates.

As the data size increases, sparsely populated regions decrease as more nodes required to model the data, thereby leaving fewer redundant nodes. However, with software limitations governing maximum map size (30 by 30) and number of classes (60), it is not surprising that the map is forced to generalise and allocate almost all nodes. With a larger map there would be more freedom to approximate the data distribution, possibly leaving more redundant nodes as probable boundaries. The 24 by 24 map trained on the 2 group data set exhibits this clearly (Fig. 5.10 & 5.11). A larger map has forced greater areas of redundancy than were apparent on the 8 by 8 map (Fig. 5.9), producing 3 very distinct clusters. However, the obvious problem is that through lack of generalisation, one of the groups has been divided. This does not mean the allocation is incorrect, but that class 2 is capable of further partitioning. This then requires some prior knowledge as to class number, or just the inference that there are 3 distinct separations. Knowing that there are 2 groups leaves the question of which two out of the three are actually one class. This is addressed in Section 5.10.4.

5.10.3 Proportional Node Response

Using this method alone for any data set, even one with distinct classes, produces very poor results. The 24 by 24 map trained using the two group data set, indicates that the two highest responding nodes are both located within the nodes allocated to class 2 (Fig. 5.41). The grouping produced thereafter is not representative of what is known to be the network's allocation.

This is reiterated with the eleven species data set. Despite consisting of well partitioned clusters, the 14 groups indicated by this method (Fig. 5.42) have little resemblance to the node allocation by the SOM (Fig. 5.17). On comparison with the two-dimensional plot (Fig. 5.15), the positions of nodes selected as centres are identified in

areas of high density, indicating as expected that high responding nodes are members of the same update region and therefore are probable members of the same cluster.

For the 11 species data set *Micromonas pusilla* and *Tetraselmis tetrathele* are allocated the least number of nodes by the SOM (Fig. 5.17, labels 7 & 9). As the number of events are equal for all classes, the internal variance of these two species must be minimal, and yet they are partitioned into 8 groups by this approach. Consequently, searching for only the required number of clusters, will inevitably partition highly populated regions into more groups than may be present, subsequently combining less populated areas that are obviously separate. For example, *Cryptomonas appendiculata*, *Chroomonas sp.* and *Porphyridium pupureum* (labels 2, 1 & 8 respectively, in the 11 species data set) .

5.10.4 Visual Population Density

When the number of classes are small this approach gives almost identical clustering to the node allocation by the SOM. There are areas of minor discrepancy for the 11 species data set (Fig. 5.50), where class membership of points that appear to lie between possible clusters is questionable (Fig. 5.15). This is particularly apparent amongst the genus *Cryptomonas*. Allocation, by the SOM, of a node to *Hemiselmis virescens* amongst majority node allocation to *Tetraselmis tetrathele*, may be due to outliers possessing similarities more akin to *Tetraselmis tetrathele* (Fig. 5.16). This is confirmed by the visual population density method, which clusters the node as *Tetraselmis tetrathele*. Although the physical location of this particular node implies its classification to be correct, the 24 by 24 map of the 2 group data set indicates that close physical location on the two-dimensional map does not necessarily imply close Euclidean distance. The two groups produced from the alternate centre selections are identical, despite the physical placement of the chosen centres, and the obvious partitioning of one of the classes (Fig. 5.50a & 5.50b). This implies that, although the SOM is topologically preserving, the projection of what in this case is seven-dimensional space onto a two-dimensional map, does not necessarily translate directly. This occurrence is especially evident in nodes located at the edge of the map, where mapping into two-dimensional has placed some away from possible neighbours. This dimensionality mismatch has been discussed in Section 5.5.

Both the 30 and 60 species data sets have some similar groupings to the SOM's node allocation, but the number of classes makes it difficult to interpret (Fig. 5.52-5.57). Again there is evidence that some classes, which are not close physical neighbours on the two-dimensional map, are actually grouped into the same cluster.

Using only 11 centres for the 60 species data set on the 30 by 30 map (Fig. 5.58), produces groups that can be interpreted with some comparability to the threshold grid (Fig. 5.29). Again the minor anomalies of the topology preservation are shown in particular areas, where for example the group constructed by the top centre position chosen, forks outwards encompassing nodes that lie between the heavier borders on the threshold plot. Although species belonging to the same taxonomic group are placed within close proximity, there are a number of probable groupings comprising a mix of taxonomic members, questioning their flow cytometric labels.

5.10.5 Agglomerative Clustering

Using 100 agglomerative clusters for the 11 species data set, groups the nodes allocated to the three species of the genus *Cryptomonas* together (Fig. 5.61, 5.17 Labels 3, 4 & 5). A slight partitioning of less than 4 nodes is present between *Cryptomonas calceiformis* and *Cryptomonas rostrella* (Fig. 5.17 Labels 3 & 5), but not between *Cryptomonas calceiformis* and *Cryptomonas reticulata* (Fig. 5.17 Labels 3 & 4). This would imply an area of lower density between the first two species. Similarities are also depicted between *Hemiselmis virescens* and *Emiliania huxleyi* B11 (Fig. 5.17 6 & 10), where the nodes allocated to these classes are merged. These two species have a common size range, but no other obvious similarities. As the number of agglomerative clusters is increased from 100 to 150, and finally to 200, the number of small groups allocated 3 nodes or less increases and appear on the outskirts of larger clusters, or surrounding redundant nodes. Also generated at these same points are smaller groups of 4/5 nodes, where clusters appear to be partitioned into finer groupings (Fig. 5.62 & 5.63). The same observations are made for both the 30 and 60 species data sets, (Figs. 5.64-5.75), however, as the number of classes increases the overlap becomes much more intense and problematic.

Although increasing the number of required groups partitions the heavily merged areas, it is minimal and dominates the borders, while smaller groups already clustered are

split into finer components. As this is a basic single linkage agglomerative technique, this result is expected. The process of adding nearest neighbours to a cluster, and the topology preserving properties of the SOM, will successively group those nodes on the outskirts first. As the required number of groups is increased, the build up of clusters halts sooner, and the relationship between individual nodes becomes gradually apparent. When the data comprises a mixture of densities, greater partitioning will be evident in sparsely populated regions, where higher Euclideans dominate the map. Thus, on completion, the clustering is more representative of the distances between individual nodes, than the map as a whole.

5.10.6 Decomposition

The results of this procedure further demonstrate the lack of similarity between the criteria used to construct morphometric groupings, and flow cytometric signatures. Using all 60 species from the 5 taxonomic groups (*i.e.* Cryptomonads, Flagellates, Diatoms, Dinoflagellates and Prymnesiomonads) produces primary reactions to only Flagellate and Prymnesiomonad species, with no more than half of any group allocated to any one node (Fig. 5.76). However, using only 5 nodes for 60 classes enforces considerable generalisation, resulting in a completely uneven distribution of all groups.

Despite the relatively distinct clusters of the 11 class data set, separation of the individual species is apparent on the 1 by 11 map (Fig. 5.77). *Cryptomonas calceiformis* (3) and *Porphyridium pupureum* (8) are not represented as the primary reaction of any node, whereas *Micromonas pusilla* and *Cryptomonas rostrella,* whose characteristics produce high supervised identification, are represented approximately 100% (Fig. 5.78). The split of the species, and apparent overlap, may be attributable to insufficient gating, However, the division is always between a species' primary node reaction and adjacent nodes as a secondary or tertiary reaction, indicating minor similarities due to the considerable generalisation.

5.11 Summary

There are both advantages and disadvantages to any boundary/cluster detection method. Whatever algorithm is employed, whether it is statistical, neural or otherwise, it will require some degree of user input. This input may be in a number of forms, but the more user defined parameters, the greater chance of error. As the purpose of the

unsupervised network is to determine underlying similarity and model the data accordingly, the output of any SOM for any suitable data set is *'correct'*, thereby alleviating the problem of inappropriate algorithm selection.

The proportional node response method gives the poorest results, even for a small separable set. A similar approach to this method was used for the analysis of residential property (Lewis, *et. al.* 1997), where a large set of historical data collected from mortgage transactions was used as inputs to a SOM. This experienced the same problem of possible centres belonging to the same group, causing incorrect clustering. In order to counter this the authors investigated the Gamma Test (Stefánsson, *et. al.* 1997). This is a data analysis routine that strives to estimate the best mean squared error that can be achieved, by any continuous or smooth modelling technique using the data. The mathematics of the algorithm will not be discussed, but the basic procedure is to select a highest responding node and, using the Gamma test, calculate the variance of all data represented by that node. Data are then sequentially added from n surrounding nodes and the Gamma test run again. Like other methods, inclusion or exclusion of the node is based on a pre-defined variance threshold, upon which the procedure continues until all neighbours within the threshold are accounted for, and the boundaries of the cluster are set. This is then repeated for the next highest responding node outside the cluster, and the process continues until either all nodes are clustered or all high responding nodes are accounted for, possibly resulting in a number of excluded outliers.

The most informative procedure appears to be the use of the boundary detection methods in conjunction with each other. The redundant node method offers an obvious partition where scarcity dominates regions of the map. Although map size is restricted by this particular software, a larger network will improve the method, offering greater areas of low population for larger more overlapping data sets, thereby indicating possible boundaries.

The visual population density method will give accurate clustering around a probable centre if the chosen centre is *appropriate.* This can be difficult with heavily overlapping data, but will again become more apparent with a larger map size. However, it must be considered, that flow cytometric characteristics of some species, may be too similar to ever allow adequate separation. As this approach is user-dependent, increased

overlap may make interpretation of probable centres more varied, and node allocations to possible groups may differ as a result of this.

Agglomerative clustering of the position vectors of nodes requires, perhaps the least amount of defined user input. Although a pre-determined number of classes are needed, no definite *a priori* knowledge is implied, except for that of user requirement. The initial selection of a larger number of groups than is assumed present, produces a range of small clusters allocated less than four nodes. Despite the appearance of some of these existing along redundant nodes or regions of high Euclidean distances, the method offers little in the way of conclusive boundaries, as the variety of inter-group relationships results in the partitioning of sparse areas and the merging of dense. This could be improved if the data was initially coarsely clustered, to remove the division of hierarchical differences, and possibly through employment of a mean vector as a measure of cluster membership. However, there is still a high risk of ambiguous groupings being produced, where some points are clustered through indirect relationships. This can result from the placement of an updated mean centre, being closer to an outsider than one of the original cluster is to another group.

While the redundant node method, Euclidean distances as boundaries and grey scale imagery give indications of the boundaries and distances relative to each other, the visual population density method compensates for any input/output dimensionality mismatch. This ensures that nodes mapped to different areas of the grid are grouped together if, in hyper-dimensional space their position vectors are neighbouring. As the methods define the areas of the map rather than individual data points or means, the problem of closeness experienced by some methods when clusters are not spherical is eliminated.

When a data set comprises both dense and sparsely populated regions, attempts to partition it may result in loss of information, especially where the within-class similarities vary greatly between different clusters. For example, the methods for visualising Euclidean distances (Section 5.7.1), indicate greater distances as either thicker borders (Section 5.7.1.1), or darker areas (Section 5.7.1.2). As the boundaries are depicted in relation to each other, only areas of maximum Euclidean distances may be illustrated, leaving denser areas, that may contain clusters, untouched. Similarly, an approach that starts with minimum Euclidean distance and builds up (Section 5.7.5) may reach the same

conclusion, as attempts to partition sparse areas will inevitably cluster dense areas in the process. To avoid this, somewhat, the decomposition method can be used as a front end procedure, where primary partitioning can discover coarse clusters, producing sub-sets of data comprising finer clusters that can be sequentially detected by one of the other methods. However, in proceeding in this manner the hierarchical structure must be maintained, and within-group and between-group similarities and dissimilarities of subsets, must be held relative to their parent or same level group. This is of particular importance when considering the possibility of an alternative structuring system where the relationships between hierarchies would be relative.

5.12 Conclusion

The class membership of the data sets used for this research were known. In a natural classification process, unless similarities are being compared, this will very rarely be the case and some form of external knowledge may be necessary. However, clustering procedures will always require some degree of *a priori* knowledge or pre-defined user requirement, such as number of clusters or a similarity criteria. Even if a number of clusters is defined during analysis, unless the data are identical, further partitioning will always be possible. If an algorithm separates data to a high (or low) degree, it does not indicate a poor method but simply that further partitioning (or merging) may be possible, and it is up to the user to define how much of a similarity is required. This has been apparent for some species of phytoplankton, where map placement is not due to translation from seven-dimensional to two-dimensional, but diversity within the strain itself. It may be possible to employ a similarity coefficient, such as distance coefficient, association coefficient or correlation coefficient, which may serve to assess the extent to which a partitioned group can be considered as one (not discussed here).

The easy and flexible detection of clusters and cluster boundaries, is not only necessary in furthering the use of SOMs, but because classification is a continual process, demanding constant revision as requirements change. The lack of similarity between the flow cytometric signatures and morphological groupings has been demonstrated, as well as how and why a possible alternative structuring system could be introduced. Automated flow cytometry groupings may provide a basis for a functional classification, and thereby introduce a more appropriate division, consequently improving supervised identification.

6 Biological Variation

6.1 Introduction

The experiments documented have been performed solely on PRiME 1 data (phytoplankton grown under a light source of 50μmol quanta $m^{-2}s^{-1}$). In order to provide an idea of natural variation, particularly in pigment content, a new data set was cultured by Plymouth Marine Laboratory (PRiME 2), using a different illumination of 12μmol quanta $m^{-2}s^{-1}$. This chapter examines the performance of both the original multi-class network and the multiple network architecture, when presented with the new data set. Overlap and differentiation between PRiME 1 and PRiME 2 data sets are also examined.

In addition to the laboratory grown cultures, the performance of both architectures was compared when analysing laboratory cultured mixtures and field samples.

The results of field studies carried out in the North Atlantic are also documented. However, these provide information on the generalisation ability of the original multi-class network only, and are not considered for the alternative multiple network architecture.

6.2 Illumination Variance

As the illumination condition under which cells are grown affects pigmentation, the optical parameters of some species will inevitably alter. To assess the generalisation ability of both architectures on this change alone, a number of experiments were performed.

6.2.1 Experimental Procedure

The PRiME 2 data set, supplied by Plymouth Marine Laboratories, consisted of the same species as that of PRiME 1, excluding *Scrippsiella trochoidea* and *Prorocentrum minimum*, which were too dilute to be analysed by the flow cytometer. Culturing conditions were also the same with the exception of illumination, which was reduced to 12μmol quanta $m^{-2}s^{-1}$. Data preparation and pre-processing was as in Chapter 2 (Section 2.6).

6.2.1.1 Multi-Class Network Architecture

Three training files were created for the original multi-class architecture. The first comprised 500 events per species (60 species) from the PRiME 2 data set only. 500

unseen events from PRiME 2 were used as a test file. The second file was constructed of 60 species from PRiME 1 and the same 60 species, but grown under different conditions, from PRiME 2. This meant each species was represented twice, but they were treated as separate classes, *i.e.* 120 in total. Each class contained 500 events, as did an unseen test file constructed in the same manner. Finally, a training file containing 60 classes was created with 600 events per class, 300 events from a PRiME 1 species and 300 events from the same species in PRiME 2. An unseen test file of the same structure was used to asses performance.

Each of the respective files was used to train an RBF network, (Chapter 2, Section 2.7.3), with 6 hidden layer nodes per class, employing a Mahalanobis distance metric. All networks were trained three times from different initialisation points.

Additionally, the optimum network trained on PRiME 1 data from Chapter 3 (Section 3.8) was presented with the PRiME 2 test file. This was repeated, using the PRiME 1 test file applied to the network above trained solely on PRiME 2 data.

6.2.1.2 Multiple Network Architecture

Nine sets of 60 single species training files were created as described in Chapter 4 (Section 4.6.1). The exact structure of each file is shown in Table 6.1, where class A refers to a particular species of interest, whether it is events from PRiME 1, PRiME 2 or both, and class B to the structure of the background class. Unseen test files, of the same content as the class of interest in the training files, were constructed to assess performance. Nine sets of 60 single species networks were trained as described in Chapter 4 (Section 4.6.2), using a random kernel placement strategy of 10 nodes for the class of interest and 1 node for the background class, all employing a Euclidean distance metric.

A separate set of 120 training files was also created, representing the 60 species from PRiME 1 and the 60 species from PRiME 2 as 120 separate classes of interest. Each training file contained 500 events for the class of interest and 250 events for each of the remaining 119 species representing the background class. A test file of unseen events was constructed, containing 500 events per species from PRiME 1 and 500 events per species from PRiME 2. 120 single species networks were trained, again using a random placement strategy of 10 nodes in the class of interest and 1 node in the background class, all employing a Euclidean distance metric.

Table 6.1 Event numbers for each of the 9 sets of 60 files constructed for training the 9 sets of 60 single species networks. Overall identification and confidence of identification from the RBF decision networks subsequently trained on the outputs of the single species networks.

Class A (Total events)		**Class B (Events per class)**		**Percentage successful identification**	**Confidence**
PRiME 1	**PRiME 2**	**PRiME 1**	**PRiME 2**		
0	500	0	500	77.8	77.9
0	500	500	0	76.9	76.7
500	0	0	500	78.1	79
250	250	500	0	66.3	67.9
250	250	0	500	66.4	67.9
500	0	250	250	77.9	78.1
0	500	250	250	77	76.5
250	250	250	250	66.8	65.6
500	500	500	500	67.1	67.5

From the outputs of the nine sets of 60 single species networks, nine input files were constructed to train RBF decision networks (Chapter 4, Section 4.4.3). These comprised 60 parameters and 300 events per class. The outputs from the 120 single species networks were converted into a 120 parameter, 300 event per class training file.

Independent test files of 500 events per class were constructed accordingly, dependent upon the primary identification requirements of the particular RBF decision network. All RBF decision networks used 3 hidden layer nodes per class, employing a Euclidean distance metric. All networks were trained three times from different initialisation points.

Additionally, the 60 single species networks trained using PRiME 2 data were presented with a PRiME 1 test file. The results were converted to form an input file for the RBF decision network trained on the 60 parameter PRiME 2 data. This was then performed in reverse, replacing PRIME 1 with PRIME 2, and vice versa.

6.2.2 Results

6.2.2.1 Multi-Class Network Architecture

The overall identification success for PRIME 2 data was 78.4% with a 77.8% confidence. Individual species identification ranged between 42.1% for *Prorocentrum nanum* and 99% for *Micromonas pusilla* (Table 6.2). Hidden layer nodes remaining after the OLS procedure (Chapter 2, Section 2.4.2.1) varied from 135 to 150. Differences were apparent between the identification of a species under PRiME 2 conditions, to that of PRiME 1. For example, *Prorocentrum nanum* was identified with 56.4% success as a PRiME 1 species, dropping to 42.1% when cultured under the illumination conditions of PRiME 2. Conversely, *Hemiselmis brunnescens* improved from 65% success under PRIME 1 conditions, to 94% success under PRiME 2. Assessing a network trained on PRiME 2 data with a PRiME 1 test file, produced an overall identification value of 40.9% and a confidence of 41.2%. The reverse produced equally poor results of 40.8% correct identification, with a confidence of 40.7%. Overall network performance when species from both culturing conditions were combined into individual classes was 69.1%, with confidence of identification at 68.3%. While treating each species from PRiME 1 as a separate class to that of PRiME 2, producing 120 classes, gave overall identification success and confidence of 59.5% and 59.4% respectively.

Table 6.2 Comparison of individual identification of species by the optimum original multi-class network (78.4%) and the optimum multiple network architecture (77.8%) for the PRiME 2 data.

Taxonomic Group and Species Name	Original Multi-class		Multiple Network		Taxonomic Group and Species Name	Original Multi-class		Multiple Network	
	Corr	Conf	Corr	Conf		Corr	Conf	Corr	Conf
Cryptomonads					**Prymnesiomonads**				
Chroomonas sp.	86.2	80.9	85.0	81.0	*Chrysochromulina camella*	70.0	89.9	81.0	89.5
Chroomonas salina	76.2	85.8	70.5	83.9	*Chrysochromulina chiton*	86.2	74.4	85.0	69.8
Cryptomonas appendiculata	94.4	83.4	95.2	94.8	*Chrysochromulina cymbium*	57.2	56.2	61.0	53.4
Cryptomonas calceiformis	80.8	88.2	90.5	88.0	*Chrysochromulina polylepis*	61.0	58.2	56.0	61.2
Cryptomonas maculata	80.6	78.8	71.4	79.1	*Emiliania huxleyi 92*	70.4	76.9	67.5	69.5
Cryptomonas reticulata	78.8	83.8	76.0	83.5	*Emiliania huxleyi B11*	98.7	98.2	99.0	97.2
Cryptomonas rostrella	74.4	89.2	81.5	84.5	*Ochrosphaera neopolitana*	67.0	67.9	67.5	60.0
Hemiselmis brunnescens	94.0	93.6	95.0	90.9	*Pavlova lutheri*	61.8	66.7	48.5	62.4
Hemiselmis rufescens	92.0	91.5	92.5	79.1	*Phaeocystis pouchetii*	60.4	62.3	61.5	58.4
Hemiselmis virescens	93.2	67.9	93.0	71.0	*Pleurochrysis carterae*	92.4	98.7	95.5	96.5
Plagioselmis punctata	84.6	85.0	79.5	92.4	*Prymnesium parvum*	54.8	61.1	40.5	62.8
Rhodomonas sp.	82.8	81.8	84.0	79.2					
Average	**84.8**	**84.2**	**84.5**	**84.0**	**Average**	**70.9**	**73.7**	**69.4**	**71.0**
Flagellates					**Diatoms**				
Chlamydomonas reginae	77.4	64.6	76	62.3	*Amphora coffaeformis*	87.0	89.8	88.0	84.2
Chlorella salina	49.4	59.8	51.5	61.7	*Chaetoceros calcitrans*	91.0	76.5	89.5	80.6
Dunaliella minuta	91	80.4	89	91.8	*Phaeodactylum tricornutum*	96.0	85.1	96.0	81.2
Dunaliella primolecta	83.6	81.6	82	82.4	*Skeletonema costatum*	69.6	62.8	63.4	59.4
Dunaliella tertiolecta	86.6	87.8	80	89.4	*Thalassiosira weissflogii*	88.9	76.7	83.5	69.5
Micromonas pusilla	99	73.2	96.5	81.5					
Nephroselmis pyriformis	62.4	56.5	63	60.5	**Average**	**86.5**	**78.2**	**84.1**	**75.0**
Nephroselmis rotunda	65.2	60.8	65	60.5					
Ochromonas sp.	43.2	66.3	38.5	66.2	**Dinoflagellates**				
Pelagococcus subviridis	82.6	80	80	85.1	*Amphidinium carterae*	82.4	77.7	80.0	76.9
Porphyridium pupureum	99	97.4	97	99	*Aureodinium pigmentosum*	91.2	79.0	88.5	84.6
Pseudopedinella sp.	48.1	59.8	46.5	52.8	*Gymodinium micrum*	70.4	79.1	70.5	78.3
Pyramimonas grossii	75.8	78.8	75	68.8	*Gymodinium simplex*	76.4	63.7	73.5	66.8
Pyramimonas obovata	72	60	63	56.8	*Gymodinium veneficum*	85.2	80.6	81.2	79.6
Rhodella maculata	94	95.3	89.5	95.7	*Gymnodinium vitiligo*	93.8	84.9	92.5	85.8
Stichococcus bacillaris	65	75.1	75	69.6	*Gyrodinium aureolum*	79.4	95.4	78.0	95.4
Tetraselmis impellucida	85.8	96.6	85.5	92	*Heterocapsa triquetra*	85.4	85.4	84.0	89.8
Tetraselmis striata	64.2	75.5	63	64.3	*Prorocentrum balticum*	63.0	72.9	58.5	79.6
Tetraselmis suecica	89.4	85.8	97	82.2	*Prorocentrum micans*	83.2	73.2	82.5	58.3
Tetraselmis tetrathele	96	95	99	96.1	*Prorocentrum nanum*	42.1	64.0	58.2	62.0
Tetraselmis verrucosa	90.4	93	88.5	97.3					
Average	**77.1**	**77.3**	**76.2**	**77.0**	**Average**	**77.5**	**77.8**	**77.0**	**77.9**

6.2.2.2 Multiple network Architecture

Using 500 events per species for the class of interest and 500 events per species in the background class, for PRiME 2 data only, produced an overall identification success of 77.8% and a confidence of identification of 77.9%. As with the original multi-class architecture, identification success varied between species, with a minimum of 38.5% for *Ochromonas* sp. and a maximum of 99% success for *Emiliania huxleyi* B11 (Table 6.2).

Again variances in identification success were evident for individual species when cultured under different conditions. For example, *Hemiselmis brunnescens* improved 100% from 40.4% success under PRiME 1 conditions, to 95% success under PRiME 2 conditions. A network trained on PRiME 2 data presented with a PRiME 1 set, produced an overall identification and confidence of identification of 37.9% and 36.8% respectively. The reverse situation was even lower, at 23.6% correct identification and 25.4% confidence of identification. Varying the background class content in the single species training files, has negligible influence on the final identification of the decision RBF network. However, combining both culturing conditions for one species into a class of interest, reduced the identification success by approximately 10%. Training 120 single species networks, *i.e.* 60 species from PRiME 1 and 60 species from PRiME 2, produced an overall identification success of 58.2% and a confidence of 61.1%.

6.2.3 Discussion

Identification of the PRiME 2 data set only, through both architectures, were comparable throughout all experiments. Differences arose in certain species where one structure performed better than the other. However, overall identification of the species cultured under PRiME 2 conditions, was approximately the same as that of PRiME 1. The poor results produced from presenting a network trained on PRiME 2 data with a PRiME 1 data file, and vice versa, indicates the variation in characteristics of species when cultured under different illumination conditions. This variation is further supported by the difference in the individual identification of some species, when cultured under PRiME 1 conditions, to their identification when cultured under PRiME 2 conditions, implying some species have become more, or less, distinct. The extent of this distinction can be dependent upon any overlapping species that may, or may not, have also experienced a change in their flow cytometric signatures. The increase in identification success of

Hemiselmis brunnescens can be attributed primarily to a variation in its orange fluorescence (phycoerythrin content) when cultured under a different illumination intensity. Under the initial culturing conditions (PRiME 1) the distributions of orange fluorescence for *Hemiselmis brunnescens* and *Hemiselmis rufescens* are very similar (Fig. 6.1a & c). When cultured under PRiME 2 conditions, a shift in the distribution of *Hemiselmis brunnescens* subsequently removes the peak from existing at the same position as that of *Hemiselmis rufescens* (Fig 6.1b & d), a species for whom orange fluorescence distribution is the same, regardless of illumination conditions. The mutual misidentity that previously existed between these two species is now reduced and both identify to >90%. This was also apparent within the *Gymnodinium* genus, where the species *Gymnodinium veneficum*, which previously identified to less than 40% has also doubled its distinction under PRiME 2 culturing conditions. Of course, this is a positive result due to a change in culturing conditions, other species exhibit a decrease in identification success due to a shift into another species' parameter range.

When treating each species cultured under different conditions as two separate classes, both architectures produce an overall identification of approximately 58%. Out of 120 classes, 22 identified to less than 40% correct, with only 11 exhibiting a misidentity with 'itself' under different culturing conditions. Misidentities were evident between groups, across both cultures. Although the overall performance seems low, this analysis involves a high number of classes. On closer investigation over half the species still identify to over 60% and naturally, when classes are added to any pattern recognition system, the overlap will increase and identification will inevitably reduce. Combining the events for a species from both data sets (*i.e.* PRiME 1 and PRiME 2) into one class improves performance in comparison to when the cultures are treated separately. Despite this appearing an acceptable approach, the data set now contains only 60 classes and in comparison to the identification of each PRiME set separately (62 classes in PRiME 1 and 60 in PRiME 2), performance has dropped by approximately 10%. Grouping those 11 species which misidentify with themselves across the two sets will improve results. However, not all species will have similarities across varying culturing conditions, and combining them simply because they are the same will prolong the training process and restrict the network from reaching a global minimum error.

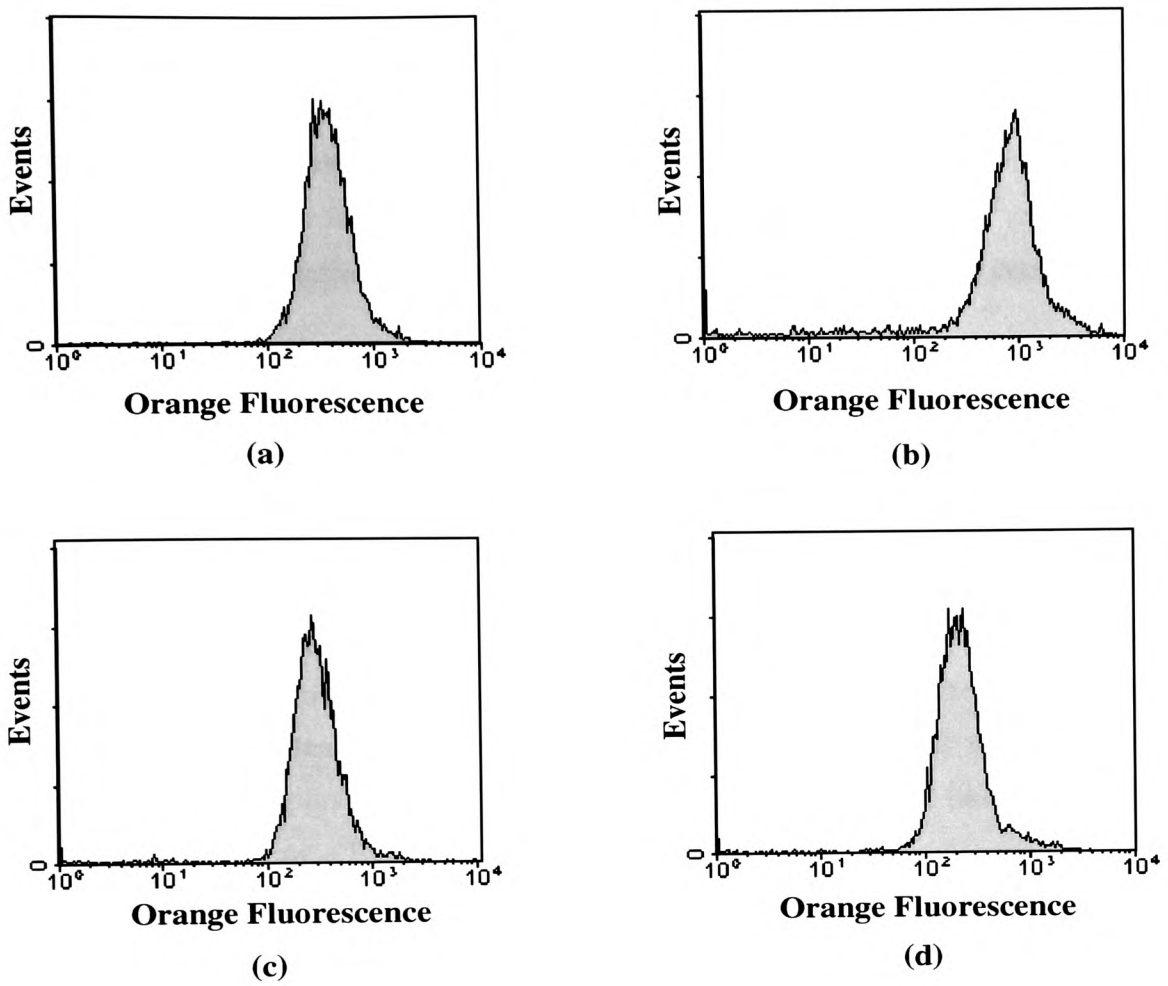


Figure 6.1 Orange Fluorescence distribution for (a) *Hemiselmis brunnescens* under PRiME 1 conditions, (b) *Hemiselmis brunnescens* under PRiME 2 conditions, (c) *Hemiselmis rufescens* under PRiME 1 conditions and (d) *Hemiselmis rufescens* under PRiME 2 conditions.

6.3 Laboratory Cultured Mixtures

To allow comparison between the two network architectures when determining the proportion of species in an unlabelled sample, seven mixtures were constructed, by Plymouth Marine Laboratories, from 27 separately cultured species (Table 6.3). With the exception of an increase in illumination intensity to 100μmol quanta $m^{-2}s^{-1}$, the culturing conditions for both the training and testing data were the same as that of PRiME 1.

6.3.1 Experimental Procedure

A 27 class data set containing 500 events per species, was used to train an original multi-class network, using 6 hidden layer nodes per class employing a Mahalanobis distance metric.

27 files were created to train 27 single species networks. Each file contained 500 events for the class of interest and 500 events per species in the background class.

Networks were trained using 10 nodes placed randomly within the class of interest, all employing a Euclidean distance metric. A 27 input parameter RBF decision network was trained on the outputs from the 27 single species networks, using 3 hidden layer nodes per class and employing a Euclidean distance metric.

As well as the supervised networks, an unsupervised 26 by 26 Kohonen map was trained, using Kohonen's recommendations (Appendix 2), to discover the natural classification of the 27 species. To provide further information regarding species overlap, a threshold of 0.07 was imposed on the boundaries between allocated nodes, which ranged between 0.014 and 0.413 Euclidean metrics (Fig. 6.2).

All networks were trained three times from different initialisation points and tested using independent test sets.

The research undertaken for this thesis was run in parallel to studies performed on the same data using Wavelet Analysis (Cohen & Kovacevic, 1996; Vidakovic, 1999). In brief, wavelets are mathematical functions that partition data into varying frequency components, which are subsequently studied with a resolution matched to its scale. Final results of the wavelet analysis for the laboratory grown mixtures were available for comparison (Collins, 2000).

Table 6.3 27 separately cultured species used to train neural networks for identification of the 7 constructed mixes. References depict species label on the Kohonen Grid (Fig. 6.2) and subsequent charts (Fig. 6.3).

Taxonomic Group	Species Name	Order	Size (μm)	Chart Ref.	Kohonen Ref.
Cryptomonad	*Cryptomonas rostrella*	Cryptomonadida	16-25	F	6
	Hemiselmis rufescens	"	4-9	N	14
	Rhodomonas sp.	"	8-13	X	24
Diatom	*Phaeodactylum tricornutum*	Bacillariophyceae	8-35	T	20
	Skeletonema costatum	"	3-5	Y	25
	Thalassiosira weissflogii	"	12-20	AA	27
Dinoflagellate	*Amphidinium carterae*	Dinoflagellida	15-20	A	1
	Aureodinium pigmentosum	"	7-12	B	2
	Gymnodinium micrum	"	8-15	I	9
	Gymnodinium simplex	"	6-10	J	10
	Gymnodinium veneficum	"	9-16	K	11
	Gymnodinium vitiligo	"	7-22	L	12
	Gyrodinium aureolum	"	35-45	M	13
	Heterocapsa triquetra	"	15-27	O	15
	Prorocentrum balticum	"	9-15	R	18
	Prorocentrum micans	"	30-40	U	21
	Prorocentrum nanum	"	8-10	V	22
Flagellate	*Chlorella salina*	Volvocida	4-8	C	3
	Dunaliella minuta	"	3-12	G	7
	Rhodella maculata	Rhodomonadida	7-24	W	23
	Tetraselmis tetrathele	Prasinomonadida	10-16	Z	26
Prymnesiomonad	*Chrysochromulina chiton*	Prymnesiida	5-9	D	4
	Chrysochromulina polylepis	"	6-8	E	5
	Emiliania huxleyi 92	"	5-6	H	8
	Isochrysis galbana	"	4-8	P	16
	Ochrosphaera neopolitana	"	8-10	Q	17
	Phaeocystis pouchetii	"	3-6	S	19

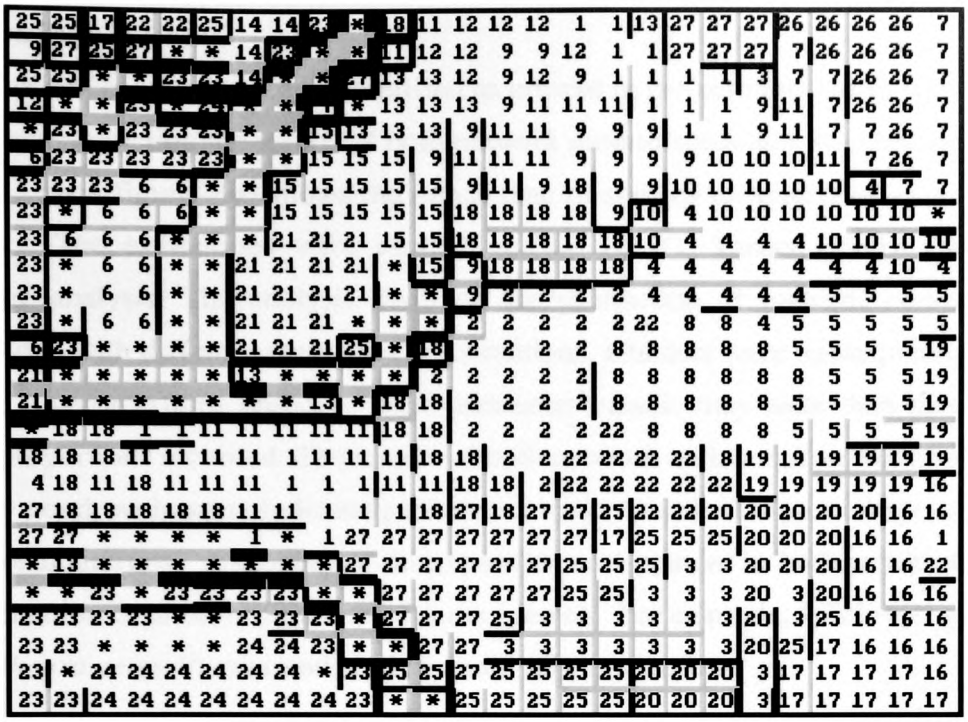


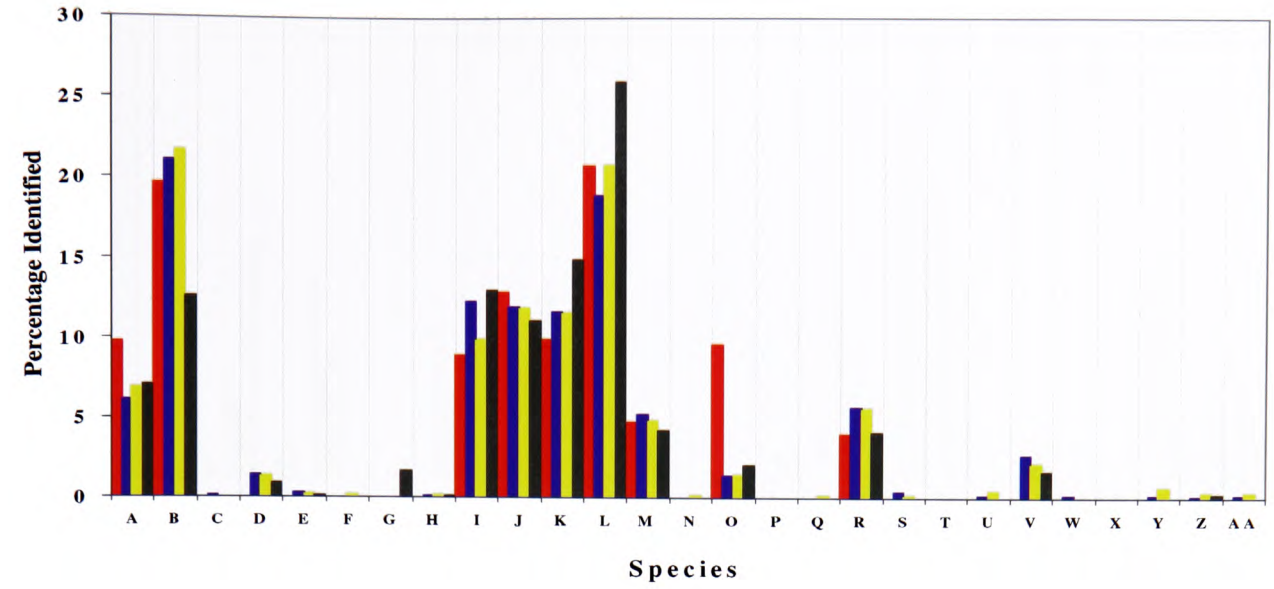
Figure 6.2 Kohonen grid produced from the 27 group data set showing borders above the threshold value (0.07) between classes allocated to different nodes in black, and between same class allocated nodes in grey.

6.3.2 Results

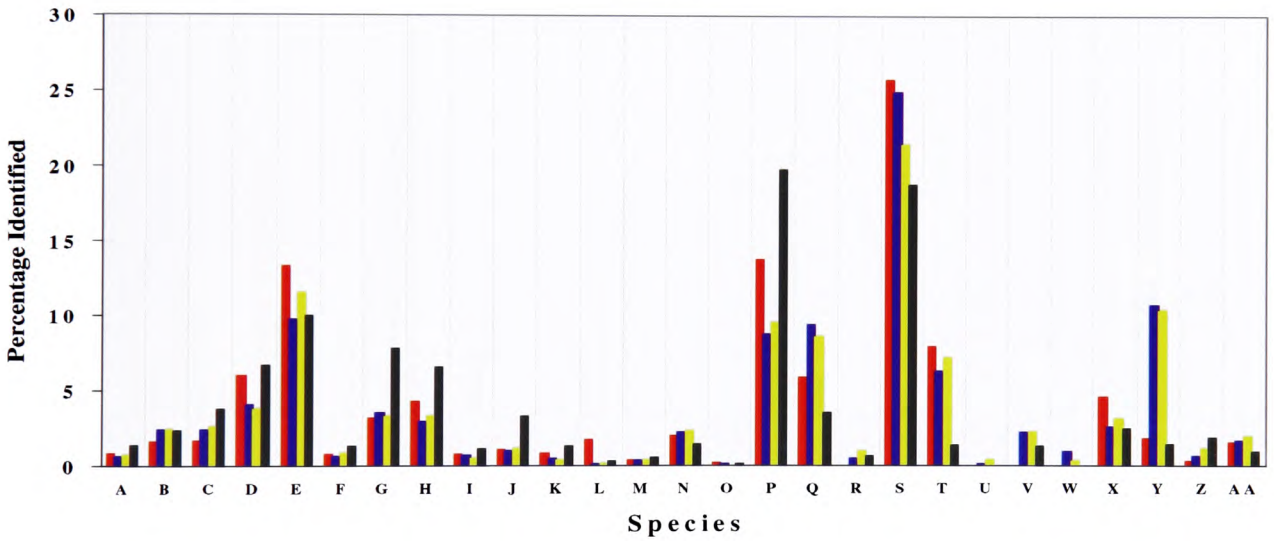
The percentage of species identified as present in the seven mixtures were compared using figures depicting the results of both network architectures and wavelet analysis (Figs. 6.3a - g). Each figure also shows the percentage of species estimated to be present in the mixtures. The estimations were carried out by Dr. G. Tarran (PML) through flow cytometric analysis. The analysis involved the production of multiple two-dimensional scatter plots, with different parameter combinations. Clusters were subsequently identified from the seven mixtures based on their parameter values, they were then gated, counted and the proportions recorded. The results of both network architectures, for all mixtures are generally good, with many indicating mixture content to a similar percentage as that of the proportional analysis. A number of discrepancies are apparent in both architectures, either over-estimation or under-estimation of some species. For example, *Isochrysis galbana* and *Ochrosphaera neopolitana* in mixes 2, 6 and 7 (Fig's.6.3b, f and g).

6.3.3 Discussion

After further investigation of the misidentification matrices, produced by both architectures (not shown), it was evident that the species for whom network assumptions were less accurate than the gated analysis, were those exhibiting low individual identification success and high misidentity with other species. For example, *Chrysochromulina chiton* is underestimated in mix 4 and 7. This species misidentifies with *Prorocentrum nanum*, which in both cases is determined present when not in the mixture. *Skeletonema costatum* is constantly overestimated in each of the mixes and is one of the less well identified species, at approximately 73% correct with a confidence of approximately 71%, by both architectures. The Kohonen map shows this species as being scattered amongst nodes that are not all within close proximity of each other. This indicates similarities with a number of other species and considerable variation within the particular strain itself. When *Isochrysis galbana* and *Ochrosphaera neopolitana* are present together, the latter is overestimated while the former underestimated. These two species are mutually misidentified by the networks and have a common location on the Kohonen map. The overlaps and misidentities of species, results in the overestimation of one at the expense of another. This again indicates that combining those species for whom overlap is high and consistent, must be a consideration in order to improve performance and definitive analysis.



(a)



(b)

Figure 6.3 Percentage of individual species assumed present in each mixture by the varying methods of analysis. ■ Gated data (Tarran, G), ■ Original multi-class architecture, ■ Multiple network Architecture, ■ Wavelet analysis (Collins, 2000) (a) Mix 1, (b) Mix 2.

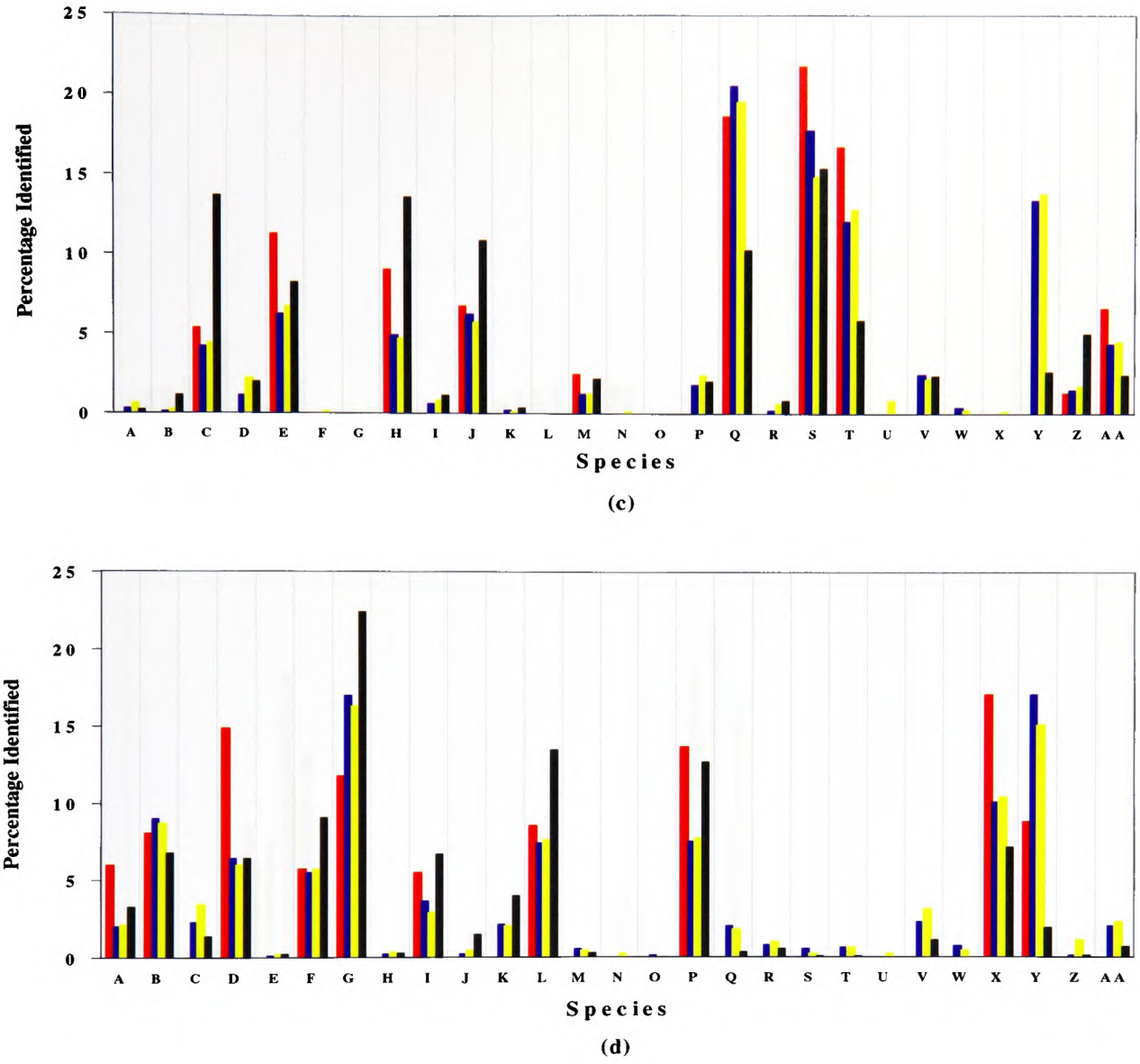
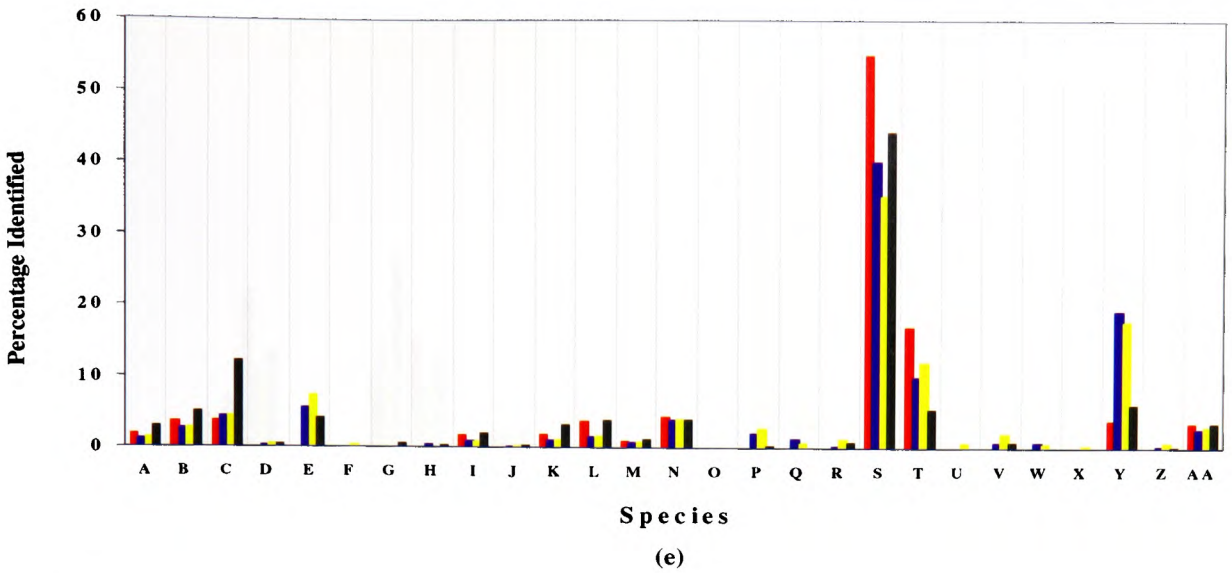
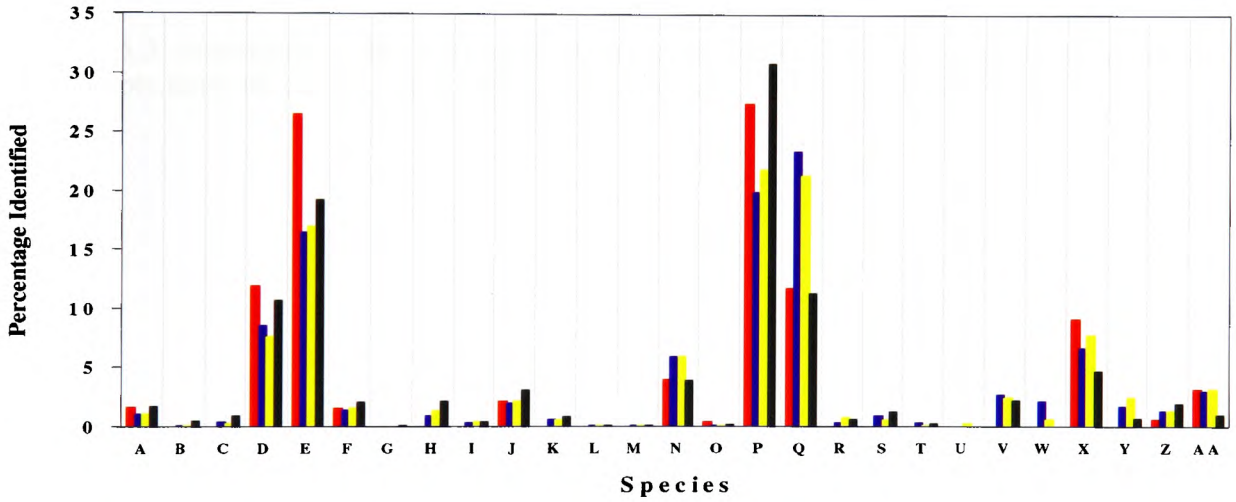


Figure 6.3 continued... Gated data (Tarran, G), Original multi-class architecture, Multiple network Architecture, Wavelet analysis (Collins, 2000) (c) Mix 3, (d) Mix 4.



(e)



(f)

Figure 6.3 continued.... ■ Gated data (Tarran, G), ■ Original multi-class architecture, ■ Multiple network Architecture, ■ Wavelet analysis (Collins, 2000) (e) Mix 5, (f) Mix 6.

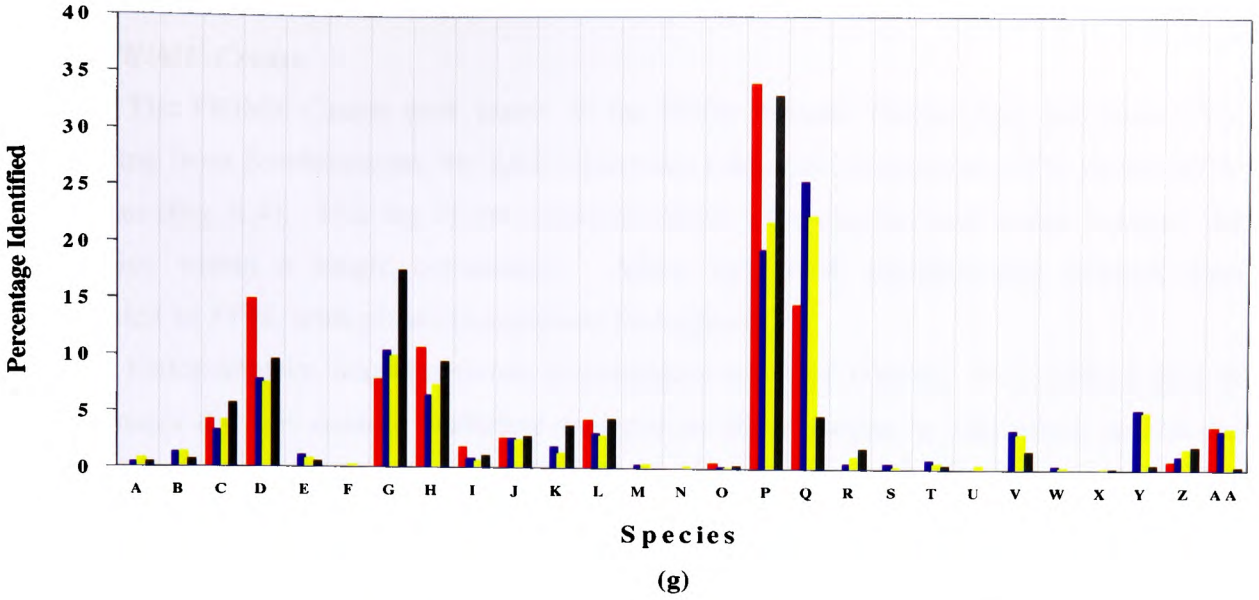


Figure 6.3 continued.... ■ Gated data (Tarran, G), ■ Original multi-class architecture, ■ Multiple network Architecture, ■ Wavelet analysis (Collins, 2000) (g) Mix 7.

6.4 Field Tests

6.4.1 PRiME Cruise

The PRiME Cruise took place in the North Atlantic during June and July 1996. Departing from Southampton, the RRS Discovery followed a transect to 59°N on the 20°W meridian (Fig. 6.4). This leg of the cruise involved a Lagrangian time series study of the dynamics within a single community. Along the 20°W meridian the transect then proceeded to 37°N, with plankton sampling throughout.

Unfortunately, neural network investigation was very limited. The southern part of the transect did not contain sufficient material to allow sorting to take place and so no analysis was performed. The remainder of the transect had abundant phytoplankton, but the species present were primarily *Synechococcus* sp., *Prochlorococcus* sp. and picoeucaryotes, none of which were present in the lab grown cultures and therefore not available for training or testing the networks ability to generalise. Other species were assumed to be present in the area by Dr. G. Tarran, and therefore, based on a brief microscopic analysis of gravity sampled seawater, 46 out of the 62 species database were chosen to train networks. This selection excluded the Diatoms and a number of species from the remaining 4 groups. As confirmation of the few phytoplankton species present was unavailable, a summary of the findings is presented for two samples, using the original multi-class network architecture to identify them. Consequently, the following comparisons were based on assumptions made by Dr. G. Tarran (PML) through two-dimensional scatter plot observations.

6.4.1.1 Procedures and Results

From a selection of samples taken at depths of 2m, 10m, 20m and 30m, a cluster of data was identified, through its distinct side scatter and red fluorescence, as *Coccolithus pelagicus* (Fig. 6.5). It was gated from the samples and used to create a new class, which was subsequently added to the training file, comprising now 47 species. Original multi-class RBF networks were trained and tested as described in Chapter 2 (Section 2.7.3) and an optimum chosen. Identification of *Coccolithus pelagicus* by the RBF network was extremely high due to its obvious separability.

The four samples taken for analysis were chosen for two reasons (Fig. 6.6 & 6.7). Firstly, the two-dimensional positioning and distinction of the clusters gave a presumed

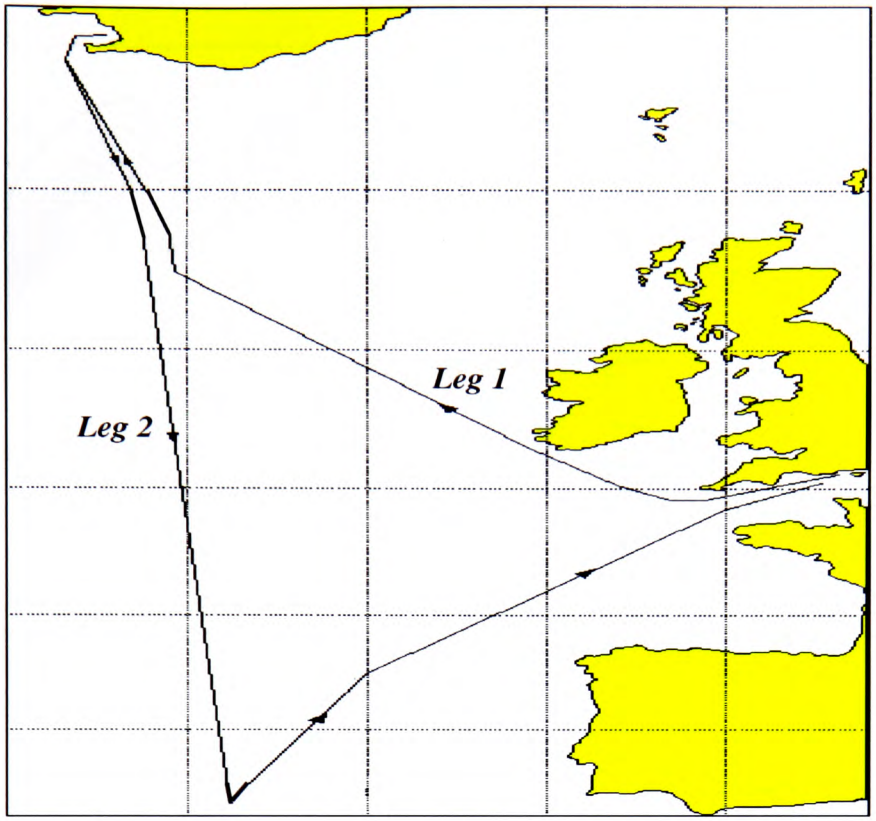


Figure 6.4 Transect of PRiME cruise from Plymouth to 59°N on the 20°W meridian (Leg 1) then South to 37°N (Leg 2).

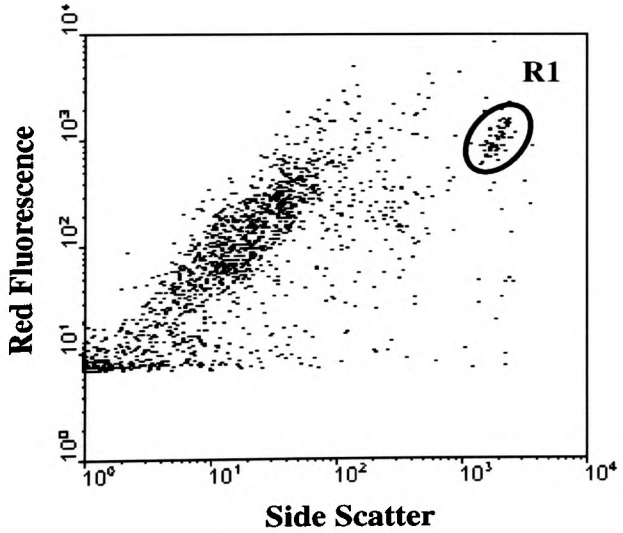


Figure 6.5 Region assumed to be *Coccolithus pelagicus*. The data cluster was subsequently gated and added to the species database.

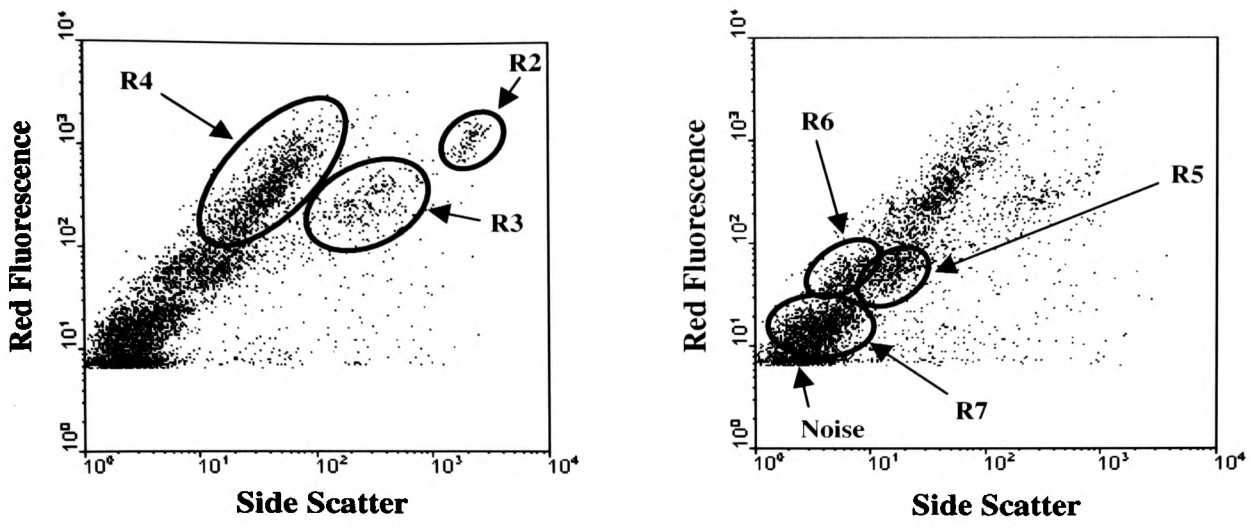


Figure 6.6 Scatter plots depicting red fluorescence against side scatter for sample 1. Regions R2 to R7 indicate data clusters of which probable identity was determined through these scatter plots.

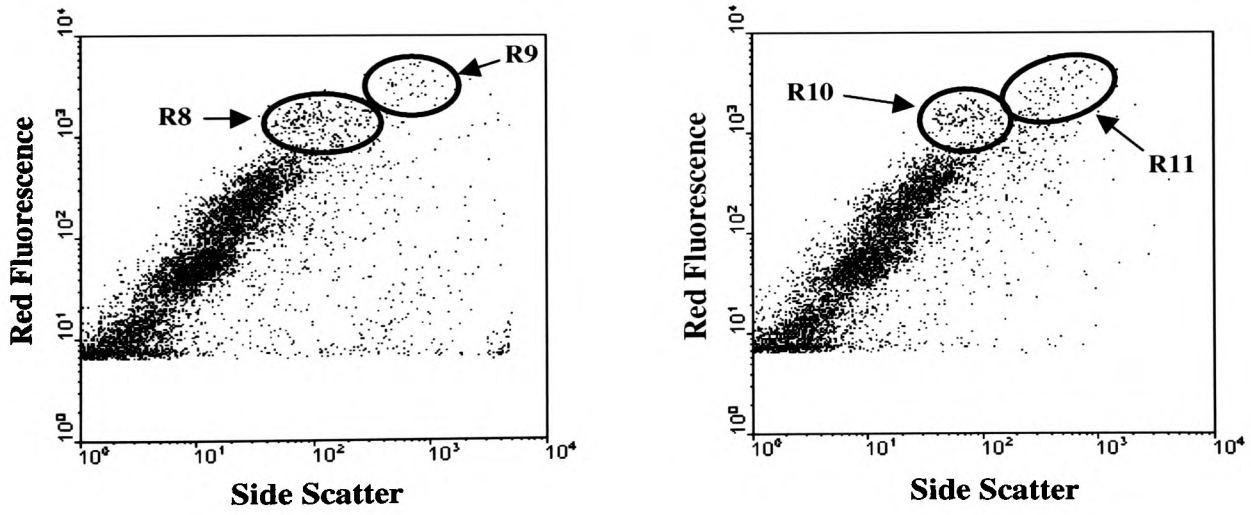


Figure 6.7 Scatter plots depicting red fluorescence against side scatter for sample 2. Regions R8 to R11 indicate data clusters of which probable identity was determined through these scatter plots.

indication of what was present. Secondly, the species suspected to be present in the samples were held on the laboratory cultured database. From the first two samples, six clusters were gated for analysis and their probable identity determined from their two-dimensional scatter plots (by Dr. G. Tarran) (Fig. 6.6). The first region identified (R2) was confirmed by the networks to be *Coccolithus pelagicus*. Although unseen data from this species identified extremely well (98%), it is with the knowledge that not only are the parameters distinct, but it is the only network class, within the database, trained on natural sea cultures. Region 3 was determined to be small coccolithophores (by Dr. G Tarran). The networks provided some confirmation of this, identifying region 3 as 68% *Emiliania huxleyi,* and a small percentage of large Coccolithophores and various others. Region 7 was assumed to be Flagellates within the size range 1-4μm, of which the network identified 98% as the species *Micromonas pusilla* (1-3μm). The remaining regions were suspected to be larger Flagellates, primarily identified by the network as *Stichococcus bacillaris,* with a mixture of others.

Dinoflagellates were suspected to be the main constituents of the four regions indicated in the second two samples (Fig. 6.7). The upper gate was assumed to be exclusively Dinoflagellates, indicated by the network to be primarily large Dinoflagellates and the lower gate both Dinoflagellates and some Flagellates.

Unfortunately, without definitive microscopic confirmation and no approximate proportions for comparisons, this field study provides limited information upon the network's ability to generalise. Particular clusters assumed to be certain species identify relatively well, but again these are distinct clusters that can be recognised visually by their scatter plots. For example, *Coccolithus pelagicus* and especially *Micromonas pusilla,* which has already shown to be a highly separable species.

6.4.2 Plymouth's Coastal Waters

With confirmation of content or quantity of the cruise field samples unavailable, the network analysis was inconclusive. Thus, further field studies were carried out at Plymouth's coastal areas. Varying depths and transects were chosen in order to study both the vertical and horizontal spatial structure of phytoplankton communities in late summer. To assess changes in composition the process was repeated over two days.

6.4.2.1 Experimental Procedure

Identical field sampling was carried out at mid-morning on both days, by Dr. G. Tarran aboard the RV Squilla (PML). The initial location was at a site SW of Plymouth (Station 1), where four samples were taken at depths of 9m, 13m, 30m and 43m. Four further sites were sampled at a depth of 2m along a transect between Station 1 and Plymouth (Table 6.4). 2 litre acid rinsed polycarbonate bottles were used to hold the samples in a refrigerator until analysed. The bottles were gently inverted several times and a 2ml sub-sample removed with a 1-5ml finnipipette and placed in a polystyrene sample tube. Samples were then analysed by the flow cytometer. For all samples 30ml of seawater was preserved with 400µl of hexamine buffered formaldehyde (=0.5% final concentration) and stored in the refrigerator for subsequent microscopic analysis.

Networks for both architectures were trained on 35 probable species from PRiME 1, suggested by Dr. G. Tarran to be present in the area. These included the species *Micromonas pusilla*, the genera *Hemiselmis* and *Emiliania*, and large and small Dinoflagellates. The original multi-class architecture used 500 events and 6 hidden layer nodes per class, employing a Mahalanobis distance metric. For the multiple network approach, the results from the appropriate 35 single species networks from Chapter 4, were combined to produce a 35 parameter input file to train a decision RBF network (Chapter 4, Section 4.4).

Unlabelled data files were constructed from the seawater samples at each station. Network analysis for the multi-class architecture was performed using the unlabelled file to produce an evaluation of sample content. For the multiple network approach each of the unlabelled data files were presented to the 35 trained single species networks, and the outputs combined to form a test file for the trained RBF decision network.

In order to achieve an improved approximation of sample content, disregarding ambiguous or borderline patterns, a threshold of 0.725 was imposed upon the hidden layer nodes of the original multi-class network, and a threshold of 0.4 upon the output layer nodes of the RBF decision network.

The values produced for comparison to the network's approximations, are achieved through the same process of gating and counting as performed for the laboratory grown mixtures (Section 6.3). For example, Figure 6.8 shows two scatter plots for a sample taken at station 1 on the second day, at a depth of 30m. Regions R1 and R2 are assumed to be

Table 6.4 Stations locations and depths for field sample analysis around Plymouth's coastal areas. Identical sampling was repeated at mid morning on both days by Dr. G. Tarran on board the RV Squilla (PML).

Station	Latitude	Longitude	Depth (m)
1a	50°15.08' N	4°12.55' W	9
1b	50°15.08' N	4°12.55' W	13
1c	50°15.08' N	4°12.55' W	30
1d	50°15.08' N	4°12.55' W	43
2	50°16.27' N	4°11.9' W	2
3	50°17.18' N	4°9.74' W	2
4	50°18.13' N	4°8.51' W	2
5	50°20.55' N	4°8.18' W	2

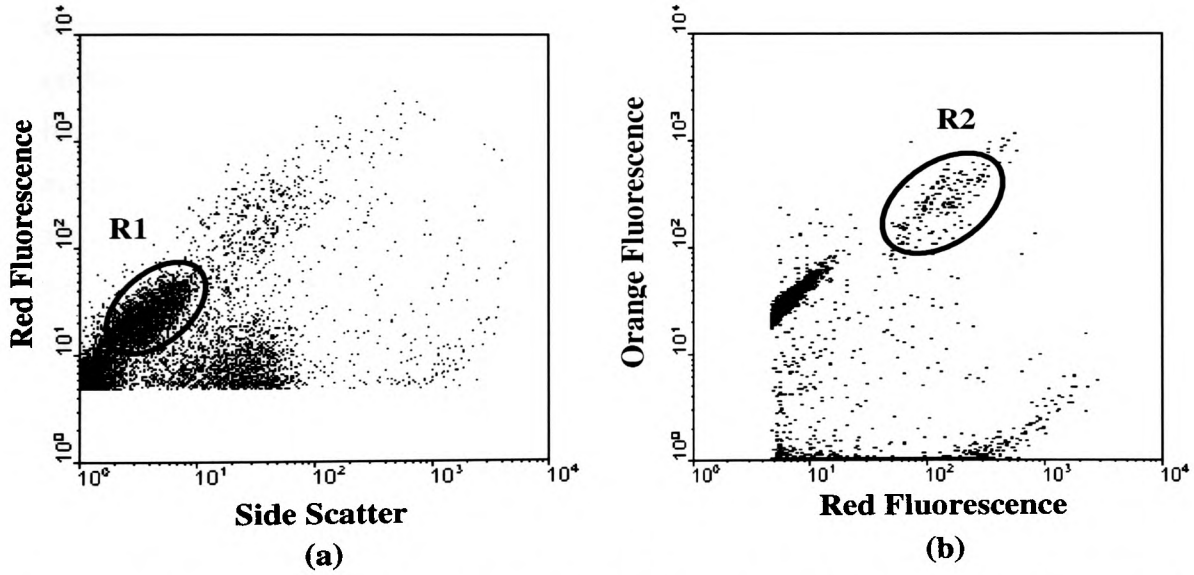


Figure 6.8 Two-dimensional scatter plots of the sample taken at station 1 at a depth of 30m on the second day. Marked regions show clusters gated within the sample assumed to be (a) *Micromonas pusilla*, R1 and (b) *Hemiselmis virescens*, R2.

Micromonas pusilla and *Hemiselmis rufescens* respectively. These two clusters were therefore gated, sorted and counted.

6.4.2.2 Results

The results of the field samplings are compared using line charts depicting the percentage of each species determined present at the various stations (Figs. 6.9a-e, 6.10a-e & 6.11a-e).

6.4.2.3 Discussion

Vertical and horizontal spatial variation of all species is within 5%, with the exception of *Micromonas pusilla* at station 1, where a reduction in the species is evident. This species is consistently the dominant component of the area across all depths and stations. As with all experiments, identification of this species has always been high. The primary results of this chapter (Section 6.2) have shown that altering the illumination value when culturing *Micromonas pusilla* has negligible variance on any of its optical parameters (Fig. 6.12 & 6.13). Although in a natural environment this will not be the only condition influencing cell formation and structure, the small size and distinct optical characteristics implies identification of this species will always be high. With the exception of the small dinoflagellates and *Micromonas pusilla* on day 2, all methods indicate less than 5% content of *Emiliania huxleyi, Hemiselmis virescens* and large Dinoflagellates. This low quantity would make separation by sorting more crucial, where missed or added species will have a large impact on percentage determination.

Identification by both network architectures are fairly consistent, with each identifying to within 2% of the other. The variation is expected, as it has already been shown that individual identification for particular species varies, depending on the architecture used. Excluding *Micromonas pusilla,* the results show species content on an exaggerated scale. This gives an impression of larger differences, between percentage determined by gating and percentage determined by the networks, than are actually present.

The species *Emilian huxleyi* and *Hemiselmis virescens* are both identified relatively well, despite the presence of only a small percentage of both. With the exception of the multiple network architecture without a threshold, the content of both is slightly

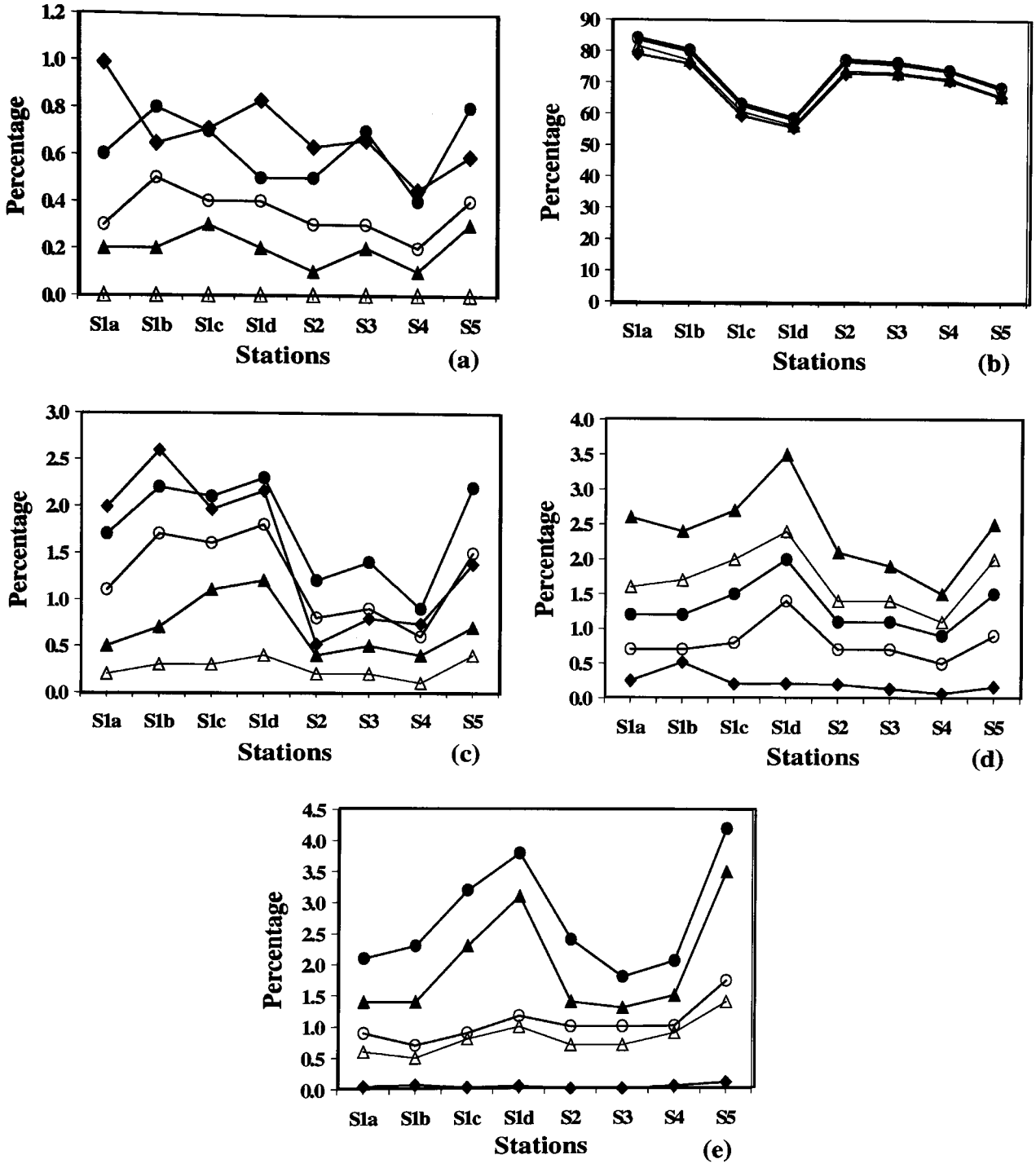


Figure 6.9 Percentage of species determined to be present in the field samples taken at the five stations on Day 1. n.b. axes scale varies between figures. ◆ Gating (Dr. G. Tarran), ▲ Original architecture, △ Original architecture with threshold, ● Multiple network architecture, ○ Multiple network architecture with threshold. (a) *Emiliania huxleyi*, (b) *Micromonas pusilla*, (c) *Hemiselmis virescens*, (d) Small Dinoflagellates (<20 μm), (e) Large Dinoflagellates (>20μm) .

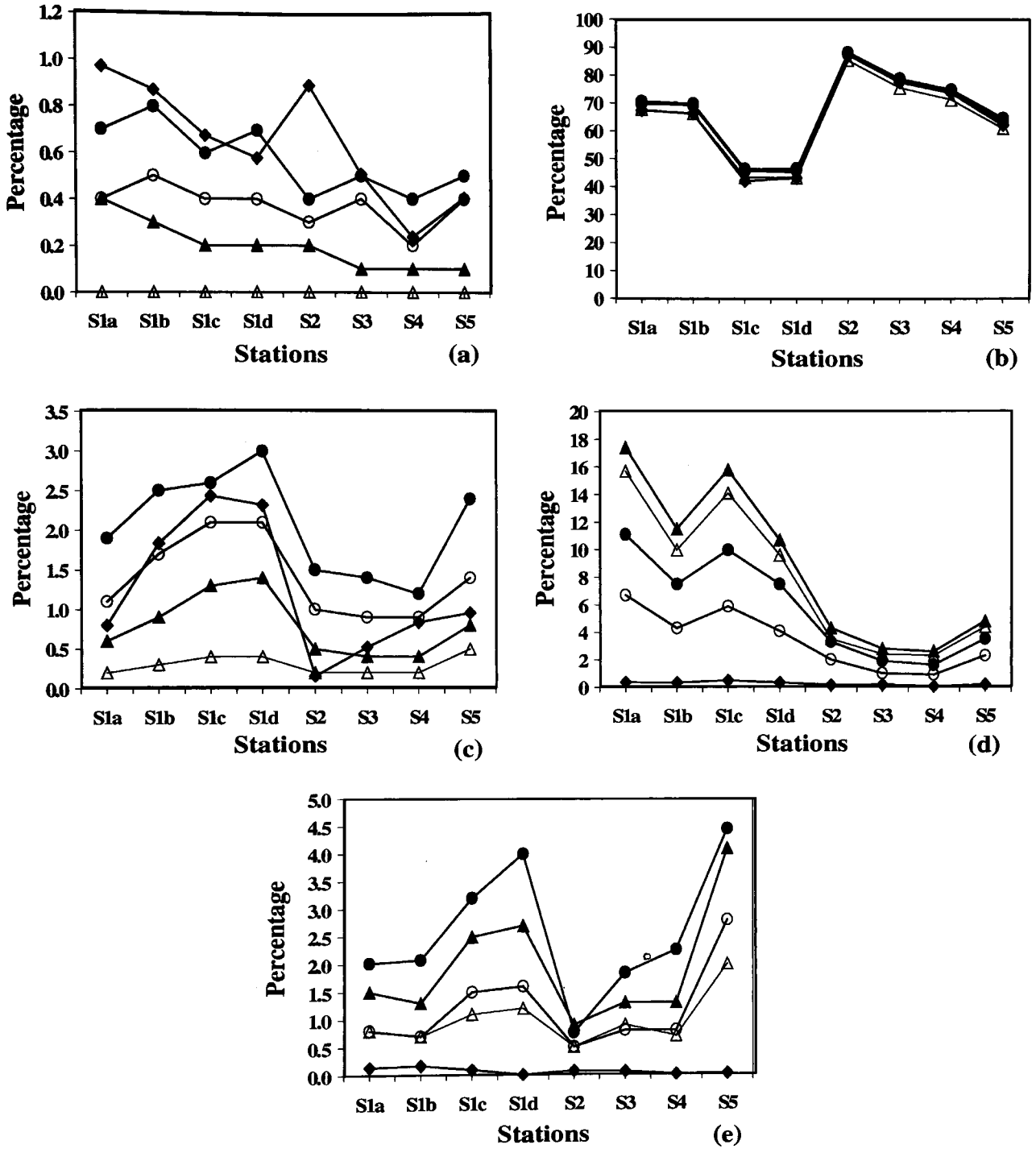


Figure 6.10 Percentage of species determined to be present in the field samples taken at the five stations on Day 2. n.b. axes scale varies between figures. ◆ Gating (Dr. G. Tarran), ▲ Original architecture, △ Original architecture with threshold, ● Multiple network architecture, ○ Multiple network architecture with threshold. (a) *Emiliania huxleyi*, (b) *Micromonas pusilla*, (c) *Hemiselmis virescens*, (d) Small Dinoflagellates (<20 μm), (e) Large Dinoflagellates (>20μm) .

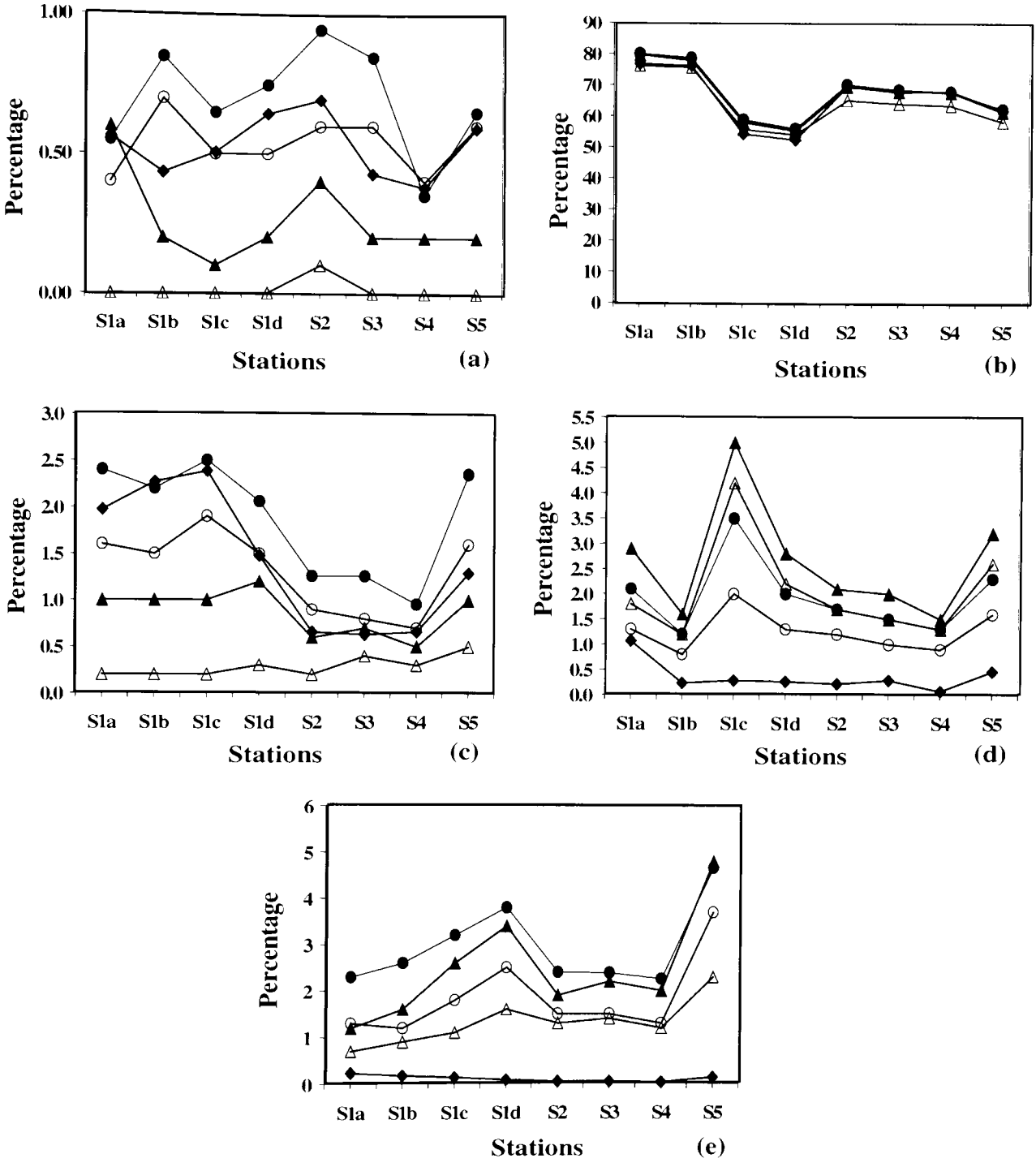


Figure 6.11 Percentage of species determined to be present in the Day 1 field samples after 2 days incubation in the refrigerator. n.b. axes scale varies between figures. ◆ Gating (Dr. G. Tarran), ▲ Original architecture, △ Original architecture with threshold, ● Multiple network architecture, ○ Multiple network architecture with threshold. (a) *Emiliania huxleyi*, (b) *Micromonas pusilla*, (c) *Hemiselmis virescens*, (d) Small Dinoflagellates (<20 μm), (e) Large Dinoflagellates (>20μm).

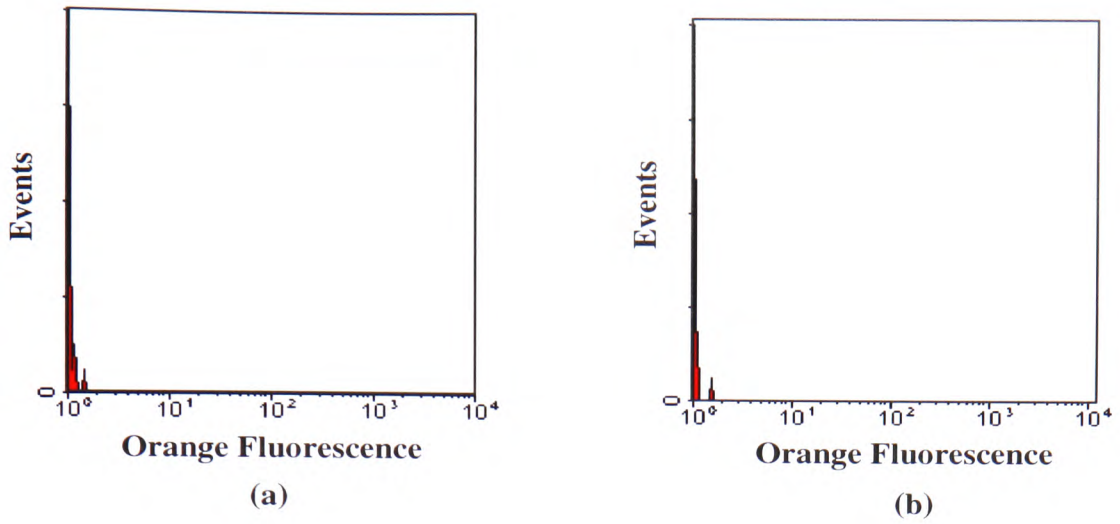


Figure 6.12 Orange fluorescence distribution against event number for *Micromonas pusilla* cultured under (a) PRiME 1 conditions and (b) PRiME 2 conditions. The distinct low phycoerythrin content in this species, is not affected by a change in illumination conditions and may account for its high identification.

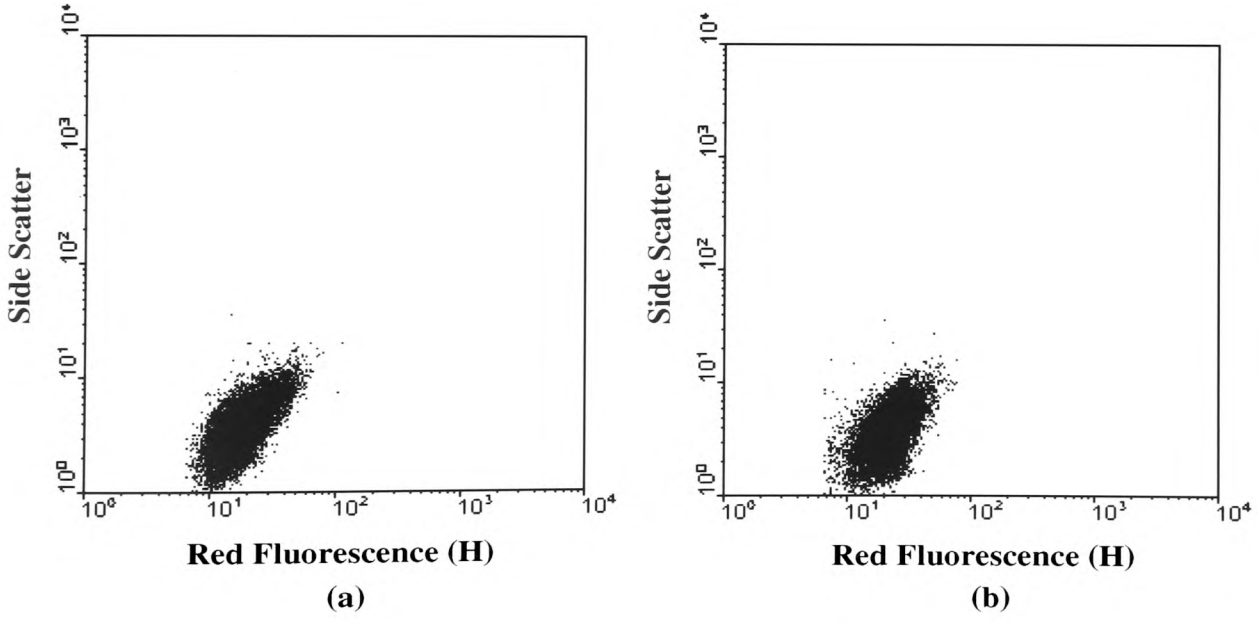


Figure 6.13 Two-dimensional scatter plot of optical flow cytometric parameters for *Micromonas pusilla* cultured under (a) PRiME 1 conditions and (b) PRiME 2 conditions, depicting little variation.

underestimated. The opposite is true in the case of small and large dinoflagellates, where overestimation by both architectures is evident. Naturally with no threshold imposed, all cells will be forced into a class, identifying them as one species or another. Introducing a threshold reduces misidentification, where species with low hidden layer node outputs that were allocated as possible Dinoflagellates were rejected. Increasing this threshold would bring the identification closer to that assumed, but may further the underestimation of *Emiliania huxleyi* and *Hemiselmis virescens*.

Identifying individual species provides less ambiguity and more information on known overlaps than assuming the existence of a size range of one group. This offers more explanation when considering those species known to constantly misidentify and overlap, indicating again the improvements that can be achieved through combinations of some species. With specific knowledge of the Dinoflagellate composition and any remaining species in the sample, more explanation could be offered to the overestimation of the group, where overlap with unnamed species may have introduced discrepancies. In all cases a percentage of Cryptomonads and Flagellates were also identified, but unfortunately no gating indication was provided for comparison. Cyanobacteria (>1μm) are also assumed to be a component of the phytoplankton, distinguished by their small size. However, they are not held on the database and thus not of interest here.

Variation between day 1 and day 2 for most species is minimal. The small Dinoflagellates are an exception to this, where network determination of this class is much higher on day 2, for all depths at station 1. Although the architectures both overestimate this class, it is to within 1-3 % for day 1 increasing to between 11 and 17% for day 2. This may be attributed to incorrect gating percentages, or species being misidentified as small Dinoflagellates by the network. However, the latter appears less likely as threshold imposition does not reduce misidentities to a level comparable to that of day 1. A third possibility may be attributed to some environmental change, therefore reducing the network's identification as species sorted are now uncharacteristic of day 1 conditions, although this seems less likely.

An important consideration when attempting to identify phytoplankton, is the extent a population may alter during incubation periods. However, day 1 repetition implies only minor fluctuations in all results, again with the exception of small Dinoflagellates. Despite the variation between percentage determined through gating, and that determined

through network analysis, the trend of the results for both network architectures are similar, implying both are identifying to their own individual level of performance, for the particular species in question.

It must be reiterated that species grown in a natural environment may, in many cases, have very different characteristics to laboratory grown samples. Determination of sample content through scatter plots, makes the assumption that parameters are the same as that of species grown under known culturing conditions. The opposite to this has already been demonstrated for particular species when illumination is varied. Additionally, manual errors introduced when gating clusters assumed to represent species, may inevitably increase or decrease the sorted sample by hundreds or even thousands of cells, possibly introducing a large error margin. Thus, although this process gives an indication of what may be present in a sample, it cannot be taken as definitive with regards to absolute percentage, and requires microscope analysis for full confirmation, as of yet unavailable.

6.5 Conclusion

The primary objectives of the field and mix data experiments were, initially, that there was consistency between the identification success of both architectures. This has been demonstrated for field and laboratory cultured data, where overall performance is close, with slight variations in favour of one method over the other for particular species.

Secondly, the translation of the network's generalisation ability to field data is of great importance if the system is to work in a real time field situation. Unfortunately, with definite confirmation of the sample composition unavailable, it is not possible to draw any strong conclusions, or offer explicit reasons as to any discrepancies. Nevertheless, the networks have demonstrated relatively good interpolation and extrapolation abilities when considering variation in a species characteristics However, the results imply that a slight change in illumination alone can severely alter the optical characteristics of some species. Therefore, in order for an analysis system to be trained to its optimum level, there must be information available on actual field samples.

7 Synthesis

7.1 Introduction

Most areas of research involving neural network analysis of flow cytometric data, has centred around small numbers of species. An unrealistic assumption in the natural environment. The abundance and complexity of phytoplankton requires constant revision of multi-class automatic identification systems (in this instance ANNs). The work reported in this thesis introduces an alternative multiple network approach, more suitable to this variable community. Additionally, the research documented illustrates how an alternative structuring system could be initiated, based on flow cytometric characteristics, through boundary and cluster recognition on the SOM. This chapter summarises the conclusions drawn from this research and makes suggestions as to how the work may progress in the future.

7.2 Paradigm Selection

In a basic analysis of 12 species, the superiority of the RBF network over the MLP was established, the two main reasons for superiority being:

1. The distinction and overlap of flow cytometric data for phytoplankton cells varies considerably between species, conforming to linearly separable clusters only when the class number is small. Unlike the hyperplanes of the MLP, boundary formation by the RBF network constructs non-linear decision regions, making it more suitable for modelling the data distribution.
2. The infinite, linear decision boundaries of the MLP will inevitably assign a high level of identification to an unknown pattern. Rejection of ambiguous patterns by threshold imposition is consequently often low, reducing overall confidence of identification. Conversely, as the boundaries of an RBF are localised and finite, identification levels are dependent upon a pattern's proximity to the centre of the basis functions, therefore assigning smaller values to unknown patterns, resulting in their rejection.

7.3 Training Set Size

The difficulties in field data acquisition will inevitably result in an under-representation of some species. When analysis is required of these species against those

that are easily obtainable, the training set constructed may contain imbalanced event numbers. For this reason, the effect on performance when event numbers are varied was considered for both balanced and imbalanced data sets. The increase of data overlap that inevitably occurs as class numbers are added, will naturally have an affect on overall performance. The research has illustrated that a balanced set of events per class is preferable, however this is not always possible. When this is the case, the distribution of classes with minimal event numbers, representative of less discriminable species, will be obscured by the abundance of data and may not be adequately modelled by the network. Despite this, the identification of a distinct species is still high, even when event numbers are few. Providing the data is sufficient enough to ensure full representation of the biological variation, and not too few so as to cause memorisation, the data distribution of a distinct species can be efficiently modelled. However, with high numbers of classes and the heterogeneous nature of phytoplankton, such seperability is rarely evident, and balanced training sets should be employed if possible.

7.4 Multi-class RBF Network

The potential of the multi-class RBF network, was further realised in the high identification success of 62 species of phytoplankton. The process of scaling up class number becomes increasingly complex, requiring greater periods of training and optimisation. Although longer training times may not be a consideration, it is envisaged that optimisation procedures may become more problematic as class numbers increase.

Despite the performance of the network, its rigid architecture restricts its potential in a number of areas. With the number of phytoplankton species unbounded, the possibility of encountering a new one is inevitable. The multi-class network is unable to encompass this novel species without complete retraining, involving long optimisation procedures. If certain species are known not to inhabit a particular body of water they can be excluded from the analysis. This is not a simple process with the multi-class network, as again it requires retraining of a complex algorithm. It is these limitations, and the inflexible nature of the original multi-class network, that has initiated the need for the alternative multiple network approach.

7.5 Multiple Network Architecture

The multiple network architecture presents a flexible system of simple, single identification networks, each responsible for an individual species, culminating in a final decision process. The single species networks are very basic RBF architectures, which require no optimisation techniques. From the two procedures investigated for combining the outputs of the single species networks for final identification, the RBF decision network was the most appropriate. This process is independent of the architecture of the single species networks, and exhibits negligible affect on overall performance, even when the single species training files comprise a heavy imbalance in event numbers.

Despite the probability of a greater number of dimensions than the seven optical parameters, the distribution of the input data generated to train the RBF decision network, are far less complex. Optimisation procedures are at an absolute minimum, encompassing very short periods of training time and no necessary expertise. The completed technique provides a library of pre-trained single species networks, to which more can be easily added each time a new species is encountered. This subsequently allows users to dynamically select subsets of single species networks as and when required, as well as the option of increased parameters possibly adding discriminatory information. The multiple network approach performs as well as the original multi-class network, but with advantage of being an adaptable, simple and flexible structure.

7.6 Morphology versus Flow Cytometry Signatures

The migration of certain species away from their taxonomic group, has been illustrated through employment of the SOM. This demonstrates the difference between the criteria used for classical morphological groupings of some species, and their flow cytometric signatures. The performance of both the multi-class and multiple network architectures, trained on the complete database, has been improved by combining some species for whom overlap is heavy and consistent, where not all constructed classes contained species from the same group or genus and in many cases comprise a mix of taxa. The construction of network groups from the supervised dendrogram analysis, allows greater improvements in identification than those achieved from forcing morphometric groupings.

The lack of correlation between the morphological characteristics and the optical properties of some species, implies that the primary, and possibly secondary, classical divisions may not be the most suitable for this analysis. A more appropriate labelling system, representative of flow cytometric similarities, will improve identification by supervised networks and offer more explanation towards species overlap.

7.7 Boundary Recognition on the SOM

When presented with a data set, the SOM is capable of adapting itself to the distribution of the input data, thus preserving the relationship between multi-dimensional variables. While this network provides the basis for classification, distinction between probable clusters in the output space can be difficult.

The methods presented for boundary detection offer an alternative approach to classical clustering, by considering the hyper-dimensional distances between the position vectors of the Kohonen nodes, and not the data. This reduces the risk of ambiguous clusters that may be a result of irregularities in the data. Once node clusters have been determined, actual data classification can be achieved by grouping the particular data allocated an individual cluster of nodes. Although definitive clustering may not be obtainable, areas of correlation exist between the methods, providing insight into the natural overlaps and distinctions of phytoplankton species. Use of the approaches iteratively, to establish coarse clustering first and then discover finer clustering, will provide a starting point for a hierarchy of divisions based on flow cytometric similarities. In addition, the data used to evaluate clustering by the SOM is limited to seven dimensions. The procedures and therefore definition of clusters may be improved, if an increased number of parameters are employed for analysis.

As classification is a continual process, it is always possible to further partition or merge a data set, thus some degree of stopping criteria or threshold determination must be defined. This can take various forms, but increasing the number of user-defined assumptions increases the chance of cluster formations that are not reflective of the true data similarities. Although some degree of input is required for the methods presented, it is no more than other algorithms, and less *a priori* information is needed as the proximity of data distribution has already been established by the SOM. This preservation of the spatial distribution of data, will improve the similarities concluded by individual users

employing the boundary detection methods. This will alleviate discrepancies caused by inappropriate algorithm selection.

7.8 Culture Variation and Field Data

Phytoplankton data are inherently variable. The multi-modal distribution of a laboratory cultured sample indicates the presence of sub-populations. This is multiplied when the species are in a natural habitat, where cells exhibit considerable variation in their optical characteristics as a result of diverse environmental conditions and varying stages of growth.

Without definitive confirmation of sample content, the field trials comparison cannot be conclusive. Both network architectures have demonstrated moderate generalisation ability. However, the poor identification of Prime 2 data by a network trained on the Prime 1 data set (and vice versa), and the increase or decrease of some species identification under different culturing conditions, indicates the extent of biodiversity when illumination alone is altered. Therefore, to achieve maximum generalisation and produce a system capable of identifying natural cultures, training data must be available for both field samples, and laboratory grown species cultured under a greater range of conditions.

7.9 Conclusions and Future Work

This research has suggested and tested an alternative multiple neural network architecture, for identifying phytoplankton from flow cytometric data. With the advantages of the multi-class RBF, the multiple network allows easy addition of novel species and the rapid selection of subsets of networks. The lack of influence of the background class on the final identification by the decision RBF network, has been demonstrated for this data. Theoretically, a relationship between output ranges should exist, providing there is consistency between the structure of the single species training files and the architecture of the associated networks. However, although the number of species analysed here (74 in total) may be more representative of a field number than previous research, it is still small compared to that encountered in the natural environment. Therefore, to fully realise the potential of the approach, further studies need to be performed on both the architecture of the single species networks and the content and

quantity of the training data. Whilst no notable problems were encountered when identifying 74 species, as numbers increase the content of the background class may influence performance.

The initiation of an alternative structuring system, would provide a more appropriate class membership reflected by the flow cytometric similarities, thereby improving identification to group and genus level. This requires further study and correlation between neural network analysts and phytoplankton taxonomists. Information regarding optical similarities can be acquired, after clustering on the SOM, by extracting flow cytometric parameters of a set of clustered data. Its variance and distribution can be assessed and similarity measures established. This procedure can also be useful in establishing similarities in data collected from the field, where some insight could be provided of the influence of environmental conditions on flow cytometric variation.

Although groupings constructed from flow cytometric similarities will improve network performance, there are some species for which overlap is such that complete separation may never be possible. Further improvements can be achieved through the combination of some species, where incorrect identification of individual points, especially in the overlapping boundaries of similar species, reduces identification and confidence of identification. Conclusive identification can then be performed on a delimited set via alternative methods such as microscopy or increased optical parameters.

References

Al-Haddad, L., Morris, C.W. and Boddy, L. (2000) Training Radial Basis Function Neural Networks: Effects of Training Set Size and Imbalanced Training Sets. *Journal of Microbiological Methods*, 43, 33-44.

Anand, R., Mehrotra, K.G., Mohan, C.K. and Ranka, S. (1993) An Improved Algorithm for Neural Network Classification of Imbalanced Training Sets. *IEEE Transactions On Neural Networks*, 4, 6, 962-969.

Balfoort, H.W., Snoek, J., Smits, J.R.M., Breedveld, L.W., Hofstraat, J.W. and Ringelberg, J. (1992) Automatic Identification of Algae: Neural Network Analysis of Flow Cytometric Data. *Journal of Plankton Research*, 14, 4, 575-589.

Baum and Haussler (1989) What Size Net Gives Valid Generalization? *Neural Computation*, 1, 151-160

Bishop, C. (1994) Neural Networks and Their Applications. *Rev. Sci. Instrum*, 65, 6, 1803-1832.

Bishop, C (1995) *Neural Networks for Pattern Recognition*. Oxford University Press Inc., New York.

Blackmore, J. and Miikkulainen, R. (1993) Incremental Grid Growing: Encoding High-Dimensional Structure Into A Two-Dimensional Feature Map. In *Proceedings, IEEE International Conference On Neural Networks*, 450-455, IEEE Press, Piscataway.

Boddy, L., Morris, C.W., Wilkins, M.F., Tarran, G.A., and Burkill, P.H. (1994a) Neural Network Analysis of Flow Cytometric Data for 40 Marine Phytoplankton Species. *Cytometry*, 15, 283-293.

Boddy, L., Morgan Gimblett, A., Morris, C.W. and Mordue, J.E.M. (1994b) Neural Network Analysis of Fungal Spore Morphometric Data for Identification of Species in the Genus *Pestalotiopsis*. In: *Intelligent Engineering Systems Through Artificial Neural Networks*, Dagli, C.H., Fernandez, B.R., Ghosh, J. and Kumara, R.T.S. (eds.) 4, 605-612. ASME Press, New York,.

Boddy L., Morris, C.W., Wilkins M.F., Al-Haddad L., Tarran G.A., Jonker R.R., and Burkill, P.H. (2000) Identification of 72 Phytoplankton Species By Radial Basis Function Neural Network Analysis of Flow Cytometric Data. *Marine Ecology Progress Series* 195, 47-59.

Boney, A.D. (1989) *Phytoplankton*. New Studies in Biology. Edward Arnold (Publishers) Ltd., London.

Broomhead, D.S. and Lowe, D (1988) Multivariate Functional Interpolation and Adaptive Networks. *Complex Systems*, 2, 321-355.

Burkill,P.H. (1987) Analytical Flow Cytometry and Its Application to Marine Microbial Ecology. In: *Microbes In The Sea*, Sleigh, M.A. (ed.) 139-166. Ellis Horwood, Chichester.

Burkill,P.H. and Mantoura, R.F.C. (1990) The Rapid Analysis of Single Marine Cells By Flow Cytometry. *Phil. Trans. Royal Soc. London*, 333, 99-112.

Carpenter, G.A. and Grossberg, S (1987a) A Massively Parallel Architecture for A Self-Organising Neural Pattern Recognition Machine. *Comput. Vision Graphics Image Process,* 37, 54-115.

Carpenter, G.A. and Grossberg, S. (1987b) ART 2: Self-Organization of Stable Category Recognition Codes for Analog Input Patterns. *Applied Optics*, 26, 23, 4919-4930.

Carr, M.R., Tarran, G.A. and Burkill, P.H. (1996) Discrimination of Marine Phytoplankton Species Through The Statistical Analysis of Their Flow Cytometric Signatures. *Journal of Plankton Research*, 18, 1225-1238.

Chandrasekaran, B. and Jain, A.K. (1975) Independence, Measurement Complexity, and Classification Performance. *IEEE Transactions On Systems, Man and Cybernetics*, 5, 2, 240-244.

Chandrasekaran, B. and Jain, A.K. (1977) Independence, Measurement Complexity, and Classification Performance: An Emendation *IEEE Transactions On Systems, Man and Cybernetics,* 7, 564-566

Chen, S., Cowan, C.F.N and Grant, P.M. (1991) Orthogonal Least Squares Learning Algorithm for Radial Basis Function Networks. *IEEE Transactions On Neural Networks*, 2, 2, 302-309.

Choueiki M.H. and Mount-Campbell, C. A. (1999) Training Data Development With The D-Optimality Criterion. *IEEE Transactions On Neural Networks*, 10, 1, 56-63.

Cohen, A. and Kovacevic, J. (1996) Wavelets: A Mathematical Background. *Proc. IEEE*, 84, 514-522.

Collins, G. S. (2000) *Multivariate Analysis of Flow Cytometry Data*, Unpublished Ph.D. Thesis, University of Exeter.

Culverhouse, P.F., Simpson, R.G., Ellis, R., Lindley, J.A., Williams, R., Parisini, T., Reguera, B., Bravo, I., Zoppoli, R., Earnshaw, G., Mccall, H. and Smith, G. (1996) Automatic Classification of Field-Collected Dinoflagellates By Artificial Neural Networks. *Marine Ecology Progress Series*, 139, 1-3, 281-287.

Demers, S., Kim, J., Legendre, P. and Legendre, L. (1992) Analysing Multivariate Flow Cytometric Data in Aquatic Sciences. *Cytometry*, 13, 291-298.

DeSieno (1988) Adding A Conscience to Competitive Learning. *Proc. Int. Conf. On Neural Networks*, 1, 117-124. IEEE Press, New York.

Dillon, W. R. and Goldstein, M. (1984) *Multivariate Analysis : Methods and Applications*. John Wiley and Sons, Inc.

Drucker, D.B. (1987) *Microbiological Applications of High Performance Liquid Chromatography*. Cambridge University Press.

Dubelaar, G.B.J., Groenewegen, A. C., Stokdijk, W. Van Den Engh, G.J. and Visser, J.W.M. (1989) Optical Plankton Analyser: A Flow Cytometer for Plankton Analysis, II: Specifications. *Cytometry*, 10.529-539.

Dubelaar, G.B.J., KöNig J.W., Cunningham, A. Groenewegen, A.C., Jonker, R.R., Wietzorrek, J., Rutten, T.P.A., Beeker, A.E.R. (1994a) EurOPA: A Novel 'High Definition' Flow Cytometer for Phytoplanktonic Cells and Colonies. In: *Conference Proceedings Oceans '94 OSATES, Brest.*

Dubelaar, G.B.J., Jonker, R.R., Meulemans, J.T.M., Van Veen, J.J.F. (1994b) Phytoplankton Analysis By (EurOPA) Flow Cytometry; Current and Future Applications in Environmental Control. In: *Conference Proceedings Oceans '94 Osates, Brest.*

Ellis, R., Simpson, R., Culverhouse, P.F. and Parisini, T. (1997). Committees, Collective and Individuals: Expert Visual Classification By Neural Network. *Neural Computing & Applications,* 5, 2, 99-105.

Foody, G.M. (1995) Training Pattern Replication and Weighted Class Allocation in Artificial Neural Networks. *Neural Computing and Applications*, 3, 178-190.

Frankel, D.S., Olson, R.J., Frankel, S.L. Chisholm, S.W. (1989) Use of A Neural Net Computer System for Analysis of Flow Cytometric Data of Phytoplankton Populations. *Cytometry*, 10, 540-550.

Frankel, D.S., Frankel, S.L., Binder, B.J. and Vogt, R.F. (1996) Application of Neural Networks to Flow Cytometry Data Analysis and Real-Time Cell Classification. *Cytometry*, 23, 290-302.

Fritzke, B. (1991) Let It Grow – Self-Organizing Feature Maps With Problem Dependent Cell Structure. In: *Proceedings International Conference On Artificial Neural Networks*, Helsinki, 403-408. Elsevier Science.

Fukunaga, K. and Hayes, R.R. (1989) Effects of Sample Size in Classifier Design. *IEEE Transactions On Pattern Analysis and Machine Intelligence*, 11, 8, 873-885.

Fukushima, H. (1993) High Reflectance Water Mass Observed By Nimbus-7 CZCS. In : *Satellite Remote Sensing of the Oceanic Environment*, Jones, S.F., Sugimori, Y. and Stewart, R.W. (eds.) Seibutsu Kenkyusha, Tokyo.

Goldstine, H. (1972) *The Computer from Pascal to von Neumann*, Princeton Univ. Press, Princeton New Jersey.

Guillard, R.R.L. (1975) Culture of phytoplankton for feeding marine invertebrates. In: *Culture of marine invertebrate animals*, Smith, W.L. and Chanley, M.H. (eds) Plenum, New York.

Harris, G.P. (1986) *Phytoplankton Ecology : Structure, Function and Fluctuation.* Chapman and Hall.

Harris, R. (1987) *Satellite Remote Sensing : An Introduction.* Routledge and Kegan Paul Ltd.

Harrington, P.D. (1993) Sigmoid Transfer Functions in Backpropagation Neural Networks. *Analytic. Chem*, 65, 2167-2168

Hashem, S (1997) Optimal Linear Combinations of Neural Networks. *Neural Networks*, 10, 599-614.

Haykin, S. (1994) *Neural Networks: A Comprehensive Foundation*, Macmillian College Publishing Company.

Heimer, L. (1995) *The Human Brain and Spinal Cord.* Springer-Verlag New York, Inc.

Huang, K., Wu, J. and Yan, H. (1997) Off-Line Writer Verification Utilising Multiple Neural Networks. *Optical Engineering*, 36, 11, 3127-33.

Hush, D.R. and Horne B.G. (1993) Progress in Supervised Neural Networks: What's New Since Lippman? *IEEE Signal Processing Magazine*, 10, 8-39.

J. Iivarinen, T. Kohonen, J. Kangas, and S. Kaski (1994) Visualizing the Clusters On the Self-Organizing Map. In: *Proc. of Conference On Artificial Intelligence Research,* Finland, 122-126.

Jain, A.K. and Chandrasekaran, B. (1982) Dimensionality and Sample Size Considerations in Pattern Recognition Practice. *Handbook of Statistics*, Krishnaiah, P.R. and Kanal, L.N. (eds.), 835-855, North Holland Publishing Company, Amsterdam.

Joliffe, I.T. (1986) *Principal Component Analysis*. Springer Series in Statistics, Springer-Verlag, New-York.

Juell, P. and Marsh, R. (1996) A Hierarchical Neural Network for Human Face Detection. *PatternRecognition.* Vol. 29, 5, 781-787.

Khotanzad, A., Afkhami-Rohani, R., Lu, T., Abaye, A., Davis, M. and Maratukulam, J.(1997) ANNSTLF – A Neural Network Based Electric Load forecasting System. *IEEE Transactions On Neural Networks.* 8, 4, 835-846

Kohonen (1988) An Introduction to Neural Computing. *Neural Networks*, 1, 3-16.

Kohonen, T. (1990) The Self-Organising Map. *Proc. IEEE*, 78, 1464-1480.

Kohonen, T. (1997) *Self-Organising Maps*. 2nd Ed., Springer Verlag, New York.

Kolomogrov, A.N. (1957) On the Representation of Continuous Functions of Several Variables By Superposition of Continuous Functions of One Variable and Addition. *Doklady Akademiia Nauk SSSR*, 114, 953-956.

Kothari, R. & Pitts, D. (1999) On Finding the Number of Clusters. *Pattern Recognition Letters,* 20, 4, 405-416.

Kraaijveld, M.A., Mao J. and Jain, A.K. (1992) A Non-Linear Projection Method Based On Kohonen's Topology Preserving Maps. *Proc. 11th IAPR International Conference On Pattern Recognition*, 2, 41-45.

Krzanowski, W.J. (1993) *Principles of Multivariate Analysis : A Users Perspective.* Oxford Statistical Science Series, Oxford University Press.

Lee, Y., Oh, S. and Kim, M. (1991) The Effect of Initial Weights On Premature Saturation in Back-Propagation Learning. *International Joint Conference On Neural Networks,* 1, 765-770.

Lewis, O.M., Ware, J.A. and Jenkins, D. (1997) A Novel Neural Network Technique for the Valuation of Residential Property. *Journal of Neural Computing and Applications*, 5, 224-229.

Lippman, R. (1987) An Introduction to Computing With Neural Nets. *IEEE Acoustics, Speech and Signal Processing Magazine*, 4, 4-22.

McLachlan, G.J. (1992) *Discriminant Analysis and Statistical Pattern Recognition.* Wiley, New York.

Mehdi, B., Stacey, D. and Harauz, G. (1994). A Hierarchical Neural Network Assembly for Classification of Cervical Cells in Automated Screening. *Analytical Cellular Pathology*, 7, 171-180.

Moody, J.E. and Darken, C. (1989) Fast Learning in Networks of Locally-Tuned Processing Units. *Neural Computation*, 1, 281-294.

Morris, C.W., Boddy, L & Wilkins, M.F. (1994) Approaches to Applying Neural Networks to the Identification of Phytoplankton Taxa From Flow Cytometry Data. In: *Intelligent Engineering Systems Through Artificial Neural Networks*, Dagli, C.H., Fernandez, B.R., Ghosh, J. Kumara, R.T. (eds.) 4, 619-629. ASME Press

Morris, C.W. and Boddy, L. (1995) Artificial Neural Networks in Identification and Systematics of Eukaryotic Microorganisms. *Binary*, 7, 70-76.

Morris, C.W. and Boddy, L. (1996) Classification As Unknown By RBF Networks: Discriminating Phytoplankton Taxa From Flow Cytometry Data. *Intelligent Systems Through Artificial Neural Networks*, 6, 629-634, ASME, Press, New York.

Murtagh, F. (1995) Interpreting the Kohonen Self-Organizing Feature Map Using Contiguity-Constrained Clustering. *Pattern Recognition Letters*, 16, 399-408.

Musavi, M. T., Ahmed, W. Chan, K.H., Faris, K.B. and Hummels, D.M. (1992) On the Training of Radial Basis Function Classifiers. *Neural Networks*, 5, 595-603.

Namphol, A., Chin, S. and Arozullah, M. (1996). Image Compression With A Hierarchical Neural Network. *IEEE Transactions On Aerospace and Electronic Systems.* 32, 1, 326-337.

Nering, E.D. (1970) *Linear Algebra and Matrix Theory*, John Wiley and Sons, Inc.

Neuralware, Inc. (1991a) *Neural Computing.* Neuralware, Inc., Pittsburg, USA.

Neuralware, Inc. (1991b) *Using Nworks.* Neuralware, Inc., Pittsburg, USA.

Olson, R.J., Frankel, S.W. & Chisholm S.W. (1983). An Inexpensive Flow Cytometer for the Analysis of Fluorescence Signals in Phytoplankton: Chlorophyll and DNA Distributions. *J. Exp. Mar. Biol. Ecol.* 68, 129-144.

Peeters, J.C.H., Dubelaar, G.B.J., Ringelberg, J. and Visser, J.W.M. (1989) Optical Plankton Analyser: A Flow Cytometer for Plankton Analysis, I: Design Considerations. *Cytometry*, 10, 522-528.

Pitas, I. (1993) *Digital Image Processing Algorithms - Prentice Hall Series in Acoustics, Speech and Signal Processing*. Prentice Hall, London.

Pitts, D., Kothari, R. and Visscher, M. (1999) An Algorithm for Automatically Determining the Number of Clusters. *Pattern Recognition Letters*, 20, 4, 405-416.

Powell, M.J.D. (1985) Radial Basis Functions for Multivariable Interpolation: A Review. *IMA Conference On Algorithms for the Approximation of Functions and Data*, RMCS, Shrivenham, UK, 143 – 167.

Preisendorfer, R.W. (1988) *Principal Component Analysis in Meteorology and Oceanography*. Elsevier, New York.

Raghavan, S., Gupta, N., Lambird, B., Lavine, D. and Kanal, L. (1991) A Layered Object Recognition System Using A Hierarchical Hybrid Neural Network Architecture. *Proc. SPIE Model Based Vision Development and Tools*, 1609, 49-60.

Randell, B. (1982) *Origins of Digital Computers: Selected Papers*, Springer-Verlag, Berlin Heidelberg.

Raudys, S. and Pikelis, V. (1980) On Dimensionality, Sample Size, Classification Error, and Complexity of Classification Algorithm in Pattern Recognition. *IEEE Transactions On Pattern Analysis and Machine Intelligence*, 2, 3, 242-252.

Refenes, A.N. and Alippi, C. (1991) Histological Image Understanding By Error Back propagation. *Microprocessing and Microprogramming*, 32, 437-446.

Refenes, A.N., Azema-Barac, M., Chen., L & Karoussos, S.A. (1993) Currency Exchange Rate Prediction and Neural Network Design Strategies. *Neural Computing & Applications*. 1, 46-58.

Richard, M.D. and Lippman, R.P. (1991) Neural Network Classifiers Estimate Bayesian *A Posteriori* Probabilities. *Neural Computation*, 3, 461-483.

Rumelhart, D.E. and Mcclelland, J.L. (eds.) (1986) Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Vol 1, Cambridge, MA, MIT Press.

Sakshaug, E. (1980) Problems in the Methodology of Studying Phytoplankton. In: *The Physiological Ecology of Phytoplankton.* Morris, I. (ed.) p57-91. Blackwell Scientific Publications.

Scardi, M. (1996) Artificial Neural Networks As Empirical Models for Estimating Phytoplankton Production. *Marine Ecology Progress Series,* 139, 1-3, p289-299.

Scardi, M. (1998) Neural Network Models of Phytoplankton Primary Production. *International Workshop On Applications of Artificial Neural Networks to Ecological Modelling*, Toulouse.

Schalkoff, R.J. (1992) *Pattern Recognition: Statistical, Structural and Neural Approaches.* Wiley International, Chichester.

Shazeer, D. (1992) Hierarchical Neural Networks for the Classification of Undersea Events. *Applications of Artificial Neural Networks.* 1709, 808-822

Simpson, R., Williams, R., Ellis, R. and Culverhouse, P.F. (1992) Biological Pattern Recognition By Neural Networks. *Marine Ecology Progress Series,* 79, 303-308.

Singer, E. and Lippmann, R. P. (1992) Improved Hidden Markov Model Speech Recognition Using Radial Basis Function Networks. In: *Neural Information Processing Systems 4,* Moody, J., Hanson, S. and Lippmann, R. (eds.), 4, 159-168. Morgan Kaufmann Publishers, Inc., California.

Smits, J.R.M., Breedveld, L.W., Derksen, M.W.J., Kateman, G., Balfoort, H.W., Snoek, J. and Hofstraat, J.W. (1992) Pattern Classification With Artificial Neural Networks: Classification of Algae, Based Upon Flow Cytometer Data. *Analytica Chimica Acta*, 258, 11-25.

Steen, H.B. (1991) Flow Cytometry Instrumentation. In :*Particle Analysis In Oceanography.* Demers, S. (ed.) p3-29. Springer-Verlag, Berlin.

Stefánsson, A., Koncar, N. and Jones, A.T. (1997) A Note On the Gamma Test. *Journal of Neural Computing and Applications*, 5, 3, 131-133.

Steidinger K.A. and Tangen, K. (1997) Dinoflagellates. In : *Identifying Marine Phytoplankton*, Tomas, C.R. (ed.). Academic Press, Inc.

Throndsen, J. (1997) The Planktonic Marine Flagellates. In : *Identifying Marine Phytoplankton*, Tomas, C.R. (ed.). Academic Press, Inc.

Tou, J.T. and Gonzalez, R.C. (1974) *Pattern Recognition Principles.* Addison-Wesley, London.

Trask, B.J., Van Den Engh, G. J., & Elgershuizen, J.H.B.W. (1982) Analysis of Phytoplankton By Flow Cytometry. *Cytometry*, 2, 4, 258-264.

Ultsch, A. and Siemon, H.P (1989) Kohonen's Self-Organizing Feature Maps for Exploratory Data Analysis. *Proc. INNC '90 Int. Neural Network Conf.*, 305-308.

Vesanto, J., Vasara, P., Helminen, R-R. and Simula, O. (1997) Intergrating Environmental, Technological and Financial Data in forest Industry Analysis. In: *Proc. of SNN.* Netherlands.

Vidakovic, B. (1999) *Statistical Modeling By Wavelets* - Wiley Series in Probability and Mathematical Statistics. Wiley, New York.

de Villiers, J. and Barnard, E. (1992) Backpropagation Neural Nets With One and Two Hidden Layers. *IEEE Transactions On Neural Networks*, 4, 1, 136-141.

von Neumann, J. (1945) First Draft of a Report on the EDVAC, Contract No. W-670-ORD-492, Between the United States Army Ordnance Department and the University of Pennsylvania Moore School of Electrical Engineering, Philadelphia.

Wang, D., Keller, J.M., Carson, C.A., Mcadoo-Edwards, K.K. and Bailey, C.W. (1998) Use of Fuzzy-Logic-Inspired Features to Improve Bacterial Recognition Through Classifier Fusion. *IEEE Transactions On Systems, Man and Cybernetics – Part B: Cybernetics*, 28, 4, 583-591.

Wigley, T.M.L. (1994) Climate Change - Outlook Becoming Hazier. *Nature*, 369, 709-710.

Wilkins, M., Morris, C. and Boddy, L. (1994a) Kohonen Maps and Learning Vector Quantisation Neural Networks for Analysis of Multivariate Biological Data. *Binary*, 6, 64-72.

Wilkins, M., Boddy, L. and Morris, C. (1994b) - A Comparison of Radial Basis Function and Backpropagation Neural Networks for Identification of Marine Phytoplankton From Multivariate Flow Cytometry Data. *CABIOS*, 10, 285-294.

Wilkins, M.F. (author) Aimsnet Software (2000) developed during the Aims project (Automated Identification and Characterisation of Microbial Populations) URL: www.cf.ac.uk/biosi/staff/wilkins/aimsnet (Date last visited Oct, 2000)

Wilkins, M.F., Boddy, L., Morris, C.W. and Jonker, R.R. (1999) Identification of Phytoplankton From Flow Cytometry Data By Using Radial Basis Function Neural Networks *Appl. Environ. Microbiol*, 65,4404-4410.

Wilson, C.L., Grother, P.J. and Barnes, C.S. (1996) Binary Decision Clustering for Neural Network Based Optical Character Recognition. *Pattern Recognition.* 29, 3, 425-437.

Witkin, A.P. (1983) Scale-Space Filtering. In: *Proc. IJCAI,* 1019-1021.

Wong, Y. (1996) Clustering Data By Melting. *Neural Computation,* 5, 89-104.

Yang, B., Carotta, M.C., Faglia, G., Ferroni, M., Guidi, V., Martinelli, G. and Sberveglieri, G. (1996). A Hybrid Neural Network Based Pattern-Recognition Engine for Out-Door Electronic Nose Application. In: *Proc ANNIE,* 449-457.

Yentsch, C.M., Horan, P.K., Muirhead, K., Dortch, Q., Haugen, E., Legendre, L., Murphy, L.S., Perry, M.J., Phinney, D.A., Pomponi, S.A., Spinrad, R.W., Wood, M., Yentsch, C.S. and Zahuranec, B.J. (1983). Flow Cytometry and Cell Sorting: A Technique for Analysis and Sorting of Aquatic Particles. *Limnology and Oceanography.* 28, 1275-1280.

Zhang, J. (1999) Developing Robust Non-Linear Models Through Bootstrap Aggregated Neural Networks. *Neurocomputing,* 25, 93-113.

Appendix 1 – Phytoplankton Characteristics

Table A1.1 Some of the primary physical features used to identify phytoplankton cells by morphology.

Cell Feature	Description
Cell Shape	Can be either a common or variable factor between cells of the same species.
Cell dimensions	Microplankton and nanoplankton consist of both zooplankton and phytoplankton, with a maximum size scale of 20-200μm and 2-20μm respectively. Picoplankton contains only algae and bacteria with sizes less than 2μm.
Cell Wall	Not always present; may be replaced by a plasma membrane. Requires high magnification to ascertain presence. Identification of some species can be evident from the walls containing inorganic substances, such as silica. This is detected through horizontal light scatter, forming either scales or a continuous cratered wall.
Mucilage Layer	Only visible by light microscopy when stained with Indian ink (Boney, 1989). An extension of the cell wall
Chloroplasts	Chlorophyll bearing cell constituents distinctive by colour, size and number giving a means of identifying a cell. The various pigments present in a particular cell are represented by the colour of the chloroplasts, for example some phytoplankton appear yellow due to the presence of xanthophyll pigments overlying the green chlorophylls. Chlorophyll *a* is the primary chlorophyll. It is the receptor of radiant energy and common to all photosynthetic organisms. It is used not only for identification of species through colour but also as a measure of phytoplankton biomass. Fluorescence per unit chlorophyll *a* is an indicator of species type but must be measured in a mixture of pigments including chlorophyll *b*, chlorophyll *c* and chlorophyll derivatives, some of which interfere with analysis (Sakshaug, 1980)
Flagella	Whip-like protrusions from a cell varying in position, number and physical appearance.

Table A1.2 Characteristics of some of the phytoplankton classes used in this research.

Group	Details
Bacillariophyceae (Diatoms)	Commonly found in marine or freshwater areas, Bacillariophyceae exist as single cells or chains, with flagella present only in male gametes. The cell forms an external silica skeleton and has one to many chloroplasts. Pigments are primarily chlorophyll *a* and *c* as well as xanthophyll, which gives the cell a yellow-brown appearance; Diatoms with a raphe are capable of independent movement.
Dinophyceae (Dinoflagellates)	Extensively found in marine and freshwater areas, the Dinoflagellates exist generally as single cells or chains, with a few being filamentous. Each cell has two dimorphic flagella, one transverse and one longitudinal, as well as a characteristic theca (cell covering) distinguishing them from other groups (Steidinger & Tangen, 1997). Cells contain one to many chloroplasts. As well as chlorophyll a and c, the cells contain fucoxanthin and peridinin (xanthophyll pigments) giving them a yellow-green or yellow-brown appearance.
Chlorophyceae (Green algae)	Characteristic of coastal and freshwater environments, Chlorophyceae can be single or colonial with a solitary chloroplast. Some species are motile, possessing two or four flagella, smooth and equal in length. Chlorophylls *a* and *b* are the photosynthetic pigments present, giving the algae their green colour.
Prasinophyceae (Green algae)	Evident in marine and freshwater areas, Prasinophyceae exist as single cells with an individual chloroplast. They commonly possess two or four flagella, which are thicker than those of Chlorophyceae due to a covering of organic scales. As with Chlorophyceae the green colour comes from the chlorophylls *a* and *b*; cells are motile.
Euglenophyceae (Euglenoid flagellates)	Found mainly in freshwaters, Euglenoid flagellates are bright green in colour, existing as individual cells with one to many chloroplasts of various shapes. The cells posses one, two or four flagella, aiding homo or heterodynamic movement (Throndsen, 1997), and contain chlorophylls *a* and *b*.
Chrysophyceae (Golden-brown phytoflagellates)	Mainly freshwater, Chrysophyceae exist either independently or in a colonial state, with some cells exhibiting a covering of silicified or organic scales. These phytoflagellates possess two rough flagella of unequal lengths and may be motile. Each cell houses one or more chloroplasts where, as well as chlorophylls *a* and *b*, the presence of a number of carotenoid pigments give the species a golden-brown or yellow appearance.

Table A1.2 continued.....

Group	Details
Prymnesiophyceae (Brown phytoflagellates)	Principally a marine class of species but can be found in freshwater areas. Existing either individually or as colonies, Prymnesiophyceae contain one or two chloroplasts and the pigments chlorophyll *a* and *b*. The species have two, generally smooth, flagella of equal lengths, aiding motility, and are yellow to golden-brown in colour.
Cryptophyceae (Brown flagellate)	This is mainly a freshwater class with few coastal inhabitants. The cells exist individually possessing two, rough, unequal flagella. Cryptophyceae normally have one or two chloroplasts and the presence of the phycobilin pigments, phycoerythrin (red) and phycocyanin (blue), gives the cells a variety of colours, including brown, green, red and blue. Cells are motile.
Cyanophyceae (Cyanobacteria; blue-green algae)	A marine and freshwater group living as single cells, colonies or filaments. The absence of a distinct nucleus in these cells produces characteristics more akin with bacteria, *i.e.* prokaryotic, than algae. Cells contain subsidiary pigments, phycobilins, phycocyanin and carotenoids accompanying the chlorophyll, giving the more common blue-green colour with variations of olive and sometimes yellow and red.
Rhodophyceae (Red algae)	Rhodophyceae are primarily marine algae and are non-motile. The cells contain carotenoids and phycobilin pigments dominated by phycoerythrin, producing the distinctive red colour.
Phaeophyceae (Brown seaweeds)	Phaeophyceae are marine inhabitants with two lateral unequal flagella. The cell colour is characterised by the presence of the yellow xanthophyll pigment, fucoxanthin.

Appendix 2 – Biological Glossary

Term	Definition
Carotenoid	Orange, brown, red or yellow photosynthetic pigments comprising Xanthophylls and Carotenes pigments.
Chlorophylls	Photosynthetic pigments present in all plants. Comprises Chlorophyll a, b, c, d and Chlorophyll derivatives.
F/2 Medium	Sterilised seawater with the addition of nutrients, trace elements and vitamins to ensure phytoplankton growth. Major nutrients are sodium nitrate and sodium hydrogen phosphate. Trace elements include iron, copper, zinc, cobalt, manganese and molybdenum as various salts. Vitamins include vitamin B12 (Cyanocobalamine), B1 (Biotin) and B6 (Thiamin hydrochloride) (Guillard, 1975).
F/10 Medium	F/2 medium with all the nutrient concentrations divided by 5 for oceanic phytoplankton (mimics oceanic conditions). (Tarran, *pers. comm.*)
Fluorochrome	Fluorescent compound
Heterodynamic movement	Flagella of individual cell used in different ways to produce movement
Homodynamic movement	Flagella of individual cell used in the same way to produce movement
Phycobilin	Photosynthetic pigments comprising Phycoerythrin (red) and Phycocyanin (blue) pigments.
Raphe	Slit found in Diatom valves
Zooplankton	Animal constituent of Plankton

Appendix 3 – Kohonen's Self Organising Map

Notation

$\mathbf{x} = (x_1,x_2,.....,x_p)$	p-dimensional feature vector of the input pattern.
$\mathbf{w_j} = (w_1,w_2,.....,w_p)$	Position (weight) vector of node j in the Kohonen layer.
$D(x,w_j)$	Euclidean distance between input pattern, x, and node j.
$D'(x,w_j)$	Euclidean distance between input pattern, x, and node j after bias has been employed.
J	Winning node.
$d(w_j,w_J)$	Euclidean distance between winning node J and node j.
$f(d(w_j,w_J))$	Neighbourhood function, generally Gaussian, where $$f(d) = \exp\left(-\frac{d^2}{2\sigma^2(t)}\right)$$
$\sigma(t)$	Neighbourhood radius – standard deviation of the Gaussian.
b_j	Bias term.
N	Number of nodes in the Kohonen layer.
F_j	Winning frequency of node j.
B	Constant chosen to ensure frequency does not mirror anomalies in the data 0<B<<1 (DeSieno, 1988).
γ	Constant controlling the bias factor.
α	Learning parameter.

Algorithm

Step 1 The weight vectors of the Kohonen nodes are initially set to small random values of the same dimension as the input data. Network parameters are set (see below – Kohonens Recommendations).

Step 2 Present randomly chosen input pattern, **x**, to network.

Step 3 For each Kohonen node j in the lattice, do steps 3.1 and 3.2.

Step 3.1 Calculate $$D(x,w_j) = \sqrt{\sum_{i=1}^{p}\left(x_i - w_{ji}(t)\right)^2}$$

Step 4 Find the node J such that $D(x,w_J)$ is a minimum.

Step 5 For all Kohonen nodes j, do steps 5.1 and 5.2.

Step 5.1 Calculate $d(w_j,w_J)=\sqrt{\sum_{i=1}^{p}(w_{ji}(t)-w_{Ji}(t))^2}$

Step 5.2 Update all nodes within the specified neighbourhood of node J.

$$w_{ji}(t+1)=w_{ji}(t)+\alpha(t)f(d(w_j,w_J))[x_i(t)-w_{ji}(t)]$$

Steps 2-5 are repeated until either stopping condition is reached or procedure is halted.

Algorithm – with conscience

Step 1 The weight vectors of the Kohonen nodes are initially set to small random values of the same dimension of the input data.

Network parameters are set (see below – Kohonens Recommendations).

Step 2 Present randomly chosen input pattern, **x**, to network.

Step 3 For each Kohonen node j in the lattice, do steps 3.1 and 3.2.

Step 3.1 Calculate $D(x,w_j)=\sqrt{\sum_{i=1}^{p}(x_i-w_{ji}(t))^2}$

Step 3.2 Calculate $D'(x,w_j)=D(x,w_j)-b_j$

where $b_j=\gamma\left[\frac{1}{N}-F_j(t)\right]$

Step 4 Find the node J such that $D'(x,w_J)$ is a minimum.

Step 5 For all Kohonen nodes j, do steps 5.1 and 5.2.

Step 5.1 Calculate $d(w_j,w_J)=\sqrt{\sum_{i=1}^{p}(w_{ji}(t)-w_{Ji}(t))^2}$

Step 5.2 Update all nodes within the specified neighbourhood of node J.

$$w_{ji}(t+1)=w_{ji}(t)+\alpha(t)f(d(w_j,w_J))[x_i(t)-w_{ji}(t)]$$

Step 6 Winning frequencies are updated for all nodes j

$$F_j(t+1)=F_j(t)(1-B)+B \quad \text{if } j=J$$

$$F_j(t+1)=F_j(t)(1-B) \quad \text{otherwise}$$

Steps 2-6 are repeated until either a stopping condition is reached, or procedure is halted.

Kohonen's Recommendations (Kohonen, 1990)

The following recommendations are set out by Kohonen for unsupervised learning by the Kohonen self-organising map.

Step 1 Initial period of global ordering

(a) Train for approximately 1000 presentations with the following parameter recommendations.

(b) The learning parameter, $\alpha(t)$, should be close to 1 and decrease monotonically. This reduction can be linear, exponential or inversely proportional to t

(c) To maximise cluster representation the training should initiate with a large neighbourhood region, which can be greater than the longest dimension of the Kohonen lattice, and shrink linearly during training to 1 node update.

Step 2 Fine adjustment

(a) Kohonen recommends training for approximately 500 x N presentations in total for good statistical accuracy. Parameter recommendations for the further training period are shown below.

(b) The learning parameter, $\alpha(t)$, should be <=0.01, with reduction either linear or exponential.

(c) The neighbourhood update region can remain at 1 node allowing those nodes immediately adjoining the winning node to continue being updated.

Appendix 4 – Publications and Posters

Papers

Al-Haddad, L.M. and Morris C. (1995) CORTEX PRO Neural Network Development System. *Binary Computing in Microbiology*, 7, 116-117.

Al-Haddad, L. Morris, C.W. and Boddy, L. (2000) Training Radial Basis Function Neural Networks: Effects of Training Set Size and Imbalanced Training Sets. *Journal of Microbiological Methods* 43:33-44.

Boddy, L., Morris, C.W., Morgan, A. and Al-Haddad, L.M. (1998) Neural Network Approaches To Interpreting Variability. In: *Molecular Variability of Fungal Pathogens* Bridge, P. D., Couteaudier, Y. and Clarkson, J. M. (Eds.), Cab International, New York, P279-290.

Boddy L., Morris C.W., Wilkins M.F., Al-Haddad L., Tarran G.A., Jonker R.R. and Burkill P.H. (2000) Identification of 72 Phytoplankton Species By Radial Basis Function Neural Network Analysis of Flow Cytometric Data. *Marine Ecology Progress Series* 195, 47-59.

Posters

Al-Haddad, L. M. (1997) Classification of Phytoplankton Through Unsupervised Artificial Neural Networks. *BIONET International Group Computer Aided Taxonomy*, Cardiff University.

Tarran, G.A., Al-Haddad, L.M., Collins, G.S., Burkill, P.H., Morris, C.W. and Krzanowski, W.J. (1998) Flow Cytometric Procedures for the Analysis of Phytoplankton Community Structure and Function. *Prime Symposium*, Bangor University.

Tarran, G.A., Al-Haddad, L.M., Collins, G.S., Burkill, P.H., Morris, C.W. and Krzanowski, W.J. (1998) Multivariate Data Analysis Developments for the Characterisation of Phytoplankton Taxa. *Prime Symposium*, Bangor University.

Tarran, G.A., Al-Haddad, L.M., Collins, G.S., Burkill, P.H., Morris, C.W. and Krzanowski, W.J. (1998) Flow Cytometric Technical Developments for the Analysis of Microbial Communities. *Prime Symposium*, Bangor University.

Tarran, G.A., Al-Haddad, L.M., Collins, G.S., Burkill, P.H., Morris, C.W. and Krzanowski, W.J. (1998) Phytoplankton Community Function : Grazing By Microzooplankton. *Prime Symposium*, Bangor University.

Presentations

Tarran, G.A., Al-Haddad, L.M., Collins, G.S., Burkill, P.H., Morris, C.W. and Krzanowski, W.J. (1997) Analysis of Phytoplankton Community Structure and Function. *PRIME Symposium*, Warwick University.