MACHINE-LEARNING APPROACHES FOR MODELLING FISH POPULATION DYNAMICS

A thesis submitted for the degree of *Doctor of Philosophy*

by

Neda Trifonova

Department of Computer Science

August 2016

Abstract

Ecosystems consist of complex dynamic interactions among species and the environment, the understanding of which has implications for predicting the environmental response to changes in climate and biodiversity. Understanding the nature of functional relationships (such as prey-predator) between species is important for building predictive models. However, modelling the interactions with external stressors over time and space is also essential for ecosystem-based approaches to fisheries management. With the recent adoption of more explorative tools, like Bayesian networks, in predictive ecology, fewer assumptions can be made about the data and complex, spatially varying interactions can be recovered from collected field data and combined with existing knowledge.

In this thesis, we explore Bayesian network modelling approaches, accounting for latent effects to reveal species dynamics within geographically different marine ecosystems. First, we introduce the concept of functional equivalence between different fish species and generalise trophic structure from different marine ecosystems in order to predict influence from natural and anthropogenic sources. The importance of a hidden variable in fish community change studies of this nature was acknowledged because it allows causes of change which are not purely found within the constrained model structure.

Then, a functional network modelling approach was developed for the region of North Sea that takes into consideration unmeasured latent effects and spatial autocorrelation to model species interactions and associations with external factors such as climate and fisheries exploitation. The proposed model was able to produce novel insights on the ecosystem's dynamics and ecological interactions mainly because it accounts for the heterogeneous nature of the driving factors within spatially differentiated areas and their

changes over time.

Finally, a modified version of this dynamic Bayesian network model was used to predict the response of different ecosystem components to change in anthropogenic and environmental factors. Through the development of fisheries catch, temperature and productivity scenarios, we explore the future of different fish and zooplankton species and examine what trends of fisheries exploitation and environmental change are potentially beneficial in terms of ecological stability and resilience. Thus, we were able to provide a new data-driven modelling approach which might be beneficial to give strategic advice on potential response of the system to pressure.

Acknowledgements

First, I would like to thank my supervisor Allan Tucker for his constructive supervision, encouragement and patience. Without his guidance, in both work and life, this work would have never been completed. It has been a privilege and a pleasure to work with him.

I also would like to thank my industrial supervisors Andrew Kenny, and David Maxwell from the Centre for Environment, Fisheries and Aquaculture Science and Daniel Duplisea from Fisheries and Oceans, Canada for their invaluable advice and support.

I am grateful to Johan Van der Molen and John Pinnegar from the Centre for Environment, Fisheries and Aquaculture Science and Jose Fernandes from Plymouth Marine Lab for providing me with constructive support. They have been great partners in collaboration.

Thanks to Eva Garnacho, Chiara Franco and Yiota Apostolaki for the interesting discussions and the collaboration.

A big thank you to my colleagues (past and present) Valeria, Mohsina, Chuang, Liang, and Ovidiu and many others for keeping the department a pleasure to work in. Thanks to all my friends for their support and amazing moments.

Last but not least a special thanks to my partner Sean and my family for the love and incredible support over all these years.

Publications

The following publications have resulted from the research presented in this thesis:

- Trifonova, N., Kenny, A., Maxwell, D., Pinnegar, J. and Tucker, A. (in review). The potential use of a dynamic Bayesian network model to predict trends of ecosystem change in response to fisheries catch, temperature and productivity scenarios. ICES Journal of Marine Science.

- Trifonova, N., Kenny, A., Maxwell, D., Duplisea, D., Fernandes, J., and Tucker, A. (2015). Spatio-temporal Bayesian network models with latent variables for revealing trophic dynamics and functional networks in fisheries ecology. Ecological Informatics, 30, 142-158.

- Trifonova, N., Kenny, A., Duplisea, D. and Tucker, A. (2015). Generalisation across different marine ecosystems for revealing trophic dynamics and structure in fisheries ecology. ICES Annual Science Conference 2015, Theme session: Operationalizing ecosystem-based fisheries management.

- Trifonova, N., Duplisea, D., Kenny, A., Maxwell, D., and Tucker, A. (2014). Incorporating Regime Metrics into Latent Variable Dynamic Models to Detect Early-Warning Signals of Functional Changes in Fisheries Ecology. In Discovery Science (pp. 301-312). Springer International Publishing.

- Trifonova, N., Duplisea, D., Kenny, A., and Tucker, A. (2014). A Spatio-temporal Bayesian Network Approach for Revealing Functional Ecological Networks in Fisheries. In Advances in Intelligent Data Analysis XIII (pp. 298-308). Springer International Publishing.

- Trifonova, N., Kenny, A., Duplisea, D. and Tucker, A. (2013). The application of statistical metrics into predictive functional Bayesian models of fisheries ecology. ICES Annual Science Conference 2013, Theme session: Advantages of Bayesian analysis for fisheries and ecological research.

Contents

List of Figures **xii**

List of Tables **xxi**

Glossary **xxiii**

1 Introduction **1**

1.1 Motivation 1

1.2 Thesis contributions 4

1.3 Thesis outline 5

2 Background **6**

2.1 Fisheries and Ecoinformatics 6

2.2 Functional Network Models 9

2.3 Bayesian Networks for Modelling Ecological Data 12

2.4 Integration of Expertise and Multiple Data Sets 15

2.5 Summary 17

3 Preliminaries **19**

3.1 Bayesian Networks 19

3.1.1 Bayesian Inference 22

3.1.2 Parameter Estimation 23

3.1.3 Structure Learning 24

3.1.4 Model Evaluation 25

3.2 Introduction of Time Series, State-space Models and Autoregressive process 26

3.3 Dynamic Bayesian Networks 29
3.3.1 Learning and Inference 31
3.4 Hidden Markov Model 34
3.4.1 Learning Hidden Markov Models 37
3.4.2 Hidden Variables 38
3.5 Autoregressive Hidden Markov Model 39
3.5.1 Inference and Learning 41
3.6 Conclusion 42

4 Exploring Regime Metrics into Bayesian Network Models with Hidden Variables to Detect Early-Warning Signals of Functional Changes in Fisheries Ecology 43

4.1 Introduction 43
4.2 Description of the Geographic Regions and Biomass Data to Model Regime Shifts 45
4.3 Methods 46
4.3.1 Introducing an Algorithm for Learning Functional Equivalence . 46
4.3.2 Hidden Variable Models to Predict Regime Shifts and Species Biomass 48
4.3.3 Learning Data-Driven Networks versus Pre-defined Diet Matrices 50
4.4 Results 51
4.4.1 Identified Functionally Equivalent Species from Different Geographic Regions 51
4.4.2 Prediction Accuracy of Hidden Variable Models in the East Scotian Shelf and North Sea 54
4.4.3 Comparison of Data-Learned Networks to Expert Diet Matrices in the East Scotian Shelf and North Sea 59
4.5 Summary 64

5 Spatio-temporal Bayesian Network Models for Revealing Trophic Dynamics and Functional Networks in Fisheries Ecology 66

5.1 Introduction 66
5.2 Description of Spatio-temporal Survey Data to Model Functional Networks across Two Geographic Regions 68
5.2.1 Gulf of St Lawrence 68
5.2.2 North Sea 69
5.3 Methods 71
5.3.1 Introducing a Technique to Learn the Structure of Bayesian Network Models for the Gulf of St Lawrence and North Sea 71
5.3.2 Spatial Autocorrelation 73
5.3.3 Generating Dynamic Bayesian Network Models to Predict Biomass in the Gulf of St Lawrence and North Sea 73
5.3.4 DBN with Two Hidden Variables and Spatial Nodes: HSDBN . . 75
5.4 Results for the Gulf of St Lawrence 77
5.4.1 Identified Functional Relationships within Spatio-temporal Scales 77
5.4.2 Prediction Performance of Dynamic and Spatial Dynamic Models 82
5.4.3 Summary for the Gulf of St Lawrence 85
5.5 Results for the North Sea 86
5.5.1 Comparative Evaluation of Biomass Predictions 86
5.5.2 Analysis of the General and Specific Hidden Variables within Spatio-temporal Scales 90
5.5.3 Analysis of the Discovered Interactions between Species and their Environment 91
5.5.4 Summary for the North Sea 94

6 A Dynamic Bayesian Network Model to Predict Trends of Ecosystem Change in Response to Fisheries Catch, Temperature and Productivity Scenarios 97
6.1 Introduction 97
6.2 Description of the Spatio-temporal Data to Model Natural and Anthropogenic Scenarios 99
6.3 Methods 99

6.3.1 Description of the Model to Predict Species Trends in Response to Natural and Anthropogenic Scenarios . . . 99
6.3.2 Modelling Species Trends in Response to Fisheries Catch, Temperature and Productivity Scenarios . . . 102
6.4 Results . . . 105
6.4.1 Fisheries Catch Scenarios . . . 105
6.4.2 Temperature and Net PP Scenarios . . . 115
6.5 Summary . . . 118

7 Models for Generalisation across Different Marine Ecosystems 121
7.1 Introduction . . . 121
7.2 Description of Biomass Data to Model Generalisation across Two Geographic Regions . . . 122
7.3 Methods . . . 123
7.3.1 Generalisation between Different Geographic Regions to Predict Species Biomass . . . 123
7.3.2 Modelling Hidden Variables to Identify Controlling Mechanisms and Drivers of Change in the North Sea . . . 125
7.4 Results . . . 125
7.4.1 Generating Biomass Predictions for the North Sea using Dynamic Models from the Gulf of St Lawrence . . . 125
7.4.2 Identified Driving Factors for the Ecosystem Dynamics in the North Sea . . . 133
7.5 Summary . . . 138

8 Conclusions 140
8.1 Thesis Contributions . . . 140
8.1.1 Functional Network Models with Hidden Variables . . . 140
8.1.2 Hidden Spatial Dynamic Bayesian Network Model . . . 141
8.1.3 Application to Several Datasets . . . 142
8.2 Limitations . . . 143

8.3 Future Work 144
8.3.1 Application to Different Kind of Data 144
8.3.2 Extension of Modelling Techniques 146

A Additional Results **147**
A.1 Chapter 5 Additional Results 147
A.2 Chapter 6 Additional Results 155

References **159**

List of Figures

2.1 A generalized marine food web showing the functional relationships between trophic levels where direction of links represents prey-predator interactions. 8

2.2 An example of a functional network model showing the functional relationships where direction of links (black) represents prey-predator interactions between fish species (sprat, herring and haddock) and a zooplankton species: *Calanus finmarchicus* and negative influence of external stressors- temperature and fisheries catch- pelagic trawlers (red) on the relevant species dynamics. 10

3.1 A simple Bayesian network over 4 variables. (a) shows the DAG component whilst (b) shows the conditional probability distribution attached to the node *Species 4*. T=True and F=False. 21

3.2 A Bayesian network representing a first-order Markov process. 29

3.3 (a) A 2TBN and (b) the same model unrolled for T=4 slices. 30

3.4 A dynamic BN where nodes represent variables at a point in time. . . . 31

3.5 A hidden Markov model, where the square nodes represent discrete hidden nodes and circle nodes represent observed nodes. 36

3.6 An autoregressive hidden Markov model, where H denotes an unmeasured hidden variable. 40

4.1 An autoregressive hidden Markov model, where H denotes an unmeasured hidden variable. 44

4.2 Regions of the three surveys (shaded area) corresponding to the three datasets: Georges Bank (GB), East Scotian Shelf (ESS) and the North Sea (NS). 46
4.3 The expected values of the discovered hidden variable from ARHMM (a), ARHMM + metrics (b) and mean variance (c) for ESS. The dashed line indicates the time of the regime shift in 1992. The solid line indicates upper and lower 95% confidence intervals, obtained from bootstrap predictions' mean and standard deviation. 55
4.4 Biomass predictions generated by ARHMM + metrics of cod (a), and silver hake (b) for ESS region. 95% confidence intervals report bootstrap predictions' mean and standard deviation. Dashed line indicates predictions by the model, whilst solid indicates standardised observed biomass for the time period 1980-2006. 56
4.5 The expected value of the discovered hidden variable from ARHMM (a), ARHMM + metrics (b) and mean variance (c) for NS. The dashed lines indicate the time period of the regime shift. The solid line indicates upper and lower 95% confidence intervals, obtained from bootstrap predictions' mean and standard deviation. 58
4.6 Biomass predictions generated by ARHMM + metrics of cod (a), and haddock (b) for NS region. 95% confidence intervals report bootstrap predictions' mean and standard deviation. Dashed line indicates predictions by the model, whilst solid indicates standardised observed biomass for the time period 1977-2009. 59
4.7 Diet matrix (a) with the network before (b) and after (c) the regime shift for ESS, generated by the data split using REVEAL. The width of edges corresponds to the computed confidence level (bold line: 0.5 and light line: 0.1). The squared nodes are significant themselves. For the diet matrix direction, of links represents predator-prey interactions. In bottom network (c): AP- American plaice and GA- Greater argentine. . 62

4.8 Diet matrix (a) with the network before (b) and after (c) the regime shift for NS, generated by the data split using REVEAL. The width of edges corresponds to the computed confidence level (bold line: 0.5 and light line: 0.1). For the diet matrix, direction of links represents predator-prey interactions. 64

5.1 Regions (shaded area) of the Gulf of St Lawrence (a) and the North Sea (f). 68

5.2 ICES statistical rectangles within the North Sea (areas 1 to 7 were used in this study). Source: ICES, Manual for the International Bottom Trawl Surveys. 70

5.3 General structural form of the HSDBN model. Solid line represents fixed edges across areas. The three spatial nodes (*P sp.*, *SP sp.*, *LP sp.*), *general HV*, *catch* and *temperature* are individually linked to either *P*, *SP* or *LP* (represented by the dotted surrounding), depending on the spatial area (grey line). Connectivity between *P*, *SP* and *LP* also differs spatially. Edges between nodes (or variables) represent dependence relationships. 77

5.4 Locations of the sampling stations before clustering (a) and after clustering in the region of St Lawrence(b). 78

5.5 The learned *I-P*, *P-SP* and *SP-LP* relationships for all 20 spatial clusters (size of scattered bubbles is equivalent to the estimated confidence by the hill-climb). The clusters mentioned in 5.4.1 are numbered. 80

5.6 The learned *I-SP*, *P-LP* and *SP-LP* relationships for clusters 5, 7 and 19 (represented by solid, dash and dot line respectively) for the time window: 2000-2013. 81

5.7 Biomass predictions generated by the spatial DBN for clusters 5 and 15 for the four trophic groups: *I*, *P*, *SP* and *LP*. Solid line indicates predictions and dash-dot line indicates standardised observed biomass. 95% confidence intervals report bootstrap predictions' mean and standard deviation. 84

5.8 Biomass predictions of *P*, *SP* and *LP* generated by the HSDBN for areas 3 (a,c,e) and 6 (b,d,f). Solid line indicates predictions and dotted line indicates standardised observed biomass. 95% confidence intervals report bootstrap predictions' mean and standard deviation. Note the negative scale is due to standardisation. 90

5.9 The expected value of the specific HV (solid line) for area 6 (a) and general HV for area 5 (b) generated by HSDBN with the observed standardised biomass for the zooplankton (dotted line). The solid line indicates upper and lower 95% confidence intervals, obtained from bootstrap predictions' overall mean and standard deviation. Note the negative scale is due to standardisation. 91

5.10 Estimated confidence by the hill-climb of each functional relationship occurring across the whole of North Sea: (a) *P*-*SP* between areas 1-3 (solid line), 1-7 (dashed) and 2-5 (dotted); (b) *P*-*LP* between areas 1-3 (solid), 2-3 (dashed) and 4-5 (dotted); (c) *SP*-*LP* between areas 1-6 (solid), 3-4 (dashed) and 4-6 (dotted). Note the time window starts from 1993 due to the windowing required during the hill-climb. 94

6.1 General structural form of the HSDBN model (a). Solid line represents fixed edges across areas. The spatial nodes (*P sp.*, *SP sp.*, *LP sp.*, *Fish species sp.*), *HV*, *catch* and *temperature* are individually linked to either *P*, *SP* or *LP* (represented by the dotted surrounding), depending on the spatial area (grey line). Connectivity between *P*, *SP* and *LP* and with the *fish species* also differs spatially. Network structure for area 4 (b) that models the dynamics of cod. The edges shown by a dotted line are defined by the expert. 101

6.2 An example matrix from a *Medium Fisheries Catch* scenario model with initial input used in model definition, the input during model testing and the generated output (a). The time window for each variable is shown in brackets. Note, the time window for the output starts from 1989. "[]" represents variables for which no evidence is introduced and which are predicted. "Z" stands for zooplankton. The observed cod catch prior to the scenario year of 1995 (solid line) and fixed catch level for the Medium Fisheries Catch scenario (dashed line) is shown in (b). 104

6.3 Recorded spatial cod data is shown in (a). The observed cod catch (live weight in tonnes) with the three fixed year levels of fisheries catch scenarios for the time window 1983-2015 is shown in (b). Recorded survey cod data (solid line) with the generated output by the *Historical* model (dotted line) for the time window 1989-2020 for area 4 (c). Recorded survey cod (solid line) with the modelled cod is shown in (d) under fisheries catch scenarios of high (black dashed line), medium (grey dashed line) and low (black dotted line) levels for the time window 1989-2020. 107

6.4 The model structure for area 3 that models the plaice dynamics is shown in (a). The dotted edges are defined by the expert. The observed plaice catch (live weight in tonnes) with the three fixed year levels of fisheries catch scenarios for the time window 1983-2015 is shown in (b). Recorded survey plaice data (solid line) with the generated output by the *Historical* model (dotted line) for the time window 1989-2020 is shown in (c). Note, the y-axis was cut off only for visual purposes. Recorded survey (solid line) with the modelled plaice data under fisheries catch scenarios of high (black dashed line), medium (grey dashed line) and low (black dotted line) levels for the time window 1989-2020 is shown in (d). 109

6.5 The observed whiting catch (live weight in tonnes) with the three fixed year levels of fisheries catch scenarios for the time window 1983-2015 is shown in (a). Recorded survey whiting data (solid line) with the generated output by the *Historical* model (dotted line) for the time window 1989-2020 is shown in (b). Recorded survey (solid line) with the modelled whiting data under fisheries catch scenarios of high (black dashed line), medium (grey dashed line) and low (black dotted line) levels for the time window 1989-2020 is shown in (c). 111

6.6 The network structure for area 6 that models the sole dynamics is shown in (a). The dotted edges are defined by the expert. The observed sole catch (live weight in tonnes) with the three fixed year levels of fisheries catch scenarios for the time window 1983-2015 is shown in (b). Recorded survey sole data (solid line) with the generated output by the *Historical* model (dotted line) for the time window 1989-2020 for area 6 is shown in (c). Recorded survey (solid line) with the modelled sole data under fisheries catch scenarios of high (black dashed line), medium (grey dashed line) and low (black dotted line) levels for the time window 1989-2020 is shown in (d). Note, the y-axis in (c) and (d) was cut off only for visualisation purposes. 113

6.7 The network structure for area 6 that models the haddock dynamics is shown in (a). The dotted edges are defined by the expert. The observed haddock catch (live weight in tonnes) with the three fixed year levels of fisheries catch scenarios for the time window 1983-2015 is shown in (b). Recorded survey haddock data (solid line) with the generated output by the *Historical* model (dotted line) for the time window 1989-2020 for area 6 is shown in (c). Recorded survey (solid line) with the modelled haddock data under fisheries catch scenarios of high (black dashed line), medium (grey dashed line) and low (black dotted line) levels for the time window 1989-2020 is shown in (d). 115

6.8 Modelled *Net PP* for areas 3 (a) and 6 (b), generated by the *Historical* model (solid line) and scenario of *T.I.* (black dashed line) for the time window: 1989-2020. Negative scale is due to standardisation. The vertical solid line indicates the year of divergence (1990), when we initiate the scenario conditions. 116

6.9 Modelled data by the *Historical* model (solid line) for *C.finmarchicus* for area 3 (a) and *C.helgolandicus* for area 6 (b) under scenarios of *T.I.* (black dashed line) and *Net.D.* (grey dotted line) for the time window: 1989-2020. Negative scale is due to standardisation. The vertical solid line indicates the year of divergence (1990), when we initiate the scenario conditions. 117

6.10 Modelled data by the *Historical* model (solid line) for herring (a) and *P* species group (b) for area 3 under scenarios from *Net PP* increase (black dashed line) and decline (grey dotted line) for the time window: 1989-2020. The vertical solid line indicates the year of divergence (1990), when we initiate the scenario conditions. 118

7.1 Locations of the spatial clusters for the Gulf of St Lawrence 123

7.2 The dynamic Bayesian network structure for cluster 2 from the Gulf of St Lawrence that was used to model the dynamics for the areas within the North Sea. 124

7.3 The spatial DBN structure for cluster 13 from the Gulf of St Lawrence that was used to model the dynamics for the areas within the North Sea. 125

7.4 The non-spatial DBN structures for clusters 11 (a) and 17 (b) from the Gulf of St Lawrence that were used to model the dynamics of species in the areas of North Sea. 126

7.5 Generated biomass predictions for species groups *P* (a), *SP* (b) and *LP* (c) for area 3 in the North Sea by the non-spatial DBN model from cluster 11 in the Gulf. Predictions are shown by the solid line and standardised observed biomass by the dotted line. 130

7.6 Generated biomass predictions for species groups P (a), SP (b) and LP (c) for area 5 in the North Sea by the non-spatial DBN model from cluster 1 in the Gulf. Predictions are shown by the solid line and standardised observed biomass by the dotted line. 131

7.7 Generated biomass predictions for species groups P (a), SP (b) and LP (c) for area 7 in the North Sea by the non-spatial DBN model from cluster 11 in the Gulf. Predictions are shown by the solid line and standardised observed biomass by the dotted line. 132

7.8 The expected value of the hidden variables (solid line) for areas 1 (a) and 4 (b), generated by the non-spatial DBN model from cluster 17 in the Gulf, with the standardised observed data (dotted line) for fisheries catch and zooplankton in the North Sea. The solid lines indicate upper and lower 95% confidence intervals. 135

7.9 The expected value of the hidden variables (solid line) for areas 4 (a) and 7 (b), generated by the spatial DBN model from cluster 11 in the Gulf, with the standardised observed data (dotted line) for fisheries catch and zooplankton in the North Sea. The solid lines indicate upper and lower 95% confidence intervals. 137

A.1 Observed standardised biomass of P, SP and LP and their predictions generated by HSDBN (a,c,e) and HDBN (b,d,f) for area 3. Note the negative scale is due to standardisation. Diagonal represents perfect prediction. 148

A.2 Observed standardised biomass of P, SP and LP and their predictions generated by HSDBN (a,c,e) and SDBN (b,d,f) for area 6. Note the negative scale is due to standardisation. Diagonal represents perfect prediction. 149

A.3 Biomass predictions of P (a), SP (b) and LP (c) generated by the HSDBN for area 1. Solid line indicates predictions and dotted line indicates standardised observed biomass. 95% confidence intervals report bootstrap predictions' mean and standard deviation. Note the negative scale is due to standardisation. 150

A.4 HSDBN network structure for areas 1 to 7 in the North Sea. *HV* stands for hidden variable. Spatial nodes are abbreviated as *P sp.*, *SP sp.* and *LP sp.* Edges between nodes (or variables) represent dependence relationships, the edges shown by a dotted line are defined by the expert. . 151

A.5 Learned trophic interactions *P-SP* (a), *P-LP* (b) and *SP-LP* (c) for all 7 spatial areas (numbered in figure). The estimated mean confidence by the hill-climb, for the time window: 1993-2014 (to recall, longer time series here), is shown above or below the links. Note the links represent dependence, not causality. The time window starts from 1993 due to the windowing required during the hill-climb. 152

A.6 The expected value of the specific HV (solid line) for areas 1 (a) and 7 (b) generated by HSDBN with the observed standardised biomass for the zooplankton (dotted line). The solid line indicates upper and lower 95% confidence intervals, obtained from bootstrap predictions' overall mean and standard deviation. Note the negative scale is due to standardisation. 154

A.7 The expected value of the general HV (solid line) for area 3 (a) and area 4 (b) generated by DBN with the observed standardised biomass for the *P* trophic group (dotted line). The solid line indicates upper and lower 95% confidence intervals, obtained from bootstrap predictions' overall mean and standard deviation. Note the negative scale is due to standardisation. 155

A.8 The modelled survey data, generated by the *Historical* model (solid line) and scenarios of *H.FC.*(black dashed line), *M.FC.* (grey dotted line) and *L.FC.* (black dotted line) for the time window: 1989-2020. The observed catch (tonnes live weight is shown next to the modelled predictions.) Note, the y-axis in some of the figures was cut off only for visual purposes. 157

A.9 The spatial values for plaice (a) and whiting (b) in area 3. 157

A.10 The expected value of the hidden variable for area 3 generated by the *Historical* model for whiting (dashed line) and from the scenario of Low Fisheries Catch (dotted line). The solid line indicates upper and lower 95% confidence intervals, obtained from bootstrap predictions' overall mean and standard deviation. 158

List of Tables

4.1 Wrapper feature selection results for GB region. 52

4.2 The functionally equivalent species to GB dataset for ESS and NS. These are each ordered based upon their relevance to species in Table 1. 53

5.1 Pre-defined Functional Relationships 72

5.2 Summary of the applied dynamic BN models on the North Sea. Models are ordered based upon complexity level in regards to number of hidden variables (HVs) and spatial nodes. 74

5.3 SSE of DBN and spatial DBN. 95% confidence intervals reported in brackets 83

5.4 Sum of Squared Error (SSE) of P, SP and LP biomass predictions generated by ARHMM (a), DBN (b), SDBN (c), HDBN (d) and HSDBN (e). The * indicates most accurate predictions for individual species groups. 87

5.5 Learned trophic associations and interactions with key stressors (*catch* and *temperature*) for each of the 7 spatial areas (the estimated mean confidence of each interaction, learned by the hill-climb for the time window: 1993-2010, is reported in brackets). The time window starts from 1993 due to the windowing required during the hill-climb. 92

7.1 Sum of Squared Error (SSE) of the overall trophic group species predictions for the North Sea, generated by the non-spatial and spatial DBNs from the Gulf of St Lawrence (GSL). The cluster from the GSL that produced the most accurate overall biomass predictions is shown in brackets. 127

7.2 Sum of Squared Error (SSE) of P, SP and LP predictions for the North Sea, generated by the non-spatial DBN from the Gulf of St Lawrence (GSL). The cluster from the GSL that produced the most accurate biomass predictions is shown in brackets. 128

7.3 Sum of Squared Error (SSE) of P, SP and LP predictions for the North Sea, generated by the spatial DBN from the Gulf of St Lawrence (GSL). The cluster from the GSL that produced the most accurate biomass predictions is shown in brackets. The * symbol indicates the clusters from the non-spatial DBN that also generated the most accurate predictions for the relevant species group within the North Sea. 129

7.4 Summary of the expected value of the hidden variable generated by the non-spatial DBN model from the Gulf of St Lawrence (GSL) for areas 1 to 7 in the North Sea. The table shows what the hidden variable is expected to model and the relevant summary statistics from the Mann-Whitney U test. 134

7.5 Summary of the expected value of the hidden variable generated by the spatial DBN model from the Gulf of St Lawrence (GSL) for areas 1 to 7 in the North Sea. The table shows what the hidden variable is expected to model and the relevant summary statistics from the Mann-Whitney U test. 136

A.1 Table that shows the influence of catch on different species groups and the relevant species of interest from that group in each area. 155

A.2 Table that shows the spatial neighbourhood of each area. 155

Glossary

- **Biomass:** the total quantity or weight of organisms in a given area or volume
- **Trophic:** relating to the feeding habits of different organisms in a food chain
- **Functional Relationships:** an interaction between two organisms of unlike species in which one of them acts as predator that captures and feeds on the other organism that serves as the prey
- **Food Web:** a network of all the predator-prey interactions in a community are interrelated
- **Pelagic:** (pelagic fish that live in the pelagic zone of ocean waters - being neither close to the bottom nor near the shore
- **Piscivorous:** fish-eating species
- **Invertebrates:** an animal lacking a backbone, such as an arthropod or mollusc
- **Plankton:** the small and microscopic organisms drifting or floating in the sea or fresh water
- **Zooplankton:** plankton consisting of small animals and the immature stages of larger animals
- **Regime Shift:** or functional collapse is an event during which the underlying food web structure dramatically changes irrevocably
- **Biotic:** relating to or resulting from living organisms
- **Abiotic:** physical rather than biological; not derived from living organisms

- **Net Primary Production:** net production of carbon by primary level producers such as phytoplankton

Key Acronyms

- **BN** Bayesian network
- **DBN** Dynamic Bayesian network
- **HMM** Hidden Markov model
- **ARHMM** Autoregressive Hidden Markov model
- **NS** North Sea
- **GB** Georges Bank
- **ESS** East Scotian Shelf
- **GSL** Gulf of St Lawrence
- **HV** hidden variable
- **SSE** Sum of Squared Error
- **Net PP** net primary production
- **P** pelagic
- **SP** small piscivorous
- **LP** large piscivorous and top predators
- **ICES** International Council for the Exploration of the Sea

Chapter 1

Introduction

1.1 Motivation

Some spectacular collapses in fish stocks have occurred in the past 20 years but the most notable is the once largest cod stock in the world, the Northern cod stock off eastern Newfoundland, which experienced a 99% decline in biomass (Tucker & Duplisea (2012)). Cod unfortunately, is not alone and there are other stocks of various species that have been reduced to only a small percentage of stock sizes in recent history. Some of these regions may have moved to an "alternative stable state" or experienced a "regime shift" and are unlikely to return to cod dominated community without some chance environmental event beyond human control (Gröger et al. (2011)). Much of this effect is due to direct mortality on fish through fishing and subsequent indirect effects or climate change affecting species recruitment and survival. Understanding multi-decadal fluctuations of fish populations and their variability related to regime shifts or other natural and anthropogenic disturbance is complicated because the abundance of fish populations is driven by both environmental forcing, fishing and species associations (Rothschild & Shannon (2004)). Therefore, in this thesis, a focus will be the use of state-of-the-art computational techniques based upon dynamic Bayesian networks with hidden variables to both integrate human expertise with extensive empirical data and model unmeasured factors to both predict functional changes in different fish communities and model their dynamics with interactions from the environment and anthropogenic forcing.

Understanding the causes and consequences of change in marine ecosystems is a fundamental requirement for effective marine management and the development of robust evidence based marine policies (Kenny et al. (2009)). It will allow precautionary management measures to be put in place as it is believed that changes in the marine environment are to become more rapid in the 21st century causing both ecological and industrial implications (Fernandes et al. (2013)). If we accept that ecosystem changes will occur in the future, then the challenge we face is how to predict such changes with sufficient advance warning. It is important to be able to understand the underlying species dynamics and ecological networks to enable any impacts from functional changes and/or future pressures to be managed and minimised in an effective way.

At present most of the management effort is directed at regulating activities and not mitigating for, or adapting to, inevitable and largely unmanageable changes. Many studies using different techniques have been undertaken to utilize environmental information and provide advice to meet management needs and understand future environmental states (Lewy & Vinther (2004); Ulrich et al. (2011); Mackinson & Daskalov (2007); Lynam & Mackinson (2015)). However such methods often limit the underlying interactions from expanding beyond the current food web paradigm. Although, the majority of traditional statistical approaches incorporate a large percentage of the higher trophic groups, they lack important extrinsic drivers, such as climate variation (for example Ecopath with Ecosim in Mackinson & Daskalov (2007)), which is fundamental for interpreting community dynamics. In addition, for such models to be valuable, they would also need to reflect the link between an input that can be managed (fisheries catch) and the response (e.g., change in species biomass), and therefore require an anthropogenic involvement (García-Carreras et al. (2015)). While such methods have been successful in providing insights into the high order dynamics and status of marine systems, the techniques are not (in themselves) suitable for testing different management scenarios or answering the "what if" type questions often asked by fisheries managers. Thus, it is clear that new assessment tools that consider multiple species interactions and their associations with external factors and can both predict and observe changes in pressure and state are needed. This would enable the strengths of the integrated ecosystem assessments to be combined with the scenario testing utility of traditional stock as-

sessment models and therefore allow predictions to be made about future significant changes in fish stock status.

Recently, an approach has been introduced to the area of biology, that is capable of inferring network structures, capturing nonlinear, dynamic and arbitrary combinatorial relationships: Bayesian Networks (BNs) (Heckerman et al. (1995)). BNs have been applied to reveal gene regulatory networks using gene microarray data (Friedman et al. (2000)) and were shown to reveal known pathways of neural information networks from brain electrophysiology data (Smith et al. (2006)). Such a flexible technique capable of identifying the complex relationships involved in bioinformatics potentially offers a valuable method in ecological studies (Milns et al. (2010)). Therefore, this work aims to adapt this novel methodology to infer the network structure directly from the collected field data.

The main aim of the research presented here is to develop dynamic models that adopt the functional network approach between species and their environment to model their dynamics allowing for predictions to be made across spatial and temporal scales. In this work, the importance of modelling unmeasured latent effects is recognised to predict functional changes and capture species variance that might not be purely constrained within the model structure. In addition, we aim at integrating the concept of functional equivalence between different datasets in order to generalise species interactions and the impact from natural and anthropogenic sources. Using a combination of structure learning techniques with some expert knowledge we introduce a model that can predict species dynamics and their interactions with external factors throughout time and within spatially differentiated scales. By applying this modelling approach, we want to understand what are the key mechanisms and influences when forming ecological networks but also to provide strategic advice on potential response of the system to pressure.

In this introductory chapter, the motivations, aims, and contributions of this thesis are fully explained.

1.2 Thesis contributions

The main contributions of this thesis are:

- **Functional Network Models** The concept of functional network models and their importance is defined in terms of modelling species dynamics and their interactions with natural and anthropogenic factors. Such models are developed to generalise between different species that have similar functions in different geographical regions in order to identify changes in underlying system structure and predict influence from natural and anthropogenic sources.

- **Exploration of Dynamic Bayesian Models with Hidden Variables** Dynamic models with hidden variables are used to predict changes in species trophic dynamics and variance. We explore the predictive qualities of recognised statistical metrics in combination with dynamic models to detect early-warning signals of a functional collapse.

- **Identification of Ecological Networks** We apply a structure learning algorithm to discover data-driven ecological interactions and associations that are used in combination with dynamic models to predict species biomass.

- **Hidden Spatial Dynamic Bayesian Network Model (HSDBN)** We develop a dynamic model that accounts for unmeasured latent effects and spatial autocorrelation to predict species dynamics and biomass across spatial and temporal scales. The proposed model is a data-driven functional network model that accounts for multispecies associations and interactions and their changes over space and time. It concerns the use of two hidden variables: one for modelling a missing variable and the other- for general overall systemic changes.

- **Exploration of the HSDBN model in combination with natural and anthropogenic scenarios** We extend the hidden spatial dynamic model to predict species dynamics and future biomass in combination with fisheries catch, temperature and productivity scenarios to explore species trends in response to change in pressures.

1.3 Thesis outline

This thesis is organized as follows.

Chapter 2 explores the state of the literature and the gaps to fill. It first explains the concepts related to marine food webs and highlights the importance of modelling functional relationships between species and their environment, using the functional network approach. Then, it moves onto emphasizing the advantages of using Bayesian networks for modelling ecological data by providing examples from the literature.

Chapter 3 explores the state-of-the-art concepts used for this work. It starts by explaining static Bayesian networks and standard techniques related to structure and parameter learning. Then, it focuses on dynamic Bayesian network models and specifically dynamic models with hidden variables and how to perform inference and learning for these methods.

Chapter 4 introduces the concept of functional equivalence and then describes the use of dynamic models with hidden variables in combination with statistical metrics to predict early-warning signals of functional changes and species biomass in different marine ecosystems.

Chapter 5 explores the use of structure learning and dynamic models to predict species biomass and functional relationships within the Gulf of St Lawrence. This chapter introduces a novel modelling approach to predict species dynamics and interactions with their environment within spatial and temporal scales across the North Sea.

Chapter 6 describes the use of the novel modelling approach to predict species trends in response to change in fisheries catch, temperature and productivity scenarios.

Chapter 7 explores the use of functionally equivalent trophic structures between species groups from different marine ecosystems to generalise influence from natural and anthropogenic sources.

Chapter 8 summarises the findings from this thesis, identifies advantages and disadvantages and explores future improvements and developments.

Chapter 2

Background

This chapter reports the state of the literature regarding modelling ecological data and specifically the analysis of functional relationships between species and their environment. The first part explains the definitions related to marine food webs and some of the threats that marine species are facing in recent years. It describes the use of functional network models to understand ecosystem structure and resilience.
The second part of this chapter highlights the advantages of using Bayesian networks for modelling ecological data by providing examples of different environmental studies and describing some of the characteristics of this approach in regards to predictive ecology.

2.1 Fisheries and Ecoinformatics

Owing to the explosion in data collection and storage made available since the dawn of parallel sequencing, there has been a demand for specialist techniques to analyse and model data (Tucker & Duplisea (2012)). The advance of research in fields such as machine learning (Bishop (2006)) and intelligent data analysis (Berthold et al. (2010)) has resulted in the development of many novel analysis tools in the area of bioinformatics (Tucker & Duplisea (2012)). Conversely, ecological data analysis has been rather much less explorative, however this is starting to change with the recent adoption of machine-learning approaches in predictive ecology.
Ecology was first defined in 1869 as the "study of the interaction of organisms with their environment" (Haeckel (1869)) and later as "the scientific study of the distribution and

abundance of organisms (Andrewartha & Browning (1961)). Krebs (1994) combined these definitions into the "scientific study of the interactions that determine the distribution and abundance of organisms". "Interactions" in the above definition refer to the interplay of organisms with biotic (of or relating to living organisms) and abiotic (physical e.g. temperature) factors (Begon et al. (1990)).

For over a century, most studies have been looking at patterns in the distribution and abundance of organisms, with respect to the study of interactions such as predator-prey (consumer-resource) relations, which has specifically received most extensive consideration (Olff et al. (2009)). However, recently, research has turned to better understanding and predicting the impacts of human induced climate change and activities such as commercial fisheries exploitation on the biodiversity and functioning of ecosystems. Specifically, some of the key questions raised are: 1) which are the most vulnerable species to extinction; 2) will the loss of some species affect other species and the structure of the ecosystem; 3) should some species be given special consideration in conservation schemes?

Although humans have been fishing for millennia, there has been a rapid development of industrialized fishing over the past 50 years, which has impacted the marine food webs through both the magnitude of the removals and the exploitation of species in waters far beyond traditional fishing grounds (Board et al. (2006)). Food webs (Fig.2.1) constitute functional relationships (such as predator-prey) by which energy and nutrients are passed on from one species to another. By removing both target and non-target species from the food web, fisheries directly influence fish abundance and ecosystem productivity, potentially altering the nature of the species interactions in marine food webs and communities.

In marine ecosystems, larger fish species or piscivorous fish (feeding on fish) have an important trophic (relating to feeding) role at or near the top of the food web, and as such they bring high economical value to the commercial fisheries (Jennings et al. (2001)). Consequently, piscivorous species (e.g. cod) become most susceptible to anthropogenic exploitation, leading to reduced abundance and consequently affecting species further down the food web (planktivorous fish) and causing alternative species to rise on top of the food web. The effect of predators (e.g. piscivorous species) has been shown to in-

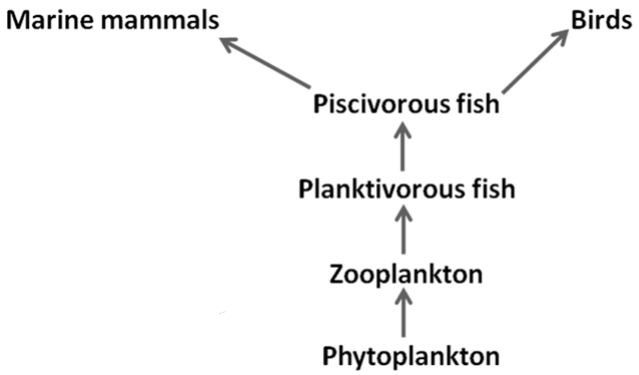


Figure 2.1: A generalized marine food web showing the functional relationships between trophic levels where direction of links represents prey-predator interactions.

fluence prey (e.g. planktivorous species) populations (a community of living organisms) and vice versa and has been described to be of the same or greater magnitude than fisheries removals in determining the dynamics of marine food webs (Tyrrell et al. (2011)). Planktivorous species (e.g. herring) are responsible for transforming the zooplankton (small animals) into food available to higher trophic level species. Zooplankton feed on phytoplankton (microscopic plants) which are the primary level produces, responsible for the production of carbon in the marine environment.

Human activities can also affect the sustainability of fish populations through the use of different management schemes (Jiao (2009)). Such changes have a direct effect on production and cause misunderstanding in management systems when it cannot be distinguished between climatic and harvesting impacts (Jiao (2009)). Subtle changes in key environmental variables like temperature, salinity, wind, ocean currents, can also significantly change the abundance, distribution and availability of fish populations, either directly or by affecting prey or predator populations (Jiao (2009)). For example, in the North Sea is now suggested that pelagic fish are partly regulated through bottom-up mechanisms as climate change has significant impacts on different phytoplankton and consequently zooplankton species through functional relationships (Kenny et al. (2009)).

Traditionally, fish stocks have been managed as separate units but recently the interest of linking stocks with other components of the ecosystem such as climate and fisheries

catch levels, has been considered as one aspect of an ecosystem-based approach to fisheries management (Kleisner et al. (2015)). The ecosystem-based approach to fisheries management acknowledges that fisheries are part of the environment and cannot be managed in isolation (Cury et al. (2005)) and requires recognition of the ecosystem dynamics and structure. This requires the need to understand indicators (e.g. species biomass and/or functional relationships between species) of ecosystem change that are easy to interpret in order to measure the impacts of fishing, climate change, and other factors across ecosystems and to provide management guidance at an ecosystem level (Kleisner et al. (2015)). Consequently, the need to examine indicators suggests the importance of applying a robust modelling technique to understand the ecosystem status and capture the dynamics of different components of the community.

2.2 Functional Network Models

Ecosystems consist of complex dynamic interactions among species and the environment, the understanding of which has significant ecological and industrial implications for predicting nature's response to changes in climate and biodiversity. Such biotic and abiotic interactions are further exacerbated by spatial and temporal variation of the system and its components. Direct observation and quantification of such relationships within real ecosystems may be beyond the scope of traditional fieldwork. Consequently, computational inference of ecological networks presents an alternate route to unravel ecosystem interactions.

Interactions among species make it difficult to predict how ecological communities will respond to environmental degradation, yet to do so we must understand the functional networks that form the systems (Dunne et al. (2002)). The functional network approach to understand community structure and resilience is an on-going approach combining known topological features of food webs with quantitative variation in species interactions with their environment and surrounding stressors to predict community stability. Different species may have similar functional roles (the functional status of an organism) within a system depending on the region. For example, one species may act as a predator of another which regulates a population in one location, but another species may

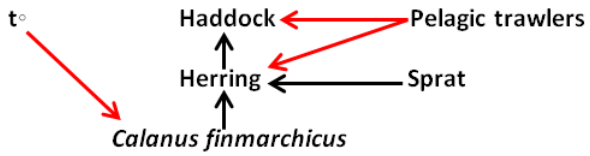


Figure 2.2: An example of a functional network model showing the functional relationships where direction of links (black) represents prey-predator interactions between fish species (sprat, herring and haddock) and a zooplankton species: *Calanus finmarchicus* and negative influence of external stressors- temperature and fisheries catch- pelagic trawlers (red) on the relevant species dynamics.

perform an almost identical role in another location. If we can model the function of the interaction rather than the species itself, data from different regions can be used to confirm key functional relationships, to generalise over systems and to predict impacts of forces such as fishing and climate change. Examining functional network models is necessary to help clarify ecological concepts and relationships and understand potential system responses to management changes, and evaluate alternative management policies (McCann et al. (2006)). Finally, applying the functional network approach provides a systematic way of presenting, describing and comparing the complexity of ecological communities, where species are represented by nodes and interactions by links between nodes.

One example of a functional network in a marine ecosystem is shown in Fig.2.2 which represents the species of interest e.g. herring and the relevant components (e.g. fisheries catch, temperature, predator-haddock, etc.) relating to the herring's dynamics. Therefore, when defining ecosystem dynamics, it is necessary to incorporate multispecies information and interactions with both physical and biological components that would reduce uncertainty in predicting the species response to change in fisheries and climate. The biological characteristics of any species stock are dependent upon and shaped over time by its interactions with other species and the rest of the ecosystem (Mackinson & Daskalov (2007)).

Ecosystem is a system that evolves both with time and space, thus the development of

spatio-temporal functional network models is needed in order to understand the mechanisms controlling ecosystem changes (Chen et al. (2011)). Real ecosystems are seen to be characterised with complex non-uniform dynamics, indicating that space and time are important domains of ecological models. Specifically, ecosystems exhibit spatial heterogeneity, forming spatial patterns of localised behaviour and species interactions that can be crucial to the overall ecological dynamics (Chen et al. (2011)). In this case, spatio-temporal models provide a powerful way to link fine scale local interactions to coarse-scale global dynamics to predict impacts from natural and anthropogenic sources and understand the underlying rules governing ecological dynamics.

Great variety of analytical models exist that describe the theory of ecological processes over space and time but there has been less attention to fitting such models to data that also describe the uncertainty associated with the data, models and parameters (Wikle (2003)). Traditionally, the generalised linear mixed model provides a stochastic modelling framework that captures complicated processes (McCulloch & Neuhaus (2001)). However, it is challenging to account for the joint spatio-temporal structure due to the complicated dynamic interactions across space and time (Wikle (2003)). Alternatively, the spatio-temporal dependence can be modelled by factoring the joint distribution of the ecological process into a series of conditional models (Wikle (2003)). These models are known as hierarchical spatio-temporal dynamic models (Wikle et al. (1998)) that take relatively simple spatial and temporal dependence attached to sub-processes in order to examine the joint spatio-temporal dependence (Wikle (2003)).

The concept of hierarchical modelling has been around for nearly 50 years but in the context of ecological modelling it is relatively new. In ecology, hierarchical modelling is based on the notion that complicated real-world processes can be modelled "in a conditional and probabilistically coherent framework that distinguishes the data, process, and parameter portions of models", according to Cressie & Wikle (2015). The method accounts for various sources of uncertainty and it is easy to implement in complicated high-dimensional settings (Wikle (2003)). In the Bayesian context, one can learn about the process and parameters given the data through: [Y,P|D] α [D|Y,P][Y|P][P], where "D" denotes data, "Y"- process, "P"- parameters and α denotes "proportional to" and the constant of proportionality ensures that the integral of the left-hand side over

the process and parameters is equal to 1 (from "Hierarchical Spatio-Temporal Models and Survey Research" by Wikle, Holan and Cressie in "Statistics Views", 2013, www.statisticsviews.com). Such a model is referred to as a Bayesian hierarchical model. This framework allows to incorporate spatial and temporal dependencies at the process-modelling stage of the hierarchy. In addition, distributions for the parameters can be specified allowing for complex dependence structures such as spatially and/or temporally dependant parameters ("Hierarchical Spatio-Temporal Models and Survey Research", 2013). Finally, this approach allows focus on the process of interest (Y) separately from the data which leads to including expertise when specifying the model and accounting for multiple sources of information about the same process ("Hierarchical Spatio-Temporal Models and Survey Research", 2013).

Conventional statistical analyses are inadequate in understanding the underlying mechanisms that link climate, ocean and biotic systems across spatial and temporal scales within an ecosystem. Traditional approaches based on system dynamics manage to represent changing behaviours in time but fail to represent spatial processes (Chen et al. (2011)). Geographic information system and remote sensing are some examples for spatial analysis but they represent a static view of the environment, failing to capture the dynamics of a system (Engelen et al. (1997)). Approaches such as the Smooth Transition AutoRegressive (STAR) models have been used to test spatial heterogeneity and non-linear time series but have failed to account for the associated uncertainty when multiple variables and factors are involved (Pede et al. (2014)). Therefore, allowing for an explorative systems level approach to provide a probability of ecological responses is required. Such approach is important as it would let people to ask the "what if" type questions, creating different management scenarios in the benefit of marine resources sustainability, and creating a balance between anticipating functional changes and the regulation of human activities.

2.3 Bayesian Networks for Modelling Ecological Data

Bayesian networks (BNs) are a relatively recent modelling approach in the area of biology that is capable of inferring network structures, capturing nonlinear, dynamic and

arbitrary combinatorial relationships (Heckerman et al. (1995)). Formerly, BNs have been applied to reveal gene regulatory networks using gene microarray data (Friedman et al. (2000)). Therefore, being able to identify the complex relationships involved in bioinformatics, the work presented onwards in this thesis, aims to adapt this novel methodology to infer the network structure directly from the collected field ecological data.

Bayesian networks are models that graphically and probabilistically represent relationships among variables. They are becoming increasingly popular for modelling uncertain and complex domains such as ecosystems and environmental management (Varis et al. (1990); Lee & Rieman (1997); Marcot et al. (2001); Uusitalo et al. (2005)). There is some research into using BNs for modelling fish populations (Hammond & O'Brien (2001); Marcot et al. (2001); Lee & Rieman (1997)). Other environmental topics include habitat suitability models, risk assessments, management evaluation, decision support and ecosystem services modelling (e.g. Chen & Pollino (2012); Haines-Young (2011); Smith et al. (2007); Tattari et al. (2003)). Lately, the BN methodology has been expanded to model the social dynamics of fishermen (Haapasaari et al. (2007); Haapasaari & Karjalainen (2010)). Most environmental BN applications often include decision analysis and encoded expert knowledge (Uusitalo (2007)). For example, Varis et al. (1990) developed a Bayesian network with decision and utility variables (an influence diagram) to evaluate management strategies of a eutrophic lake. Kuikka & Varis (1997) estimated the effect of climatic change on watersheds, and Marcot et al. (2001) combined some data with their mostly expert-opinion-based analysis of population viability. Uusitalo et al. (2005) used BNs to obtain an estimate on the maximum salmon smolt production capacity in the Baltic Sea, which is an important management variable on which no actual data exist.

As applied in ecology, the BNs represent an approach of analysing multiple associations between groups of species and their environment which is a more comprehensive route to revealing interactions within the ecosystem (Aderhold et al. (2012)) directly from the data, rather than taking an "existing" network structure and analysing it in terms of summary statistics. BNs are efficient in integrating variables presented at different scales (Wooldridge (2005)), allow empirical data to be combined with existing knowledge

(Uusitalo (2007)), operate within a data poor environment (Uusitalo (2007)), integrate the uncertainty associated with species dynamics due to the action of multiple driving factors and they can be easily combined with decision analytic tools to aid management (Marcot et al. (2001)). These features make BNs attractive for modelling environmental systems where uncertainty is pervasive (Hamilton et al. (2015)).

What makes BNs attractive to environmental management is that they represent probabilistic dependencies among species and ecosystem factors that influence the variables likelihood in an intuitive, graphic form (Jensen (2001)), therefore the manager has a quantitative indication of the range of possible scenarios consistent with the data to enable management advice (Cooke (1999)). The use of BNs methodology facilitates the communication of modelling results and the representation of a variety of perspectives as a means of modelling likelihoods of natural and anthropogenic effects (Levontin et al. (2011)). By having the ability to demonstrate graphically how assumptions affect the probability of outcome, BNs aid improving communication among ecologists, decision-makers, and stakeholders who may lack formal training in the underlying scientific disciplines (Kuikka et al. (1999)). Additionally, by incorporating nodes in their graphical structure, BNs can investigate potential management decisions as they calculate and update probabilities of conditions and outcomes, providing managers with alternative solutions or strategies (McCann et al. (2006)).

Essentially, BNs serve as tools to quantify the relationships between ecological-response variables such as species biomass, distribution and environmental predictor variables such as their habitat and/or fisheries exploitation, climate. Probabilistic methods such as BNs provide estimates of the uncertainty associated with predictions, as demonstrated by Fernandes et al. (2010). With the recent adoption of BNs in predictive ecology, few assumptions can be made about the data and complex, spatially varying interactions can be recovered from collected field data, as demonstrated by Trifonova et al. (2015). In BN modelling, some sources of uncertainty include: lack of clear understanding of the system and processes within, missing and/or biased data (Hamilton et al. (2015)). The way BNs manage with such uncertainties is through developing alternative model structures and accounting for these uncertainties in the distribution of probabilities across the BN nodes and their states (Chen & Pollino (2012)). The

approach of creating multiple models, which reflect the different hypothesis of the system, deals with structural uncertainties and accepts that several different models can adequately represent the same system albeit from different perspectives (Krueger et al. (2012)). A fundamental advantage of Bayesian inference is that it allows for the use of informative prior distributions on uncertain model parameters (Uusitalo et al. (2012)). For example, literature data may be used to define priors, with study-specific data being used to update those priors.

2.4 Integration of Expertise and Multiple Data Sets

A practical approach to overcome problems derived from low data quality, noise and measurement errors typical during the collection of ecological datasets is to incorporate existing knowledge into a computational framework. This way statistical inference can increase the knowledge in the areas that are still lacking evidence and help construct more precise models. Expertise is often applied to develop BN modelling structure. Expert knowledge can be combined with data regarding variables on which different levels of accuracy (e.g. absence/presence and quantity data) exist (Marcot et al. (2001)). Specifically, the expert judgement can be used to parameterize the BN models, which is of significant importance for environmental studies which do not have enough quantitative data for use by statistical modelling approaches (Smith et al. (2007)). The use of expert knowledge is especially advantageous when the data are scarce or missing, though this should be incorporated with caution as often expertise can be biased in the decision making (Morgan & Henrion (1990)). However, expert bias can be reduced by using rigorous methods for eliciting and applying expert judgement (Krueger et al. (2012)). Finally, expertise is used to assist selecting variables and measuring uncertainties associated with those variables (Marcot et al. (2001)).

The number of experts formally consulted in environmental modelling studies varies (Krueger et al. (2012)). BNs constructed for water quality and watershed management have been conceptualised with the inputs from 3 to 6 experts (Marcot et al. (2001); de Santa Olalla et al. (2007); Lynam et al. (2010)), while expert numbers increase when the models were constructed and developed during workshops (Vilizzi & Copp (2013)).

Generally, a participatory approach with workshops including experts, stakeholders and final users has been used to construct and develop BN models (Ticehurst et al. (2007); Inman et al. (2011); Richards et al. (2013)), whilst some other studies have validated the structure using questionnaires (de Santa Olalla et al. (2007); Franco (2014)).
Another way to incorporate expertise is by integrating prior knowledge into the system. Priors reflect our knowledge of the subject before new data is collected, and can be highly informative, when there is a lot of knowledge about the subject already, or very uninformative, if not much is known (Uusitalo (2007)). These priors are then updated with data, to obtain a synthesis of old knowledge and new data, which can then be applied as a prior in a different study (Uusitalo (2007)). In this way, the approach of scientific learning is explicit, providing the opportunity for assumptions made by the experts to be openly discussed. The use of prior knowledge reduces the need for large datasets and most importantly, the use of BN priors also provides the opportunity to identify key gaps in the ecosystem knowledge and thus inform future research priorities.
Another way to overcome noisy data issue is to integrate heterogeneous data sources to improve the reliability and accuracy of the modelled results and predictions. BNs can be informed simultaneously with different types of data and sources, quantitative and qualitative data (McCann et al. (2006)). For example, these can include datasets from field monitoring or laboratory studies; process equations, derived from peer-reviewed studies; datasets, derived from models (Chen & Pollino (2012)). Some example environmental studies include Wooldridge & Done (2004) who predicted coral bleaching in the Great Barrier Reef using a BN based on various data (e.g. temperature, reef shape, bathymetry). They used a BN as a framework to investigate dependencies among potential drivers and responses to ultimately update prior beliefs. Another example related to reef studies comes from Franco (2014) in which multi-scale scenario-based analyses, conducted for two different geographical reef datasets, quantified the effects of multiple stressors (e.g. anthropogenic and climatic stressors) on the reefal components, providing information on the actual state and possible future state of the ecological framework. Milns et al. (2010) evaluated the potential usefulness of BN algorithms by applying inference to avian count and habitat data. They examined properties of the revealed ecological functional networks and evaluated them against known features

of the ecosystem. Hamilton et al. (2015) used a BN approach to develop a habitat suitability model. Expert knowledge was elicited through surveys designed to describe the strength of relationships between the habitat suitability of the species of interest-crayfish and three other environmental variables. A series of 18 alternative models were developed based on the same model structure but parameterised using different sources expert judgement, field data or a combination of the two.

2.5 Summary

In this chapter, we explained the main definitions relating to food webs and functional relationships. We described the impact of fisheries on the marine food webs, highlighting the importance of ecosystem-based approach to fisheries management to better understand ecosystem structure and dynamics. Additionally, we explained the functional network approach as a potential method to examine community structure and resilience and the recent adoption of machine-learning techniques such as BNs that are becoming increasingly important in predictive ecology to understand ecological relationships in response to anthropogenic exploitation and climate change. We highlighted the recent use of BNs in the field of environmental studies, providing some specific examples in the area of fisheries. Also, we described some of the BN characteristics which have assisted in the application of this approach to become so advantageous in the area of ecological modelling, specifically outlining the ability of BNs to integrate multispecies information and interactions with both physical and biological components throughout space and time. Conventional statistical studies have failed to link response variables to multiple stressors through an ecosystem. Therefore, the need to adopt a more explorative systems level approach such as BNs capable of capturing the underlying mechanisms in a complex dynamic environment is needed. Hence, in this research we want to develop an approach to robustly identify key functional interactions that are of importance to the underlying ecosystem and its components in order to understand the ecological dynamics and structure and predict the long-term ecosystem change in response to anthropogenic disturbance and climate.

In the next chapter, we explore in detail some of the modelling techniques to build

functional network models that will be used later.

Chapter 3

Preliminaries

In this chapter, we illustrate the fundamentals of the Bayesian network framework. As discussed in the previous chapter, they may be particularly appropriate for investigating many questions related to modelling ecological data. We also illustrate an extension of static BNs- dynamic Bayesian networks for modelling temporal data and specifically focus on describing the characteristics of a type of a DBN- hidden Markov model. Finally, we explain the use of a hidden variable and how it can be used when modelling fish population data.

3.1 Bayesian Networks

Bayesian Networks (BNs) (Nielsen & Jensen (2009); Friedman et al. (2000)) use graph theory in combination with statistics to capture relationships or dependencies among independent variables. The dependencies are qualitatively represented through graph-based structures (networks) while the strength of the relationships is quantitatively shown by conditional probability tables (CPTs) for continuous or discrete data. The graph-based structure in combination with the CPTs provides ecologists and other expertise with a user-friendly framework to communicate the results (Chen & Pollino (2012)). The graphical structure of BNs is particularly convenient when we aim in describing an ecological network to model all the interactions between species and their environment.

A BN is defined as a probabilistic graphical model that encodes a joint probability

distribution over a set of variables, $X_1 ... X_N$ by exploiting conditional independence relationships, represented by a directed acyclic graph (DAG) (Friedman et al. (1999)). In the DAG, the nodes are the variables and the edges (or links) represent the conditional relationships between them. It is relevant to think of the BN as a "graph", describing groups of species as the "nodes" within the graph, and interactions as the "edges" that join the nodes (Faisal et al. (2010)). The edges are connected from a parent node to a child node, and the lack of possible links in the DAG encodes conditional independences. Conditional independence between the random variables X and Y given Z is defined as

$$p(X|\, Y, Z) = p(X|Z) \tag{3.1}$$

Therefore, once the value of Z is known, X and Y are independent.

The DAG represents the set of conditional independence relationships, which are explained by the directed separation (*d-separation*) criterion (**?**). Given three subsets of disjoint nodes: V_1, V_2 and V_3, it is V_3 that "d-separate" V_1 from V_2 if among the arcs between V_1 and V_2, there is a node V for which one of the following applies:

- v has converging arcs (all the arcs from the adjacent nodes point to v) and none of v or its children are in V_3;
- v is in V_3 and does not have converging arcs.

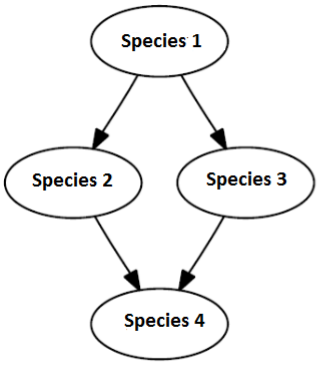
In general, a BN can encode a set of conditional independence relationships between variables, and it can be shown that a node is conditionally independent of its non-descendants given its parents (**?**), which defines the *Markov property* of a BN. The Markov property of BNs allows showing the global distribution of a network G as the product from the local distributions (CPDs) associated with each variable X_i. In a BN, a node is conditionally independent of all the other nodes in the graph given its *Markov blanket*, which is defined as the node's parents, children and children's parents. Given that, the CPD associated with each variable (either continuous or discrete) X encodes the probability of observing its values given the values of its parents, and is described

in the CPT and all the CPTs in a BN together provide an efficient factorisation of the joint probability:

$$p(\mathrm{x}) = \prod_{i=1}^{n} p(x_i|\mathbf{pa_i}) \tag{3.2}$$

where $\mathbf{pa_i}$ are the parents of the node x_i (which denotes both node and variable).

A simple example of a generic Bayesian network is shown in Figure 3.1 (the example is adapted from Murphy (2001)). The DAG (Fig.3.1a) shows 4 nodes, each representing a variable called *Species* with discrete values. The links between the nodes indicate that *Species 1* directly influences *Species 2* and *Species 3* which in turn influence *Species 4*. The conditional probability table (Fig.3.1b) quantifies the strength of each link. Although BNs are defined in the aspect of conditional independence, it does not necessarily mean that the arcs should represent cause-effect relationships. In this research, the links between variables represent dependence, which are predictive in an informative, not causal aspect (Milns et al. (2010)).



(a) DAG

Sp2	Sp3	p(Sp4=F)	p(Sp4=T)
F	F	0.9	0.1
T	F	0.2	0.8
F	T	0.2	0.8
T	T	0.1	0.9

(b) Conditional probabilities for Species 4

Figure 3.1: A simple Bayesian network over 4 variables. (a) shows the DAG component whilst (b) shows the conditional probability distribution attached to the node *Species 4*. T=True and F=False.

3.1.1 Bayesian Inference

Once the CPTs are defined, another useful aspect of BNs is to examine inference. Bayesian inference also called probabilistic reasoning or belief updating allows to determine the state of a set of variables given the state of others as evidence. Each node in the DAG is characterised by a state which can change depending on the state of other nodes and information about those states propagated through the DAG. By using this kind of inference, one can change the state or introduce new data or evidence (change a state or confront the DAG with new data) into the network, apply inference and inspect the posterior probabilities (Koller & Friedman (2009)). Given a BN B with graphical structure G and parameters Θ, we want to analyse the effect of a new evidence $\mathbf{E}$ on the distribution of $\mathbf{X}$ using the knowledge encoded in the BN, which means to analyse the posterior distribution: P($\mathbf{X}$|$\mathbf{E}$,B) = P($\mathbf{X}$|$\mathbf{E}$,G,Θ), which represents the distributions of the variables given in the observed evidence.

BNs use Bayes rule to perform inference (Murphy (2001)). Using Bayes rule, the posterior probability can be calculated $P(A|B)$, based on the values $P(A)$- the prior probability, $P(B|A)$- the likelihood and $P(B)$. If we want to calculate $P(A|B)$ given the observed data for $P(B|A)$, Bayes rule is:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{3.3}$$

Using Bayes rule together with the joint distribution described by a BN, we can infer the probability distribution on the value of a target variable, given the observed data of other variables in the network. For example, in the network from Fig.3.1, if it is observed that *Species 4* is True, *Sp4*=*T*, and we want to find the probability that it is *Species 3* is True, *Sp3*=*T*, then by Bayes rule:

$$P(Sp3 = T|Sp4 = T) = \frac{P(Sp4 = T|Sp3 = T)P(Sp3 = T)}{P(Sp4 = T)} \tag{3.4}$$

$$= \frac{\sum_{sp1,sp2} P(Sp1 = sp1, Sp2 = sp2, Sp3 = T, Sp4 = T)}{\sum_{sp1,sp2,sp3} P(Sp1 = sp1, Sp2 = sp2, Sp3 = sp3, Sp4 = T)} \tag{3.5}$$

$$= \frac{0.4138}{0.6125} \tag{3.6}$$

$$= 0.6755 \tag{3.7}$$

Similarly, if we want to find the probability that *Species 2* is True, *Sp2=T*, given that *Species 4* is True, *Sp4=T*:

$$P(Sp2 = T|Sp4 = T) = \frac{P(Sp4 = T|Sp2 = T)P(Sp2 = T)}{P(Sp4 = T)} \tag{3.8}$$

$$= \frac{\sum_{sp1,sp3} P(Sp1 = sp1, Sp3 = sp3, Sp2 = T, Sp4 = T)}{\sum_{sp1,sp2,sp3} P(Sp1 = sp1, Sp2 = sp2, Sp3 = sp3, Sp4 = T)} \tag{3.9}$$

$$= \frac{0.2938}{0.6125} \tag{3.10}$$

$$= 0.4796 \tag{3.11}$$

In this case, $P(Sp3{=}T|Sp4{=}T) > P(Sp2{=}T|Sp4{=}T)$ so it is more likely that *Species 3* has a bigger influence on *Species 4*.

Exact inference in BNs can be a complicated task. Therefore many alternative exact and approximated methods are often used in practice, for example some methods include: Monte-Carlo sampling and variational methods. In this work, the *junction tree* inference algorithm was applied, which uses a greedy search procedure to find a good ordering for variable elimination. (Murphy, 1998). The research presented in later chapters uses the Bayes Net Toolbox (Murphy et al. (2001)) to implement the junction three engine for the prediction of node values in BNs.

3.1.2 Parameter Estimation

The aim, now, is to calculate the parameters of a distribution (or a set of distributions, depending on the dimension of the variables) for each node in the network and for every configuration of its parents. Given a probability distribution X and a dataset $D = x_1,x_2,\ldots,x_n$, we want to learn a set of parameters Θ for X that maximizes the likelihood $L(\theta)$ that the data D comes from X. We need to compute the posterior distribution of the parameters $p(\theta|\boldsymbol{D})$, thus this is often problematic for high dimensional data and approximate methods are applied. For example, the parameters θ can be calculated by maximising the posterior probability of the parameters given the data D:

$$\hat{\theta} = \arg\max_{\theta} p(\theta|\mathbf{D}) = \arg\max_{\theta} p(\mathbf{D}|\theta)p(\theta) \tag{3.12}$$

However, this estimation applies only when there are no missing data. The latter equation can be simplified by maximising the marginal likelihood function, i.e. $\arg\max_{\theta} p(\mathbf{D}|\theta)$

given the uniform prior distribution of the parameters. Often, the log-likelihood is considered so that the log-likelihood of the dataset D_i, $i = 1...M$, is a sum of terms, one for each node k:

$$L_k = \frac{1}{M} \sum_{i=1}^{M} logp(x_k|pa(x_k), D_i) \tag{3.13}$$

When the data are missing and/or incomplete, the posterior distribution becomes intractable as the log-likelihood is not decomposable and approximation methods such as Monte-Carlo, Gaussian approximation and the Expectation Maximisation (EM) algorithm need to be used (Dempster et al. (1977)). In this work, we focus on using EM which is an iterative procedure in two steps that aims to estimate the parameters. In a BN, this extends to including unobserved or hidden variables, which are variables for which no instances of the dataset are specified. In the first step of the EM, the hidden values for the unobserved data are inferred using the applied model structure, while in the second step the estimated likelihood function is maximised. When the algorithm converges to a local maximum, the parameters are estimated. Hidden variables will be discussed later on in the chapter.

3.1.3 Structure Learning

The network structure can be imposed by the expert (who is constructing the model architecture) or learned from the data. In the case of using data, structure learning is a *NP-hard* problem because the number of DAGs on N variables is super-exponential in N. Some methods to look for local optima in the structure space include score-based learning algorithms, for example hill-climbing. Score-based approaches consist of forming a set of possible network structures, each represented by a score of how well it fits the data and then select the one with the highest score. Specifically, the hill-climb belongs to the family of local search which begins with an empty network. In each stage of the search, networks in the current neighbourhood are found by applying a single change to a link in the current network such as *add* arc, *delete* arc or *reverse* arc and choose the one change that improves the score the most. The hill-climb can be performed with a random restart which conducts several hill-climbing runs, perturbing the result of each one as the initial network for the next (Bouckaert (1995)). At each

iteration, a score is computed to determine if the new network fits the data better than the previous one. The algorithm stops when there is no more improvement and the final network is selected. The Bayesian Information Criterion (BIC, Schwarz et al. (1978)) is used for scoring candidate networks. The BIC is a combination of a log likelihood model and a penalization term which penalizes complicated models against simpler ones:

$$BIC = \log P(\Theta) + \log P(\Theta|D) - 0.5k\log(n) \tag{3.14}$$

where Θ represents the model, D is the data, n is the number of observations (sample size) and k is the number of parameters. $\log P(\Theta)$ is the prior probability of the network model Θ, $\log P(\Theta|D)$ is the log-likelihood while the term $k \log(n)$ is a penalty term, which helps to prevent over-fitting by biasing towards simpler, less complex models. Another option to calculate the score is to use the Akaike Information Criteria (AIC) (Akaike (1974)) which illustrates a penalization term of $2k$ instead of $0.5k\log(n)$. In this thesis, we apply the BIC as being more suitable because it takes into account the number of observations (n). Other score-based approaches include simulated annealing (Bouckaert (1995)) and genetic algorithms (Larranaga et al. (1997)).

3.1.4 Model Evaluation

After learning a BN, we want to evaluate the performance of the model. We present two methods which are used to evaluate the research in this thesis.

The first method makes a comparison of the network structure to documented knowledge. Specifically, in this thesis we compare interactions learned by the BN model to documented interactions by the expert which have been confirmed, usually through biological experiments from collected field data. A "true" network is formed from modelled interactions, where a directed edge $X \rightarrow Y$, between two species X and Y , exists if a confirmed species interaction between that pair of species can be found through expertise knowledge. The learnt BN structure can be compared to the true network in terms of true and false positives and negatives. A *true positive* (TP) is an edge that is present in both the learnt and true networks. A *false positive* (FP) is an edge that is present in the learnt network but not in the true network. A *false negative* (FN) is an edge that is in the true network but not in the learnt network, whilst a *true negative*

(TN) is an edge that is not in the true or learnt network. In terms of directionality of the edges in the learnt network, we assume dependence between the nodes or variables, so unless stated otherwise, the direction of an edge in the learned network is not used to represent causality but the equivalent classes of each identified link or edge.
Another method of BN performance evaluation is the prediction of node values. In the case of BN that represents a functional network of species interactions, we predict species biomass (more on prediction further below).

3.2 Introduction of Time Series, State-space Models and Autoregressive process

A time series is a chronological sequence of observations on a variable of interest (Xuan (2004)). A mathematical description of the time sequence could be a sequence of random variables $\{x_t | t \epsilon T\}$, where T is an index set of integers, e.g. {1,2,3..}. The distribution of this sequence is specified by the joint distribution of every finite subset of $\{x_t | t \epsilon T\}$ where $\{x_{t1}, x_{t2},...,x_{tk}\}$ for all integer k (Xuan (2004)).
A time series is said to be *stationary* if the distribution of $\{x_{t1}, x_{t2},...,x_{tk}\}$ is the same as the distribution of $\{x_{t1+h}, x_{t2+h},...,x_{tk+h}\}$ for all choices of $\{t_1,t_2,...,t_k\}$ and h such that $t_1,t_2,...,t_k \epsilon T$ and $t_1+h, t_2+h,...,t_k+h \epsilon T$ (Xuan (2004)). In another way, a time series is said to be stationary if there are no systematic change in the mean and variance. A time series which is not stationary, is called *non-stationary*. One way to capture periodic non-stationarities is to add extra hidden nodes to represent the current "regime", thereby creating mixtures of models (Murphy (2002)), this will be addressed in more detail further below.
A simple example of a stationary time series is a white noise series. $\{u_t \mid t \epsilon T\}$ is a collection of identical-distributed and mutually independent random variables with common mean zero and constant variance σ_2 (Xuan (2004)). Actually, white noise time series is purely random and it is often included in more complicated probabilistic models as the random error.
Let $\{u_t \mid t \epsilon T\}$ be a white noise process with mean zero and variance σ_2. A process

$\{x_t \mid t \epsilon T\}$ is said to be an autoregressive time series of order p: AR(p), if

$$x_t = \delta + \beta_1 x_{t-1} + \beta_2 x_{t-2} + ... + \beta_p x_{t-p} + u_t \tag{3.15}$$

where $\{\beta_i\}$ are constants (Xuan (2004)). The AR format is similar to the multiple regression model. Specifically, the prefix "auto" comes from the fact that x_t is regressed on the past values of itself.

Let us look at an example of an AR process with δ=0 (first-order case) which becomes:

$$x_t = \beta x_{t-1} + u_t \tag{3.16}$$

When $|\beta| = 1$, x_t is called a random process and then:

$$x_t = x_0 + \sum_{i=1}^{t} u_i \tag{3.17}$$

It follows that $Var(x_t) = Var(x_0) + t\sigma_2$. As the variance changes with t, the process becomes non-stationary (Xuan (2004)).

When $|\beta| > 1$, the random term u_t will eventually disappear and the equation becomes:

$$x_t = \beta x_{t-1} \tag{3.18}$$

and the process will follow a non-stationary deterministic path.

Only when $|\beta| < 1$,

$$Var(x_t) = \frac{\sigma_2}{1 - \beta^2} \tag{3.19}$$

the process is stationary.

In this thesis, the data is a time-series, generated by a dynamic system. One may be interested either in online analysis, where the data arrives in real-time, or in offline, where the data has already been collected (Murphy (2002)). Here, we are interested in the offline analysis, where one common task is to predict future observations given all the observations up to the present time, which we will denote by $y_{1:t}$= (y_1,...,y_t). We are genuinely not sure about the future so we would like to compute a best estimate. Additionally, we would like to know how confident is the best estimate. Hence, we will try to compute a probability distribution over the possible future observations, which we denote by: $P(y_{t+h}|y_{1:t})$, where $h > 0$ is the horizon, meaning how far into the future we want to predict. Specifically, we are interested to predict future outcomes as a

function of the inputs. Let $u_{1:t}$ denote our past inputs and $u_{t+1:t+n}$ denote our next n inputs, which leads to compute $P(y_{t+n}|u_{t+1:t+n}, y_{1:t})$.

"Classical" approaches that predict time-series data include linear models such as ARMAX or non-linear models such as neural networks. However, in this thesis we apply probabilistic models (or generally, *state-space models* that relate to modelling time-series data) because such techniques offer a natural mechanism for incorporating expert knowledge relating to the network structure. They also allow predictions to be made across very different platforms and organisms (Tucker & Duplisea (2012)) through the use of a network structure and inference that allow us to ask "what if ?" type questions of the data. They can handle multi-dimensional (multi-variate) inputs and/or outputs. Finally, state-space models allow for hidden variables to be included in the model, these are variables that we cannot measure (missing or unobserved) but we would like to estimate their state. By incorporating hidden variables into the model, we might be able to define structure much closer to the "true" structure of the domain we are modelling (Murphy (2002)).

In a state-space model, we can assume that there is some underlying hidden state that generates the observed data and that this hidden state evolves in time as a function of our inputs (Murphy (2002)). Here, the goal becomes to infer the hidden state given the observed data up to the present time. We will denote the hidden state as X_t at time t, which consequently defines computing: $P(X_t|y_{1:t},u_{1:t})$, this is called the *belief state*. Astrom (1965) has proven that the belief state is a sufficient statistic for prediction/control purposes, so we do not need to keep the past observations. We can update the belief state recursively using Bayes rule. Similarly to prediction, we have a probability distribution over X_t, rather than a best estimate, in order to efficiently reflect on our uncertainty about the "true" state of the world (Murphy (2002)).

We now discuss the more general state-space models, called dynamic Bayesian networks and one of the most common kind of state-space models: the hidden Markov model.

3.3 Dynamic Bayesian Networks

Modelling time series is achieved by using an extension of the BN known as the Dynamic Bayesian Network (DBN), where nodes represent variables at particular time slices. DBNs are directed graphical models of stochastic processes that characterise the unobserved and observed state in terms of state variables (as they include a temporal dimension), which can have complex interdependencies (Murphy (2002)). The DBN structure provides an easy way to specify such conditional independencies between the acting variables and hence to provide a compact parameterization of the complex real-time ecological data in this thesis. Also, DBNs allow expertise to monitor and update a system as time progresses and make robust predictions of the behaviour of the system (Mihajlovic & Petkovic (2001)). The terminology of static BNs: *nodes*, *edges* and *probabilities*, also applies to DBNs. Note that "dynamic" refers to modelling a dynamic system and not the assumption that model structure changes in time.

The states of a DBN satisfy the *Markovian condition* which is defined as: the state of a system at time t depends only on its immediate past, its state at time *t-1* (Mihajlovic & Petkovic (2001)). This property is often considered as a definition of *First order Markov property*: the future is independent of the past given the present (Murphy et al. (1999)) meaning each variable is directly influenced only by the previous variable (Fig.3.2)

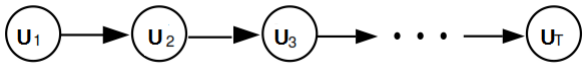


Figure 3.2: A Bayesian network representing a first-order Markov process.

A DBN is defined to be a pair (B_1, $B\rightarrow$) where B_1 is a BN with a prior $P(Z_1)$ and $B\rightarrow$ is a two-slice temporal BN (2TBN) which defines $P(Z_t|Z_{t-1})$ by means of a DAG:

$$P(Z_t|Z_{t-1}) = \prod_{i=1}^{N} P(Z_t^i|\mathbf{Pa}(Z_t^i)) \tag{3.20}$$

where Z_t^i is the i'th node at time t and $\mathbf{Pa}(Z_t^i)$ are the parents of Z_t^i in the graph (Murphy (2002)). The i'th node at time t could be a component of the input, hidden or output variables since: Z_t=(U_t,X_t,Y_t) which represent the input, hidden and output variables respectively. The nodes in the first temporal slice of a 2TBN do not have any

parameters associated with them, but each node in the second slice of the network has an associated CPD which defines $\mathrm{P}(\mathrm{Z}_t^i|\mathbf{Pa}(\mathrm{Z}_t^i))$ for all $t > 1$.

The semantics of a DBN can be defined by "unrolling" the 2TBN which provides T time slices (T is the length of the sequence) resulting in a joint distribution:

$$P(Z_{1:T}) = \prod_{t=1}^{T}\prod_{i=1}^{N} P(Z_t^i|\mathbf{Pa}(Z_t^i)) \tag{3.21}$$

An example of a 2TBN and an unrolled version for a sequence of length T=4 is shown in Fig.3.3. Here, $\mathrm{Z}_t^{(1)}$=U_t, $\mathrm{Z}_t^{(2)}$=Y_t, $\mathrm{Pa}(\mathrm{U}_t)$=$\mathrm{U}_{t-1}$ and $\mathrm{Pa}(\mathrm{Y}_t)$=$\mathrm{U}_t$, so the joint becomes:

$$P(U_{1:T}, Y_{1:T}) = P(U_1)P(Y_1|U_1)\prod_{t=2}^{T} P(U_t|U_{t-1})P(Y_t|U_t) \tag{3.22}$$

In addition, the parameters for slices t = t, *t+1*, .. do not change over time, i.e., the model is time invariant which allows unbounded amount of data to be modelled with a finite number of parameters (Murphy (2002)).

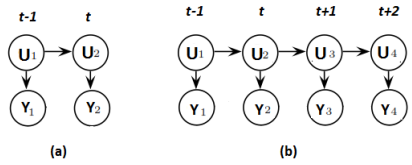


Figure 3.3: (a) A 2TBN and (b) the same model unrolled for T=4 slices.

We can see from Fig.3.3 that the DBN framework not only allows for connections within time slices: intra-slice connections (e.g. $\mathrm{U}_1 \rightarrow \mathrm{Y}_1$ at *t-1* Fig.3.3a) but also for connections between time slices (inter-slice connection e.g. $\mathrm{U}_1 \rightarrow \mathrm{U}_2$ from *t-1* to t respectively Fig.3.3b) which incorporate condition probabilities between variables from different time slices (Mihajlovic & Petkovic (2001)). In this way, we extend the first-order Markovian condition by allowing for higher order interactions between variables. For example, a τ^{th}-order model allows arcs from $\{\mathrm{U}_{t-\tau},...,\mathrm{U}_{t-1}\}$ to U_t. Another example of a DBN is shown in Fig.3.4 where the network includes only inter-slice connections between variables. Not necessarily all the states of a DBN need to be observed. States that are not directly observed (hidden states) may influence other variables that we can measure or calculate. This is another way of extending beyond the first-order models: define observations to be dependant on a hidden variable. So overall, in a DBN, each state at time t

may depend on states at time *t-1* and/or on current states of some other nodes at time *t*.

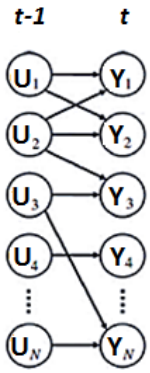


Figure 3.4: A dynamic BN where nodes represent variables at a point in time.

3.3.1 Learning and Inference

The techniques for inference and learning DBNs are mostly straightforward extensions of the techniques for inference and learning static BNs by unrolling the DBN for T slices and apply static BN algorithms.

Similarly to static BNs, learning starts with some a *priori* knowledge about the model structure and parameters. The prior probability distribution over the model parameters and structure is used to represent this initial knowledge, which is updated using data to obtain a posterior probability distribution over model parameters and structure (Ghahramani & Jordan (1997)). Specifically, assuming a prior distribution over model structure and parameters, a data set is used to obtain the posterior distribution using Bayes rule (we provided an example of how to use Bayes rule in 3.1.1 *Bayesian Inference*).

Since only a subset of states can be observed at each time slice, we can do inference by calculating all the unknown states in the network. This inference problem can be represented as finding $\mathrm{P}(X_0^{t-1}|\,U_0^{t-1})$, where U_0^{t-1} is a finite set of continuous observations, $U_0^{t-1} = \{u_0,u_1,...,u_{t-1}\}$ and X_0^{t-1} is the corresponding set of hidden variables X_0^{t-1} $=\{x_0,x_1,...,x_{t-1}\}$. (Mihajlovic & Petkovic (2001)).

According to the type of DBN and its structure, different calculations might be needed. That is why sometimes it could be more appropriate to estimate the summary statistics for $\mathrm{P}(X_0^{t-1}|\,U_0^{t-1})$ by choosing some number of states or a single state and calculate only their values throughout the different time slices (Mihajlovic & Petkovic (2001)). In this case, the $\mathrm{P}(X_0^{t-1}|\,U_0^{t-1})$ is expressed as a *Gaussian function* and the *mean* and *variance* for x_t, denoted as $\langle x_t|\,U_0^{t-1}\rangle$ and $\langle x_t x_t'|\,U_0^{t-1}\rangle$ respectively for every t are estimated (Mihajlovic & Petkovic (2001)).

Another interesting inference problem is predicting future observation and/or estimating hidden states based on past observations. Having observed $\{\mathrm{u}_1...\mathrm{u}_t\}$, to predict the next observation U_{t+1}, based on the data D and model M, the prediction will be:

$$P(U_{t+1}|D) = \int P(U_{t+1}|\theta, M, D)P(\theta|M, D)P(M|D)\; d\theta\; dM \tag{3.23}$$

which integrates out the uncertainty associated with the structure and parameters θ (Ghahramani & Jordan (1997)).

The prediction of a hidden state in the next time slice can be defined as an inference calculation of $\mathrm{P}(x_{t+1}|\,U_0^t)$ or $\mathrm{P}(u_{t+1}|\,U_0^t)$, according to Mihajlovic & Petkovic (2001) and they have shown:

$$P(x_{t+1}|\,U_0^t) = \frac{\sum_{x_t} P(x_{t+1}|x_t)\alpha_t(x_t)}{\sum_{x_t} \alpha_t(x_t)} \tag{3.24}$$

and similarly:

$$P(u_{t+1}|\,U_0^t) = \frac{\sum_{x_{t+1}} \alpha_{t+1}(x_{t+1})}{\sum_{x_t} \alpha_t(x_t)} \tag{3.25}$$

Often, it is more appropriate to express the prediction problem in the sense of expected or maximum likelihood estimates. Assuming a data set D with observations $\{U^{(1)}...U^{(N)}\}$, the likelihood of the data set, according to Ghahramani & Jordan (1997) is defined as:

$$P(D|\theta, M) = \prod_{i=1}^{N} P(U^{(i)}|\theta, M) \tag{3.26}$$

The maximum likelihood parameters are obtained by maximising the log likelihood:

$$L(\theta) = \sum_{i=1}^{N} \log P(U^{(i)}|\theta) \tag{3.27}$$

If the observation vector includes all the variables in the BN, we further factor each term in the log likelihood:

$$\log P(U^i|\theta) = \log \prod_j P(U_j^{(i)}|U_{pa(j)}^{(i)}, \theta_j) \tag{3.28}$$

$$= \sum_j \log P(U_j^{(i)}|U_{pa(j)}^{(i)}, \theta_j) \tag{3.29}$$

where j indexes over the DBN nodes, pa_j are the parents of j and θ_j are the parameters that define the conditional probability of U_j given its parents (Ghahramani & Jordan (1997)). In this way, the likelihood splits into local terms of each node and its parents. When there are hidden variables present, the log likelihood cannot be decoupled as in (3.24) but it is:

$$L(\theta) = \log P(U|\theta) = \log \sum_X P(U, X|\theta) \tag{3.30}$$

where $\sum_X$ is the sum over the set of hidden variables X, required to obtain the marginal probability of the data (Ghahramani & Jordan (1997)). Here, the EM algorithm (more detail in 3.1.2 *Parameter Estimation*) is applied which alternates between maximising the log likelihood with respect to some distribution over the hidden variables and θ respectively.

Although, inference can measure the performance of the network, to measure the goodness of inference, the *prediction accuracy* is usually calculated which is the proportion of variables' (fish species in this thesis) values that have been predicted correctly. In this work, the prediction accuracy is calculated after applying non-parametric bootstrap approach (re-sampling with replacement from the training set, Fernandes et al. (2013)) to obtain statistical validation in the predictions and calculating the sum of squared error (SSE) relating to model performance. This allowed for a quick analysis and comparison of the prediction accuracies of the modelling results with the real species data. The bootstrapping technique allows to obtain an unknown characteristic of an unspecified distribution by drawing subsets from the observed data iteratively, and computing a statistic for each subset. For a great number of iterations, the bootstrap distribution approximates the actual distribution.

3.4 Hidden Markov Model

Hidden Markov models (HMMs) have been extensively used as a data-driven modelling approach in speech recognition (Juang & Rabiner (1991); Rabiner (1989)), where signals are encoded as temporal variation of a short time power spectrum (Al-ani (2011)). Other HMM applications include computational biology (Krogh et al. (1994)), fault detection (Smyth (1994)), signal processing (Vaseghi (2008)) and telecommunication (Hirsch (2001)). The fact that these models represent in a natural way the individual component states of a dynamic system makes them so advantageous in different fields. Most importantly, HMMs can be used to describe discrete stochastic changes in a system with ongoing continuous dynamics. That is why, specifically for this thesis, this kind of model was chosen to model data that vary over time and can be thought of as having been generated by a process that iterates between different states at different time points. Of particular interest is the missing set of indicator variables that describes which state gave rise to a particular observation. Such missing or hidden indicator variables are not independent but are governed by a stationary Markov chain (McGrory & Titterington (2009)).

In an HMM, the sequence of observations $\{\mathrm{U}_t\}$ is modelled by assuming that each observation depends on a *discrete* hidden state S_t and that the sequences of hidden states are distributed according to a Markov process (Ghahramani & Jordan (1997)). The joint probability for the sequences of states and observations can be factorised, according to Ghahramani & Jordan (1997):

$$P(\{S_t, U_t\}) = P(S_1)P(U_1|S_1)\prod_{t=2}^{T} P(S_t|S_{t-1})P(U_t|S_t) \tag{3.31}$$

The hidden state is represented by a single multinomial variable that can have one of K discrete values, $\mathrm{S}_t \in \{1,..., \mathrm{K}\}$ (Ghahramani & Jordan (1997)). The state transition probabilities, $P(\mathrm{S}_t|\mathrm{S}_{t-1})$ are specified by a single $K\mathrm{x}K$ transition matrix (Ghahramani & Jordan (1997)). For real-valued observation input, $P(\mathrm{U}_t|\mathrm{S}_t)$ is modelled in a Gaussian form:

$$P(U_t = u|S_t = i) = N(u; \mu_i, \Sigma_i) \tag{3.32}$$

where $N(u;\mu, \Sigma)$ is the Gaussian density with *mean* μ and *covariance* Σ or modelled as a mixture of Gaussians:

$$P(U_t = u|S_t = i) = \sum_{m=1}^{M} P(M_t = m|S_t = i)N(u; \mu_{m,i}, \Sigma_{m,i}) \tag{3.33}$$

where M is a hidden variable that specifies which mixture component to use and $P(M_t{=}m|S_t{=}i){=}\ C(i,m)$ is the conditional prior weight of each mixture component (Murphy (2002)).

Rabiner (1989) speaks of an HMM "generating" a sequence. The HMM is composed of number of states which are interconnected by state-transition probabilities. Starting from the initial state, a sequence of states is generated by moving from state to state according to the state transition probabilities until an end state is reached (Eddy (1996)). The sequence of states is a *Markov chain* because the choice of the next state is dependant on the current state (Eddy (1996)). However, in the case of HMMs the state sequence is hidden. Only the sequence that the hidden states generate is observed. Let us consider an example: assuming we have a video sequence and we would like to decide whether a speaker is in a frame (example taken from Jia Li, *Hidden Markov Model*. Department of Statistics, The Pennsylvania State University, www.sites.stat.psu.edu). The underlying states are: *with a speaker* (state 1) vs. *without a speaker* (state 2). Let S_t denote whether there is a speaker in the frame for $t{=}\ 1...,T$. If one frame has a speaker, it is very likely that the next frame would also contain a speaker because of the frame-to-frame dependence. Conversely, a frame without a speaker is much more likely to be followed by another frame without a speaker. Thus, the S_t's are not independent but we assume that the state sequence follows a *Markov chain*. Here, the goal is to find out the state sequence, given the observed sequence of feature vectors that can be extracted for each frame. It is expected that the probability distribution of a feature vector differs according to the state. By using the dependence among states, we can make better predictions of the states than guessing each state separately using only the feature vector of that frame.

An HMM can be represented as an instance of a DBN, for example in Fig.3.5 which defines the conditional independence assumptions: $\mathrm{S}_{t+1} \perp \mathrm{S}_{t-1}|\mathrm{S}_t$ (the Markov property) and $\mathrm{U}_t \perp \mathrm{U}_{t'}\ |\mathrm{S}_t$, for $t' \neq t$.

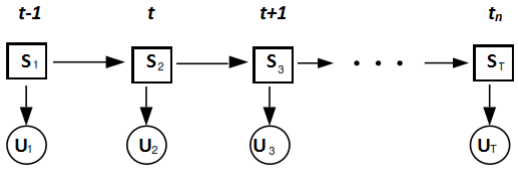


Figure 3.5: A hidden Markov model, where the square nodes represent discrete hidden nodes and circle nodes represent observed nodes.

The HMM consists of one hidden discrete node and one continuous or discrete observed node per slice. We assume the distributions $\mathrm{P}(S_{(t)}|S_{(t-1)})$ and $\mathrm{P}(U_{(t)}|S_{(t)})$ are independent of t for $t > 1$. We also assume the observation model $\mathrm{P}(U_{(t)}|S_{(t)})$ is independent of t for $t > 1$, but it is conventional to assume this is true for all t. As with static BNs, the CPD of each node is defined given its parents, that is defining $P(S_1)$, $P(\mathrm{S}_t|S_{t-1})$ and $P(\mathrm{U}_t|S_t)$ which was explained above.

The difference between a DBN and an HMM is that the DBN represents the hidden variable in terms of a set of random variables, X_t^1,...,$X_t^{N_h}$ whilst in the HMM, the state space consists of a single random variable S_t and the HMM has a fixed structure whilst the DBN allows for more general graph structures (Murphy (2002)). In an HMM, we have to define the transition model $P(\mathrm{S}_t|\mathrm{S}_{t-1})$, the observation model $P(\mathrm{U}_t|\mathrm{S}_t)$ and the initial state distribution $P(S_1)$, whilst in a DBN U_t also represents sets of variables. Similarly to DBNs, the HMM models the conditional distribution of a sequence of outputs, given a sequence of input observations. To be able to do that, the HMM can be augmented to allow for input variables U_t in a way that there is an input dependant state transition probability, $P(S_t|S_{t-1}, U_t)$ (Bengio & Frasconi (1995)). Like DBNs, we can assume that the parameters are time-invariant, thus we only need to specify $P(S_1)$, $P(S_2|S_1)$ and $P(U_1|S_1)$. Consequent CPDs are assumed to be the same as in the first two slices.

3.4.1 Learning Hidden Markov Models

The log probability of the hidden variables and observations for an HMM following Ghahramani & Jordan (1997) is defined as:

$$\log P(\{S_t, Y_t\}) = \log P(S_1) + \sum_{t=1}^{T} \log P(Y_t|S_t) + \sum_{t=2}^{T} \log P(S_t|S_{t-1}) \tag{3.34}$$

Each of the terms from (3.34) can be decoupled into summations over S. According to Ghahramani & Jordan (1997), the transition probability is:

$$P(S_t|S_{t-1}) = \prod_{i=1}^{K}\prod_{j=1}^{K}(P_{ij})^{S_{t,i}S_{t-1,j}} \tag{3.35}$$

with P_{ij} being the probability of transitioning from state j to state i, arranged in a KxK matrix P. Then

$$\begin{aligned}\log P(S_t|S_{t-1}) &= \sum_{i=1}^{K}\sum_{j=1}^{K} S_{t,i}S_{t-1,j}\log P_{ij} \\ &= S_t'(\log P)S_{t-1}\end{aligned} \tag{3.36}$$

using matrix notation (Ghahramani & Jordan (1997)). If we assume a vector of initial state probabilities, π, then

$$\log P(S_1) = S_1' \log \pi \tag{3.37}$$

The emission probabilities depend on the form of the observations. If Y_t is a discrete variable which can have D values, we represent it using D-dimensional unit vectors to receive:

$$\log P(Y_t|S_t) = Y_t'(\log E)S_t \tag{3.38}$$

where E is a DxK emission probability matrix (Ghahramani & Jordan (1997)).

We cannot compute (3.34) directly because of the hidden state variables. However, the EM algorithm allows solving this problem by computing the expectation of (3.34) under the posterior distribution of the hidden states given the observation (Ghahramani & Jordan (1997)). The expectation can be expressed as a function of $\langle \mathrm{S}_t \rangle$ and $\langle \mathrm{S}_t \mathrm{S}'_{t-1} \rangle$ ($1 \leq t \leq T$), where $\langle \mathrm{S}_t \rangle$ is a vector containing the probability that the model was in each of the K states at time t given its current parameters and sequence of observations (Dean & Kanazawa (1989)). The second term, $\langle \mathrm{S}_t \mathrm{S}'_{t-1} \rangle$, is a matrix containing the

joint probability that the model was in each of the K^2 pairs of states at times *t-1* and t (Ghahramani & Jordan (1997)). Then, the **M** step of the EM is straightforward: we take derivatives of (3.34) with respect to the parameters, set to 0, and solve subject to the sum-to-one constraints that provide valid transition, emission and initial state probabilities (Ghahramani & Jordan (1997)).

3.4.2 Hidden Variables

Variables that are not observed are the so-called *hidden* variables and are used to model missing data and/or unmeasured effects. Current methods for learning hidden variables require experts to choose a fixed network structure or a small set of possible structures but sometimes this might not be feasible within some domains (Friedman et al. (1997)). For example in a medical domain, in which we have the observed symptoms (e.g., fever, headache, blood pressure etc.), and the medication prescribed by the doctor but the disease is a hidden quantity. If we were to know the disease, the treatment becomes independent of most of the symptoms. When we do not know the disease, all symptoms are related to each other.

By introducing hidden variables into the models, simpler models can be learned that are less prone to overfitting and more efficient for inference. For example, the apparent complexity of a predicted variable can be explained imagining it is a result of two simple processes, the "true" underlying state, which may evolve deterministically, and our measurement of the state, which is often noisy (Murphy (2002)). We can then "explain away" unexpected outliers in the observations, as opposed to strange fluctuations in "reality".

In most domains, the observed variables represent only some characteristics of a system, which can have a negative effect on the learning procedure. It has been shown that the hidden variable is only beneficial when it is connected to other observed variables in the network (Friedman et al. (1997)) as the hidden variables are parameterised through the observed variables by using the EM algorithm. For example, look at the structure of the Autoregressive Hidden Markov model (ARHMM), where the hidden variable is a parent to each observed variable (Fig.3.6). Russell et al. (1995) and Binder et al.

(1997) applied the EM to learn parameters with possible hidden variables. They also extended this to the case of continuous nodes and dynamic Bayesian networks. Kwoh and Gillies (1996) applied the same idea, but also described the technique of inventing hidden nodes to describe dependencies between variables. If all the hidden nodes are discrete, we can use the *junction tree algorithm* to perform inference which unrolls the DBN into a static network and applies the jtree algorithm.

In this thesis, the hidden variable is used to model unobserved variables and missing data which can infer some underlying state of the series when applied through an autoregressive link (ARHMM, Fig.3.6) that can capture relationships of a higher order (Murphy et al. (2001)). In this work, the hidden variable was chosen to most easily reflect the complex interdependencies between and among species groups and their environment. Specifically, we used the hidden variable in our models to capture changes in the species variance and/or to reflect on temporal changes of the underlying processes in different ecosystems. In addition, the hidden variable allows examining unmeasured effects that are often not constrained within the model structure or measured data. For example, we used the hidden variable to model a group of species for which we did not have measured data but we predicted the state of the hidden variable with the help of experts who defined part of the network structure.

3.5 Autoregressive Hidden Markov Model

Sometimes time series may consist of observations generated by different mechanisms at different time points (Xuan (2004)). In this case, the time series observations would alternate back and forth between different states. When changing into a different state, the time series could be characterised with a significant change in their means or frequencies. The *Autoregressive Hidden Markov model* (ARHMM) can handle such fluctuations of the time series. It is a combination of an autoregressive time series model and a hidden Markov model. It is also referred as *time series with change in regime* (or *states*) by the econometricians (Xuan (2004)). The ARHMM often leads to models with a higher likelihood of convergence than general HMMs, especially when modelling complex and noisy time-series data. The standard HMM assumption $Y_t \perp \mathrm{Y}_{t'} \mid X_t$ can be relaxed as

shown in Fig.3.6, which shows an ARHMM. This type model allows for Y_{t-1} to predict Y_t (not only from X_t as with the general HMM) as well which often results in models with higher likelihood (Murphy (2002)). If Y is discrete, the CPD can be represented as a table and if Y is continuous, one way to show the CPD is:

$$P(Y_t = y_t | X_t = i, Y_{t-1} = y_{t-1}) = N(y_t; W_i y_{t-1} + \mu_i, \Sigma_i) \tag{3.39}$$

where W_i is the regression matrix given that X_t is in state i (Murphy (2002)).

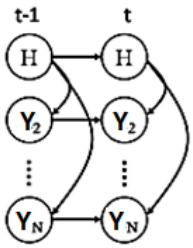


Figure 3.6: An autoregressive hidden Markov model, where H denotes an unmeasured hidden variable.

The ARHMM is specifically applied to model correlations in sequential data (Xuan (2004)). The hidden variable (H in the ARHMM) is used to model unobserved variables and missing data and can infer some underlying state of the series when applied through an autoregressive link (Fig.3.6) that can capture relationships of a higher order (Murphy (2002)). In this way, the ARHMM enhances the HMM by introducing a direct stochastic dependence between observations (Murphy (2002)). Specifically, the observation vector Y with components $\{y_0, y_1, y_2, ..., y_{t-1}\}$ is drawn from an autoregressive process (Rabiner (1989)). So the current observation Y_t not only depends on the last observations but also on the current states. Each of the observations is modelled as a *Markov chain* and observations are assumed to be conditionally independent given the state (Stanculescu et al. (2014)):

$$P(Y_t | Y_{t-1}, X_t) = \prod_{n=1}^{N} P(Y_t^N | Y_{t-1}^{N, X_t}) \tag{3.40}$$

The probability density function for Y is Gaussian autoregressive (or order p), the components of Y are related by:

$$Y_k = -\sum_{i=1}^{p} a_i Y_{k-i} + e_k \tag{3.41}$$

where e_k,k=0,1,2,...,*K-1* are Gaussian, independent identically distributed random variables with zero *mean* and *covariance* σ^2 and a_i,i=1,2,...,p are the autoregression or coefficient predictors (Rabiner (1989)).

$X = \{x_1,x_2,...,x_t\}$is the hidden state sequence with N possible states. X is assumed to be a *Markov chain* with transition matrix $A = [a_{ij}]$ and initial distribution vector $\pi = [\pi_i]$. The hidden states of an ARHMM can take one of J values and are also organised as a *first-order Markov chain* with parameters $\theta_{j|i} = P(X_t = j|X_{t-1} = i)$ and $\pi_j = P(X_1 = j)$ (Stanculescu et al. (2014)). Conditioned on the state X_t, the emission process is also a *first-order Markov chain*, parametrised by $\phi_{l|kj} = P\ (Y_t = l|\, Y_{t-1} = k,X_t = j)$ and $\pi_{l|j} = P(\, Y_1 = l|X_1 = j)$ (Stanculescu et al. (2014)). According to Stanculescu et al. (2014), the joint probability distribution is given for a sequence of length T:

$$P(X_{1:T}, Y_{1:T}) = \pi_{x1}\pi_{y1|x1}\prod_{t=2}^{T}\theta_{xt|xt-1}\phi_{y_t|y_{t-1}xt'} \tag{3.42}$$

Similarly to the HMM, the ARHMM models hidden discrete states X_t and observations Y_t. To recall, in the HMM, the current observation is independent of all the other observations given the current state. Here, the ARHMM allows for correlation amongst observations by adding direct dependencies between them which makes the ARHMM a genuinely better model for handling temporal data (Stanculescu et al. (2014)).

3.5.1 Inference and Learning

Similarly, the main inference problem for ARHMMs is computing $P(X_t|\, Y_{1:t})$. We use the EM algorithm for parameter estimation with hidden variables. We use $\theta = \{\pi, A, \phi\}$ to define the set of parameters. The likelihood function is obtained from the joint distribution by marginalising over the hidden variables (Bishop (2006)):

$$P(Y|\theta) = \sum_{X} P(Y, X|\theta) \tag{3.43}$$

We start the EM algorithm with some initial value set in the parameters, which we denote by θ^{old}. In the E-step of EM, the objective is to find the expectation of the likelihood. Thus, we use the posterior distribution of the hidden variables to evaluate the expectation of the complete-data likelihood function, as a function of the parameters θ, to give the function Q(θ, θ^{old}), according to Bishop (2006). The remaining part of the E-step is to compute marginal posterior distribution of the hidden variable and the joint posterior distribution of two successive hidden variables. To do this, the forward-backward algorithm can be applied. The idea behind this algorithm is to find out a recursive way to represent the variable sequence. The next task is to maximize the expectation likelihood, which is the so called M-step of EM algorithm or specifically to maximise the logarithm expectation likelihood Q(θ,θ^{old}).

Solving the problem of HMM parameter learning is the generation of all possible sequences of observed events and hidden states with their probabilities. The joint probability of two sequences, given the model, is calculated by multiplying the corresponding probabilities in the matrices. Such procedure has a complexity of O($2\mathbf{TN}^T$) where $\mathbf{T}$ is the length of the sequences of observed events and $\mathbf{N}$ is the is the total number of symbols in the state alphabet (Bishop (2006)).

3.6 Conclusion

Several algorithms and modelling approaches have been described to detect the best possible model to build functional network models that describe species associations and interactions within their environment. Each method with its strengths and weaknesses can be applied to discover the underlying mechanism and the relationship hidden in the data under analysis.

In the next chapter, we show how we can use the hidden Markov model across geographic regions to make predictions of complex real time ecological data. Specifically, we show how we can use a hidden variable to detect unmeasured effects about the underlying state of different ecosystems.

Chapter 4

Exploring Regime Metrics into Bayesian Network Models with Hidden Variables to Detect Early-Warning Signals of Functional Changes in Fisheries Ecology

4.1 Introduction

In this chapter, dynamic Bayesian networks have been applied to predict future biomass of geographically different but functionally equivalent fish species. A hidden variable is incorporated to model functional collapse, where the underlying food web structure dramatically changes irrevocably (known as a *regime shift*). Regime (functional) changes can affect the abundance and distribution of fish populations, either directly or by affecting prey or predator populations (Jiao (2009)). The main question for environmental management is whether such changes could have been detected by early-warning signals.

There is a growing literature that addresses indicators that can be used as early-warning signals of an approaching critical transition (or a regime shift) (**?**). We examined if the use of a hidden variable can reflect changes in the trophic dynamics of the system and also whether the inclusion of recognised statistical metrics would improve predictive accuracy of the dynamic models.

We explored functional relationships (such as predator, prey) that are generalizable between different oceanic regions allowing predictions to be made about future biomass. In particular, we exploited multiple fisheries datasets in order to identify species with similar functional roles in different fish communities. The species were then used to predict functional collapse in their respective regions through the use of Dynamic Bayesian Networks (DBNs) with hidden variables. This represents the most challenging inference problem here as we make computationally complex predictions involving dynamic processes. However, the hidden variable is chosen to most easily reflect complex interdependencies between the acting variables. Thus, we are specifically interested in using the architecture of an Autoregressive hidden Markov model (ARHMM) that incorporates a hidden variable (Fig.4.1).

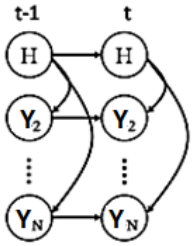


Figure 4.1: An autoregressive hidden Markov model, where H denotes an unmeasured hidden variable.

In this chapter, we investigate the reliability of our modelling approach in detecting *early-warning* signals of functional change across *different geographic regions*. We explore how the hidden variable reflects the *regime metrics* (the applied statistical indicators of functional changes in the study) and the variability of exploited fisheries and to what extent including them in our models impacts the expected values of the hidden

variable. We also explore how these models can be used to identify species that are key to regime shifts in different regions. An earlier work by Tucker & Duplisea (2012) explores functionally equivalent species but here we further adopt the approach to predict functional collapse by investigating *early-warning signals* and comparing learned Bayesian network (BN) topology prior to and after suspected regime changes. At larger spatial scales, although fishing can still be the dominant driver of regime changes, the consequences of fishing are not predictable without understanding the trophic dynamics (Jiao (2009)). A clear example is the Scotian Shelf, where fishing has led to a restructuring of the ecosystem (Jiao (2009)). We investigate whether exploiting expert knowledge (in the form of diet matrix, that represents the prey-predator functional relationships between species) of this region or learning model structure from the data alone is a better approach when predicting food web dynamics.

4.2 Description of the Geographic Regions and Biomass Data to Model Regime Shifts

We apply our modelling approach to predict species biomass and functional change in three different geographical regions: North Sea (NS), Georges Bank (GB) and East Scotian Shelf (ESS) (Fig.4.2). For all of the datasets, the biomass was determined from research vessel fish trawling surveys assuring consistent sampling from year to year, resulting in 44 species for NS (1967-2009), 44 species for GB (1963-2008), and 42 for ESS (1970-2006). In addition to survey data on fish biomass, grey seal abundance and plankton time series were also included in the analysis.

GB is a relatively small productive fishing bank historically supporting large catches of common groundfish such as cod and haddock. The ESS, though geographically not far from GB is a much different system with lower productivity and diversity. A key characteristic of the ESS is the presence of a small sandy arc 200 km offshore called Sable Island, which is the largest grey seal breeding colony in the world. The NS is a shallow warm sea with high fish community diversity and productive multi-species fisheries.

Large groundfish declines occurred on GB and ESS which resulted in the year 1988 being

designated as a collapse year for GB and 1992 for ESS. Despite the extremely high fishing pressure in NS and complex climate-ocean interactions, it is difficult to distinguish a radical switch in the system that might be termed a regime shift. However, experts refer to some ecosystem changes in the period of late 1980s to mid-1990s.

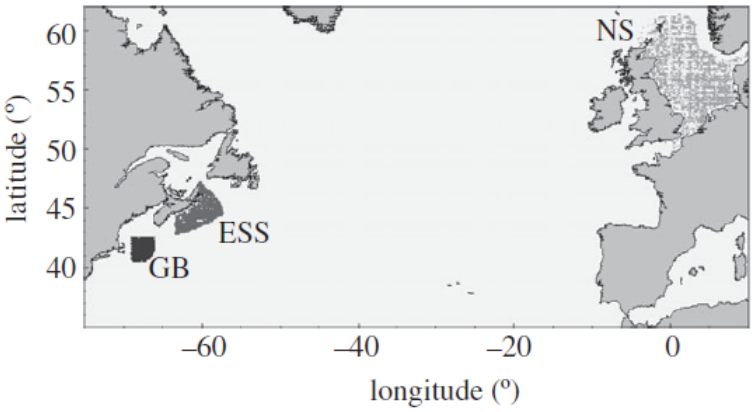


Figure 4.2: Regions of the three surveys (shaded area) corresponding to the three datasets: Georges Bank (GB), East Scotian Shelf (ESS) and the North Sea (NS).

4.3 Methods

4.3.1 Introducing an Algorithm for Learning Functional Equivalence

This part of the experiments involves applying classification. Feature selection is used to identify the relevant species for optimal classification.

We select a number of species that are associated with cod collapse by using wrapper feature selection with a Bayesian Network Classifier on GB data where the class node is a binary variable that represents functional collapse in GB. The greedy K2 search algorithm (Cooper & Herskovits (1992)) is used to build the BN classifiers. This involves a greedy search technique where links are incrementally added to an initially unconnected graph and scored using the metric given in equation (4.1), where n is the number of nodes, F_{ijk} is the frequency of occurrences in the dataset that the node x_i takes on the value vik (where there are r_i possible instantiations) and the parent nodes π_i take on

the instantiation w_{ij} (where there are q_i possible instantiations).

$$\log \sum_{i=1}^{n} \sum_{j=1}^{q_i} \frac{(r_i - 1)!}{(F_{ij} + r_i - 1)!} \sum_{k=1}^{r_i} F_{ijk} \tag{4.1}$$

A bootstrap approach (Efron & Tibshirani (1995)) is employed to repeat the following 1000 times:

- score each species using the likelihood score from (4.1) and take the mean over the bootstrap, this is known as filter feature selection (Inza et al. (2004)) which scores each variable independently;
- learn BN structure with the K2 algorithm and score the proportion of times that links are associated with the class node during the bootstrap. This is a form of wrapper feature selection (Inza et al. (2004)) which scores each variable by taking into account their interaction with other variables through the use of a classifier model.

Species are ranked based upon the two feature selection approaches to examine their relevance to functional collapse in GB. Next, we identify the equivalent species in the other two datasets (NS and ESS) using species identified by feature selection from GB in combination with the features discovered using Algorithm 1. The algorithm is applied to both the NS and ESS to identify equivalent species. The *functional equivalence* search algorithm (Tucker & Duplisea (2012)) works by using a BN model, where the given function is in the form of a predefined structure, BN_1, and a set of variables, $vars_1$, parameterised on $data_1$, (here a BN model parameterised on the GB data). Simulated annealing (Kirkpatrick et al. (1983)) is applied to identify variables in another dataset, $data_2$, (here species in the ESS and NS datasets) that best fit this model. We set $iterations = 1000$ and $t_{start} = 1000$ as these were found through experimentation to allow convergence to a good solution. The fit is scored using the log likelihood score (Cooper & Herskovits (1992)). In Algorithm 1 *UnifRand* represents a random value generated from a uniform distribution with limits between (0,1).

Algorithm 1 The *functional equivalence* search algorithm.

Input: $t_{start}, iterations, data_1, data_2, vars_1, BN_1$
Parameterize Bayesian Network, BN_1, from $data_1$
Generate randomly selected variables in $data_2 : vars_2$
Use $vars_2$ to score the fit with selected model $BN_1 : score$
Set $bestscore = score$
Set initial temperature: $t = t_{start}$
for $i = 1$ to *iterations* **do**
    Randomly replace one selected variable in $data_2$ and rescore: *rescore*
    dscore = rescore - bestscore
    if *dscore* ≥ 0 OR *UnifRand and* (0,1) $< \exp^{(dscore/t)}$ **then**
        bestscore = rescore
    else
        Undo variable switch in $vars_2$
    end if
    Update the temperature: $t = t$ x 0.9
end for
Output: $vars_2$

4.3.2 Hidden Variable Models to Predict Regime Shifts and Species Biomass

The experiments involve the prediction of a pre-selected variable (here functional collapse, represented by the hidden variable) based on the values of other variables (here species biomass). After choosing the species, we want to predict their biomass and the functional collapse in the relevant geographic region. Given a graphical structure (Fig.4.1), and a probability distribution over $\mathbf{Y}[t]$ where $\mathbf{Y} = Y_1 ... Y_n$ are the n variables observed along time t, DBNs naturally perform prediction using inference. To predict functional collapse, we compute $P(X^t|Y^t, Y^{t-1})$, where X^t represents the hidden variable (functional collapse) and Y^t represents all observed variables at times t. First, we infer the biomass at time t by using the observed evidence (or species biomass) from *t-1* and then use the predicted variable states (or species biomass) to infer the hidden

state at time t. The hidden variable was parameterised using the EM algorithm. We used an exact inference method: the junction tree algorithm.

The statistical metrics we chose to apply as "regime metrics", that would help us in identifying a functional collapse by modelling early-warning signals in the time-series, were variance and autocorrelation (**?**). According to theoretical expectations of critical slowing down (or an approaching functional collapse), the metrics would appear to increase prior to the expected regime shift and follow a consistent decline throughout time following the collapse potentially resulting in a clear early-warning signal to forewarn a major ecosystem change.

The metrics were calculated on a window of data, set to size 10, so that each metric captures the value of interest over the previous 10 years. Two sets of experiments were then conducted: one that excludes the regime metrics to examine the expected state of the fitted hidden variable (ARHMM) and check predictive accuracy of the model. In the other, metrics were included in the model (ARHMM + metrics) to see if they improve prediction of species biomass and whether the hidden variable would capture the predictive qualities of the metrics in terms of detecting a regime shift. Non-parametric bootstrap analysis (Friedman et al. (1999)) was applied 250 times for each variant of the model to obtain statistical validation in the predictions. We found the window size and number of iterations through experimentation to be optimum when dealing with limited time series.

An F-test was performed over a sliding window of five years to detect any significant changes in the slope of the hidden variable from both models before and after the time of the expected collapse (Gröger et al. (2011)). Given a breakpoint in the time series, the minimum of this sequence of p-values gives a potential estimate of the first signals of ecosystem change in time. Levene's test on homoscedasticity was performed on the variance before and after the predicted functional change (Gröger et al. (2011)). Note, if the p value is small ($p \leq .05$), we accept the null hypothesis and conclude that the two distributions (before and after the regime shift) are distinct, however if the p value is large ($p \geq .05$), the data do not give us any reason to accept the null hypothesis, which does not necessarily mean that the two distributions are identical but there is no compelling evidence that they differ significantly. All statistical tests were reported at 5%

significance level. To highlight, there are many approaches such as Benjamini Hochberg to adjust the false discovery rate, although, here this was not performed. However, the techniques of sliding window and bootstrap of sampling with re-placement were preferred in order to maximise the confidence in the learned models and obtain statistical validation in the predictions. In addition, the expected value of the hidden variable was explored with and without the influence of recognised statistical metrics that are known to predict functional change. Comparison of the expert knowledge versus data alone was also performed, as explained below. The results are explored and discussed with caution in terms of the wider fisheries literature.

4.3.3 Learning Data-Driven Networks versus Pre-defined Diet Matrices

Finally, we explore whether exploiting expert knowledge in the form of diet matrices based upon stomach surveys is a better approach to learning model structure than using biomass data alone when predicting food web dynamics. Specifically, we wanted to learn data-driven features of networks to see whether we would be able identify "true" relations between species. Then, we would compare the learned features versus a diet matrix. The diet matrix reports a predator and the biomass of its main fish prey. The matrix is build for a number of predators based on a survey of stomach contents data for each one of the regions.

In this part of the experiments, learning the dynamic BN structure, the species biomass data was discretised and a greedy search algorithm: REVEAL (Liang et al. (1998)) was applied to learn the structure of the DBN model for each region. The purpose of the discretization process was to transform a range of numeric attributes in the dataset into nominal attributes, assigning each interval to the class distribution. Automatic discretization to bin the data, using splits based on quantiles of the data itself, was applied. The discretization was needed here as there is currently only one structure learning algorithm (REVEAL) for DBNs that the BN toolbox in Matlab can handle (Murphy et al. (2001)). This assumes all nodes are discrete and observed, and that there are no intra-slice connections. Hence, we can find the optimal set of parents for

each node separately, without worrying about directed cycles or node orderings. The algorithm sets a penalty term to 0, which gives the maximum likelihood model. This is equivalent to maximizing the mutual information between parents and child. A non-zero penalty invokes the BIC criterion, which lessens the chance of over-fitting (Murphy et al. (2001)).

Similarly, to prediction, a non-parametric bootstrap was also applied 250 times to identify statistical confidence in the discovered network links with *threshold* ≥ 0.5. Features with statistical confidence above the threshold are labelled "positive" or "negative" if the confidence is below the threshold. We measure the number of "true positives", correct features of the generating network (based upon a pre-defined diet matrix, established by stomach content surveys for the relevant region) or "false negatives", correct features labelled as negatives (Friedman et al. (1999)).

4.4 Results

4.4.1 Identified Functionally Equivalent Species from Different Geographic Regions

The wrapper feature selection approach managed to identify the species likely to be associated with cod collapse on GB, Table 4.1 illustrates the resulting ordered list of most relevant variables (BN wrapper confidence reported in brackets). For example, herring (*Clupea harengus*) was identified as a key species and it is known that there were large abundance changes in the late 1980s (Beaugrand et al. (2003)). Similarly, the approach identified two zooplankton species (*Calanus* and *Pseudocalanus*) as key to the functional collapse and it is known from other sources that there were relatively large changes then (Kane (2007)), and these changes can have bottom-up effects which affect species such as cod (Beaugrand et al. (2003)). The confidences resulting from the Bayesian wrapper method applied to GB showed that thorny skate was the most important species implicated in the functional decline due to an ecosystem regime shift which coincided with an increase in abundance (Fogarty & Murawski (1998)). When this structure is imposed on the ESS and NS using the functional equivalence search, a small number of equivalent species are recognised in both regions with high confidence.

Table 4.1: Wrapper feature selection results for GB region.

GB Wrapper Feature Selection	
1. Thorny skate (1.0)	14. Lady crab (0.24)
2. Blackbelly rosefish (0.98)	15. Spotted flounder (0.23)
3. Herring (0.97)	16. *Calanus* spp. (0.20)
4. Fourbeard rockling (0.82)	17. American lobster (0.13)
5. Cusk (0.75)	18. American plaice (0.13)
6. *Pseudocalanus* spp. (0.65)	19. Ocean pout (0.09)
7. Gulf stream flounder (0.47)	20. Little skate (0.07)
8. *Centropages typicus* (0.44)	21. Sea scallop (0.07)
9. Atlantic rock crab (0.41)	22. Sand lance (0.05)
10. Witch flounder (0.29)	23. Winter flounder (0.03)
11. American angler (0.28)	24. Moustache sculpin (0.02)
12. White hake (0.26)	25. Silver hake (0.02)
13. Krill (0.25)	26. Longfin hake (0.02)

The species from ESS and NS that were identified by the *functional equivalence* search algorithm are ranked based upon the confidence associated with their equivalent species in GB (Table 4.2, confidence reported in brackets).

Table 4.2: The functionally equivalent species to GB dataset for ESS and NS. These are each ordered based upon their relevance to species in Table 1.

Functionally Equivalent Species	
ESS	**NS**
1. Cod (1.0)	1. European plaice (0.98)
2. Pollock (0.58)	2. Atlantic halibut (0.93)
3. Grenadier (0.51)	3. Cod (0.87)
4. White hake (0.50)	4. Lumpfish (0.78)
5. Mackerel (0.23)	5. Thorny skate (0.53)
6. Rockfish (0.22)	6. Whiting (0.50)
7. Grey seals (0.20)	7. Argentine (0.42)
8. Argentine (0.15)	8. Megrim (0.35)
9. Atlantic halibut (0.12)	9. Haddock (0.29)
10. Spiny dogfish (0.11)	10. Atlantic wolfish (0.24)
11. Little skate (0.09)	11. American plaice (0.21)
12. Atlantic wolfish (0.07)	12. Common dragonet (0.20)
13. American plaice (0.04)	13. Solenette (0.16)
14. Red hake (0.04)	14. Poor cod (0.13)
15. Silver hake (0.03)	15. Sprat (0.05)
16. Little hake (0.01)	16. Pollock (0.02)

A striking feature of the identified ESS species is the presence of many deepwater species like argentine (*Argenti silus*) and grenadier (*Nezumia bairdi*). That could be an indication of the water cooling that occurred in the late 1980s and early 1990s. Grey seals were also included as they were implicated in the functional decline and lack of recovery of many groundsh stocks on the ESS. In the NS, most of the selected species are commercially desirable and some experienced large declines in biomass in this period, though the nature of the species is not dissimilar to GB when compared with ESS, which showed the appearance of some qualitatively different species. For example,

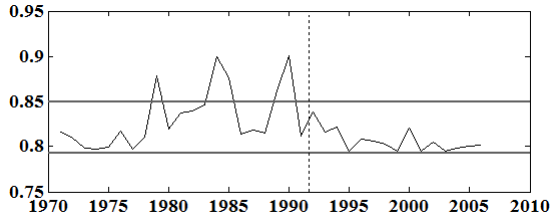
megrim (*Lepidorhombus whiffiagonis*) and solenette (*Buglossidium luteum*), not fished commercially, are also selected as being implicated by other groundfish decline. Such species would probably be less likely to be considered as indicator species of regime shifts elsewhere. Cod and haddock appeared to be more important in the NS, compared to ESS, which might suggest that catch is a more significant factor driving change in the NS, while on the ESS, it may be that other factors cause functional change in the ecosystem. However, here the *functional equivalence* search algorithm performed well in terms of identifying key species, associated with functional changes in the relevant regions which would be potentially beneficial when investigating the reliability of our modelling approach in terms of detecting signals of functional change.

4.4.2 Prediction Accuracy of Hidden Variable Models in the East Scotian Shelf and North Sea

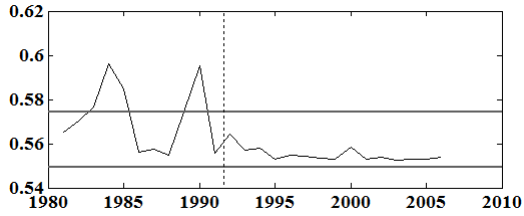
We now explore the hidden variable models for East Scotian Shelf (ESS) and North Sea (NS) learnt from the selected functionally equivalent species. We focus on the relationship of the hidden variable to the two regime metrics and the accuracy of the models in predicting species biomass.

The expected value of the hidden variable for ESS managed to capture some of the key predictive qualities of the metrics in terms of identifying a regime shift that is known to have occurred. The ESS hidden variable model (ARHMM) from Fig.4.3a demonstrates a large fluctuation between 1980 and 1990 with a steep increase in 1984 and 1989 prior to the time of the expected regime shift and it was then followed by a consistent decline following the collapse in 1992. The hidden variable increase coincides with a steep increase in variance (Fig.4.3c) in 1985, all above the 95% confidence upper interval. However, lowest p-value ($F_{21,13} = 11.59, p < .0001$) was reported at the time of the known collapse in 1992 (balanced design of the sliding window during the F-test). The assumption of homoscedasticity was not met ($F_{1,25} = 4.05, p < .05$), indicating variance inequality before and after the collapse. The autocorrelation showed little variation and it remained close to 1, as already illustrated for ecosystems undergoing a transition (Dakos et al. (2012)). To recall, according to theoretical expectations of critical slowing down, both

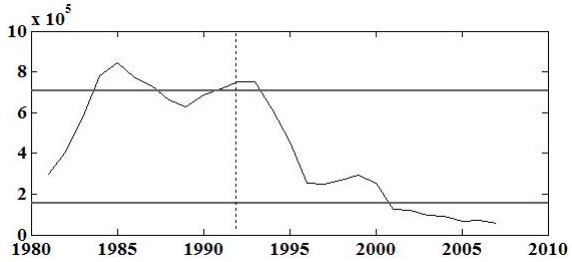
hidden variable and variance appear to increase prior to the expected regime shift and follow a consistent decline throughout time following the collapse correctly resulting in a clear early-warning signal to forewarn a major ecosystem change. After the addition of the metrics in the model, the hidden variable was more stable and still reflective of capturing the correct dynamics and characterised by rising trends in time prior to the expected transition (ARHMM + metrics in Fig.4.3b - Note this starts from 1980 due to the windowing required for calculating the metrics). In 1990, the lowest p-value was recorded ($F_{9,15} = 23.90, p < .0001$) which was actually lower than the p-value reported by the ARHMM, suggesting that the hidden variable in combination with the metrics might be performing better in earlier detection of change in the time series.



(a) Expected hidden value (ARHMM)



(b) Expected hidden value (ARHMM + metrics)



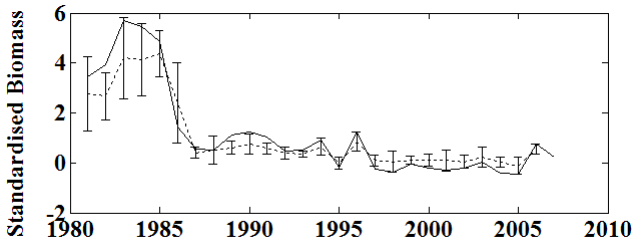
(c) Mean variance

Figure 4.3: The expected values of the discovered hidden variable from ARHMM (a), ARHMM + metrics (b) and mean variance (c) for ESS. The dashed line indicates the time of the regime shift in 1992. The solid line indicates upper and lower 95% confidence intervals, obtained from bootstrap predictions' mean and standard deviation.

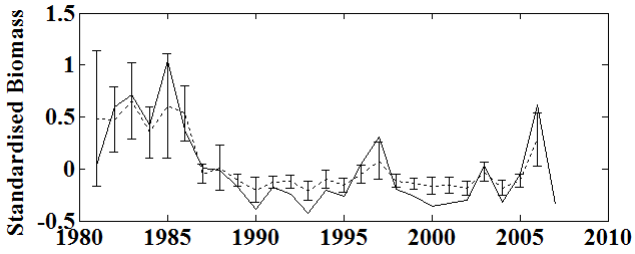
Although, the hidden variable in combination with the metrics performed well in terms of capturing early-warning signals of a regime shift that is known to have occurred,

the ARHMM + metrics model had some negative impact on the predictive performance of species biomass. Genuine species biomass variations throughout time were captured by both models, however, the ARHMM alone produced more accurate overall results of species biomass (SSE ARHMM: 4.83 and SSE ARHMM + metrics: 13.65).

The ARHMM + metrics managed to capture the variance in cod biomass, specifically the steep drop from late 1980s to early 1990s, followed by consistently low biomass predictions, due to the underlying functional changes occurring in the ecosystem (Fig.4.4a). Similar predictions were modelled for another species- the silver hake, prior the time of the regime shift the model captured the genuinely stable biomass but after the collapse, the model reflected on the declining and fluctuating trend of the biomass throughout time (Fig.4.4b).



(a) Cod, ESS

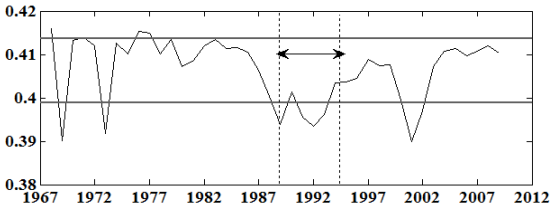


(b) Silver hake, ESS

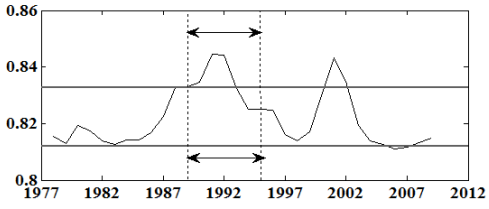
Figure 4.4: Biomass predictions generated by ARHMM + metrics of cod (a), and silver hake (b) for ESS region. 95% confidence intervals report bootstrap predictions' mean and standard deviation. Dashed line indicates predictions by the model, whilst solid indicates standardised observed biomass for the time period 1980-2006.

The expected value of the hidden variable for NS (Fig.4.5a) was characterised by some fluctuation up to early 1980s followed by a small decrease below the lower con-

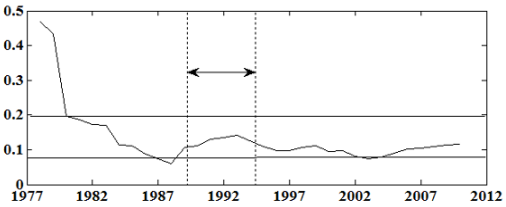
fidence level coinciding with the time around the functional changes in late 1980s to mid-1990s. Nevertheless, the F-test did not detect any significantly different changes in the slope of the hidden variable. These values are much smaller than for the expected values of the hidden variables in GB and ESS. Perhaps this is not surprising as it was found in (Tucker & Duplisea (2012)) that the hidden variable in the NS data did not seem to reflect a distinct regime shift and this fits with the general consensus that the NS has not suffered such a radical switch as the other two regions. Both hidden variable and variance (Fig.4.5c) show a trough in late 1980s which could be a reflection of the end of the "gadoid outburst" where groundfish were very abundant for about the previous 25 years (Beaugrand et al. (2003)). Here, the condition for equality of variance before and after the predicted functional change was fulfilled ($F_{1,31} = 1.40, p = 0.08$). The hidden variable from the ARHMM + metrics (Fig.4.5b) was more explicit and clear, finding the lowest p-value ($F_{10,12} = 0.27, p < .05$) in 1988 when first functional changes are believed to have occurred in the system according to experts. NS is a diverse system, subject to external anthropogenic forcing and internal environmental variation and as such, it is suggested that it seems to exhibit a range of discontinuous disturbances which would be more difficult to interpret by the hidden variable alone (**?**). However, the effect of the metrics on the hidden variable assisted in the correct identification of the time period where we would expect some functional change or disturbance in NS.



(a) Expected hidden value (ARHMM)



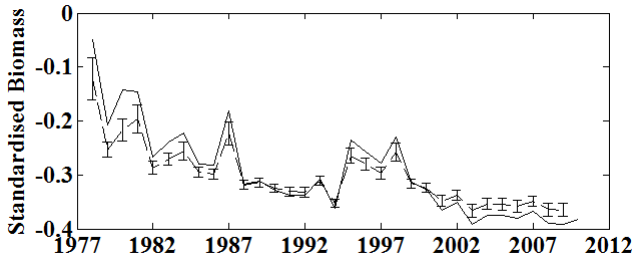
(b) Expected hidden value (ARHMM + metrics)



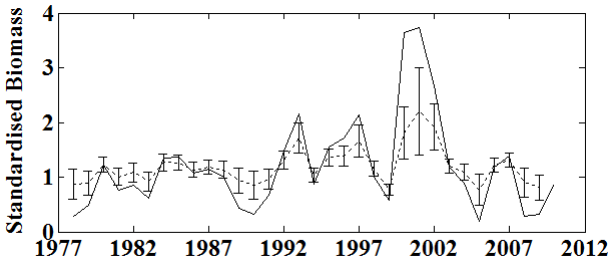
(c) Mean variance

Figure 4.5: The expected value of the discovered hidden variable from ARHMM (a), ARHMM + metrics (b) and mean variance (c) for NS. The dashed lines indicate the time period of the regime shift. The solid line indicates upper and lower 95% confidence intervals, obtained from bootstrap predictions' mean and standard deviation.

Results for NS showed reliable prediction of species biomass, with improved ability of the dynamic models when used in combination with the published metrics (SSE model: 5.50, SSE model+ metrics: 2.82). Let us look at the predictions for cod: we notice that the model in combination with the metrics has successfully managed to capture the genuine declining trend of cod biomass (Fig.4.6a), whilst for the haddock the model managed to reflect on the fluctuating trend of the biomass in time with some individual year effects in most recent years (Fig.4.6b).



(a) Cod, NS



(b) Haddock, NS

Figure 4.6: Biomass predictions generated by ARHMM + metrics of cod (a), and haddock (b) for NS region. 95% confidence intervals report bootstrap predictions' mean and standard deviation. Dashed line indicates predictions by the model, whilst solid indicates standardised observed biomass for the time period 1977-2009.

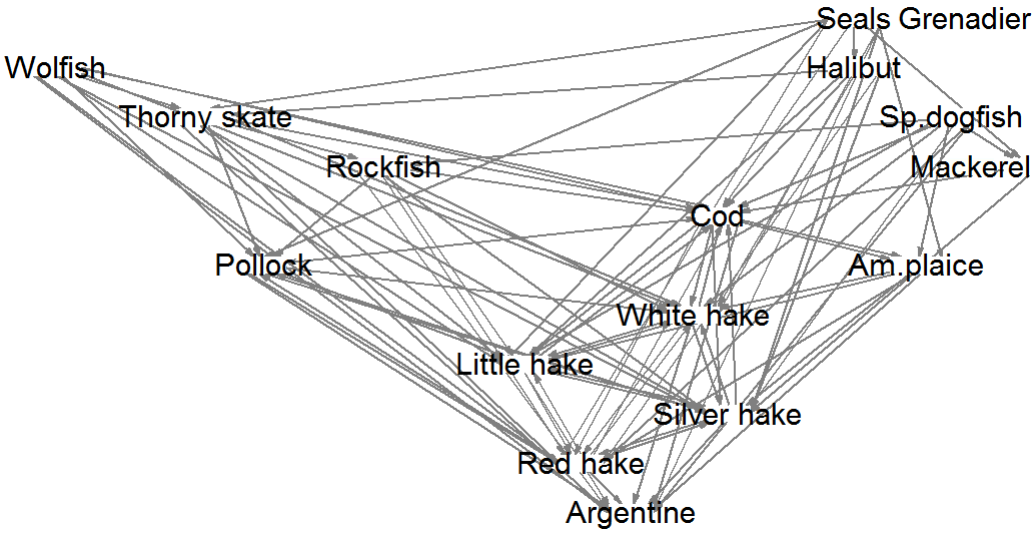
To summarise, the models that included the regime metrics performed better in terms of capturing the correct dynamics earlier in the time series. The hidden variable alone managed to reflect the ecosystem dynamics but that was more evident in the ESS region with a larger regime shift.

4.4.3 Comparison of Data-Learned Networks to Expert Diet Matrices in the East Scotian Shelf and North Sea

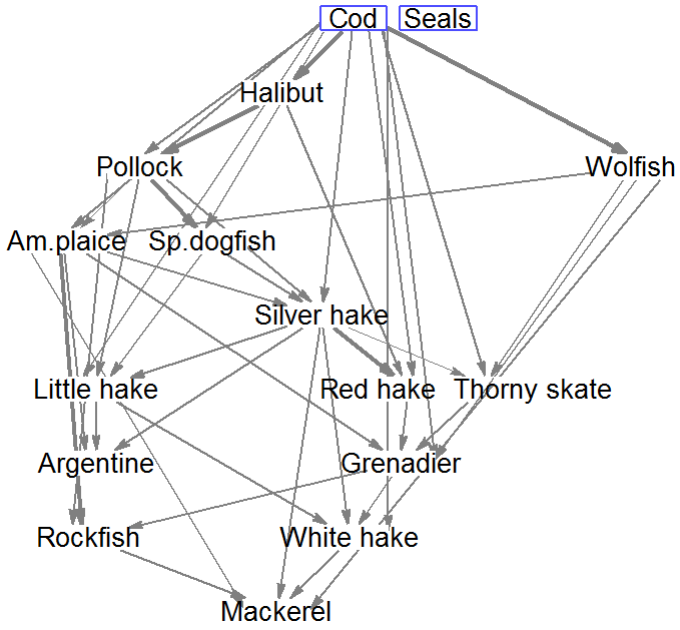
We now turn to the analysis of the learned networks by separating the data before and after the regime shift according to experts and comparing them to the networks generated by data split from our hidden variable models (timing identified from F-test significant results in the first part of the study). Some high confidence relationships were identified which represent likely models of the functional interactions between species.

The direction of the discovered significant links did not mean causation and it was not considered in the comparison with the generating diet matrix as we were interested in finding correctly identified species *associations*. Note that some of the discovered links, not directly relating to the diet matrix, could have been explained by either intermediate variables not included in the model or common observed effects acting on the model variables, however this was not the purpose here.

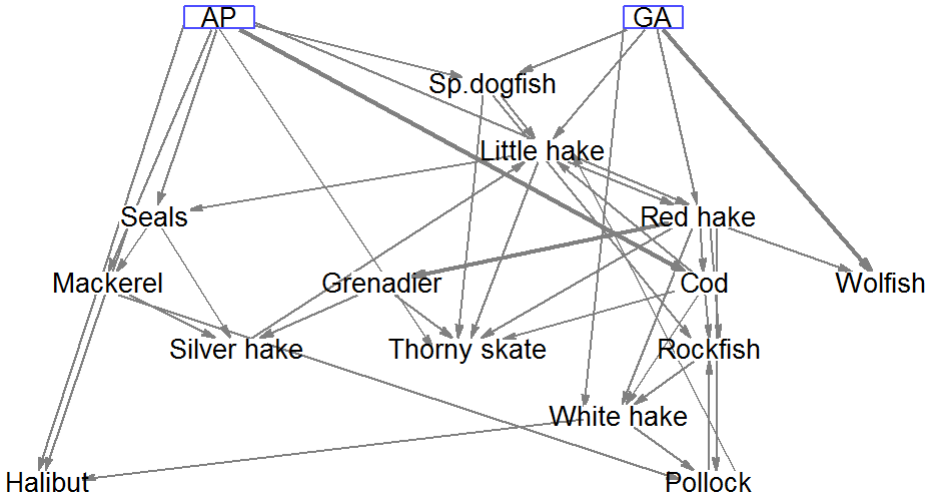
For ESS the learned network before the regime shift based on the experts' split was complex, identifying 7 significant features (four true positives) whilst the network after was rather simplified, finding only two significant links (one true positive), suggesting the influence of a radical switch in the system following the fisheries collapse. The network before 1990 (Fig.4.7b) (as found by ARHMM + metrics) identified 8 significant links (four true positives) and after (Fig.4.7c)- five significant links (four true positives). When comparing the networks of experts' split and data split, three of the significant links were preserved, one of them was a true positive. Learning the structure before and after the data split for ESS was a much better case in terms of detecting more correct associations with the diet matrix (Fig.4.7a). To recap, species selected in ESS were based on a regime shift in GB using the *functional equivalence* search, suggesting the successful algorithm performance in terms of capturing the correct structure and food web dynamics.



(a) Diet matrix, ESS



(b) Network before regime shift, ESS

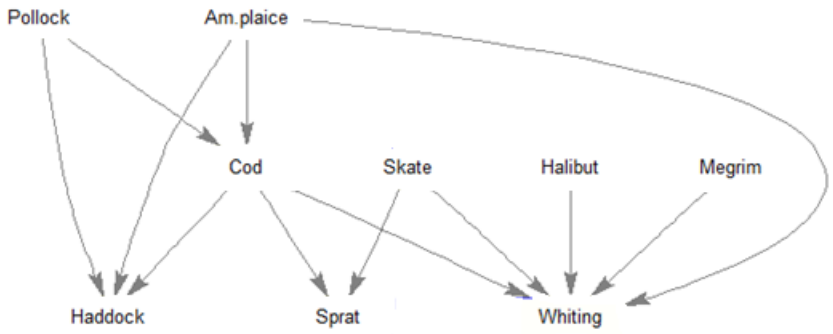


(c) Network after regime shift, ESS

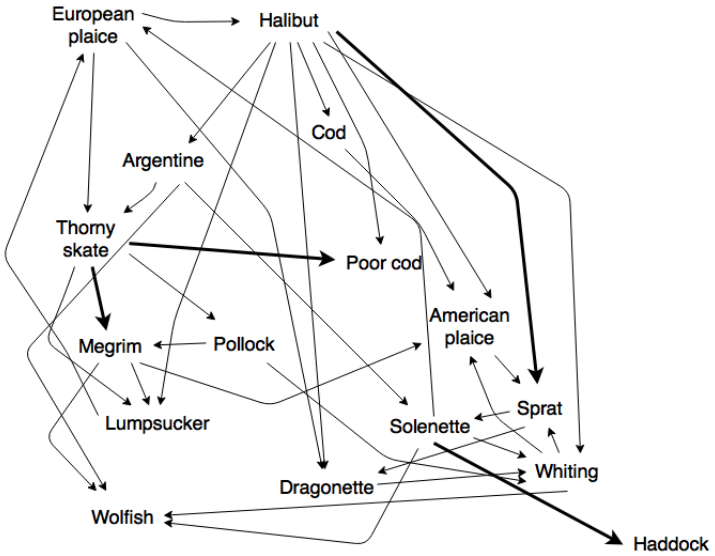
Figure 4.7: Diet matrix (a) with the network before (b) and after (c) the regime shift for ESS, generated by the data split using REVEAL. The width of edges corresponds to the computed confidence level (bold line: 0.5 and light line: 0.1). The squared nodes are significant themselves. For the diet matrix direction, of links represents predator-prey interactions. In bottom network (c): AP- American plaice and GA- Greater argentine.

For NS, the learned network before the experts' split identified five significant features (one true positive) and the network after- 7 significant features but none of them were true positives. The network before 1988 (as found by ARHMM + metrics) identified four significant links (Fig.4.8b) and after: one significant link, no true positives (Fig.4.8c). The relatively low number of correctly identified associations of the NS networks with the diet matrix compared to ESS, could be due to the possible influence of factors such as climate or fisheries exploitation that might have some common effects on different variables. But most importantly, the NS diet matrix was also relatively "poor" compared to ESS in terms of quantity of species recorded. However, we can see that the number of significant links was lower in the network after the regime shift, similarly to ESS, suggesting the potential influence of a functional collapse in this ecosystem. Finally, some of the discovered features of these networks could be due to indirect effects of modifications of food webs which would not be straightforward to detect by data collection alone.

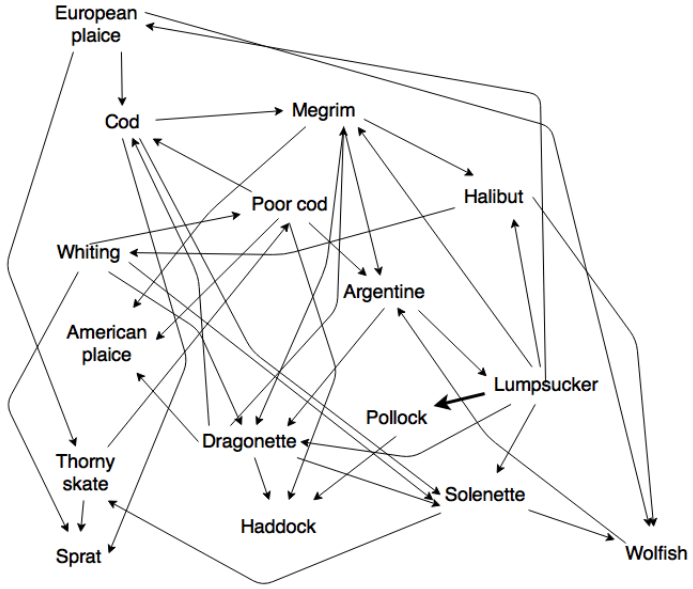
To summarise, the bootstrap methodology of learning the model structure in combination with the data split from our hidden variable models managed to detect pairwise relations of high confidence between species providing us with assumptions about the relevant food web structure and dynamics. Also, in both regions, significant links found before the data split, were generally reduced after, implying a signal of functional changes in the ecosystems.



(a) Diet matrix, NS



(b) Network before regime shift, NS



(c) Network after regime shift, NS

Figure 4.8: Diet matrix (a) with the network before (b) and after (c) the regime shift for NS, generated by the data split using REVEAL. The width of edges corresponds to the computed confidence level (bold line: 0.5 and light line: 0.1). For the diet matrix, direction of links represents predator-prey interactions.

4.5 Summary

In this chapter, we have explored the use of regime metrics in conjunction with hidden variables which proved useful (compared to those without) in terms of detecting *early-warning* signals (significantly rising variance and hidden variable fluctuations) of functional changes but it seemed to have an impact on biomass prediction. The hidden variables fitted to models that exclude these metrics appear to reflect some of their characteristics in terms of capturing the correct trophic dynamics.

Here, the use of machine-learning techniques is unique because it exploits functional equivalence between different datasets and uses the identified species in combination with a hidden dynamic model to predict functional collapse and species biomass. The recognition of a hidden variable in ecological studies is important because it may partially represent something external to the fish community which is not purely found

within the constrained model structure (Tucker & Duplisea (2012)). This is very different from more traditional statistical approaches whose fitting is conditioned completely upon the model structure. Changes in conditions external to the fish community such as oceanographic conditions may have driven the collapse in East Scotian Shelf. For the North Sea, what is occurring is less clear. It is a highly exploited and complex environment, which may be more variable and able to cope with disturbances, compared to other regions.

Here, the modelling approach differs from other methods in how correlative structures discovered in one region can be imposed upon another. The Bayesian network topology allows us to explore such structures explicitly prior to and after suspected regime changes. Overall, the learned network features managed to find some overlap with the diet matrices, though not many, maybe due to implicit correlations (and so more hidden variables may need to be structured into the models to deal with this). Nevertheless, the general finding was that prior to collapse there were more correctly identified links and these seemed to disappear after the regime shift. This can provide insights into why fished ecosystems collapse and why they sometimes do not recover. It may even give us an insight into signals of an imminent collapse perhaps while there is still time to prevent it.

In the following chapter, we further explore the use of a hidden variable by using data of external factors such as temperature and fisheries catch that might be influencing the fish community. We continue to use the functional network approach but now account for species interactions and associations with their environment across spatial scales. In addition, we develop a novel approach of modelling species dynamics by incorporating a second hidden variable to model unmeasured spatial effect.

Chapter 5

Spatio-temporal Bayesian Network Models for Revealing Trophic Dynamics and Functional Networks in Fisheries Ecology

5.1 Introduction

In recent decades, it has become clear that ecosystem structure and function can change over relatively short time scales (Scheffer et al. (2001)). Fishing pressure can change the structure of marine populations and consequently influence the nature of their responses to climate (Planque et al. (2010)), which could have impacts on the value of commercial fisheries (Perry et al. (2005)). Therefore being able to predict the dynamics of the species and their environment at spatially and temporally resolved scales, is of growing importance for the protection of natural biodiversity and human resources which poses new challenges for analytical tools and computational statistics (Aderhold et al. (2012)).

One way to understand ecosystem dynamics is examination of the functional relationships (such as prey-predator) between species along with their interaction with stressors such as temperature change and fisheries exploitation in their potential habitat (space)

and across time. In this way, learning functional relationships can provide a metric for assessing community structure and resilience in response to natural and anthropogenic influences, which is influential for future management options and long-term viability of populations (Gaston et al. (2000)).

The objective of this chapter is to model species dynamics and their interactions at geographically and temporally varied areas within two separate geographic regions: Gulf of St Lawrence and North Sea. Firstly, we examine how aggregated species interact at different spatial scales and over time within the Gulf to understand what mechanisms are involved in shaping the ecological networks and functional dynamics of food webs. Specifically, we explore how pre-defined functional relationships vary in time and space in order to better understand community structure.

Then, we move onto evaluating the potential usefulness of Bayesian inference for ecological data by examining the predictive capability of *different* dynamic Bayesian network (DBN) architectures for the North Sea. We correct for spatial autocorrelation by introducing a *spatial node*- a parent node representing the spatial neighbourhood of a node. We also account for latent effects by introducing two hidden variables- one *general* to detect overall change in the species biomass and another *specific* to capture spatial unmeasured effects. We produce a novel approach of modelling ecosystem dynamics that accounts for the *heterogeneous* nature of the driving factors within *each spatial area* and their changes over *time*. We examine the models' accuracy in predicting biomass, in response to any changes in temperature and fisheries catch or given there is a change in another species group biomass and therefore aid towards the better understanding of trophic dynamics in this complex ecosystem.

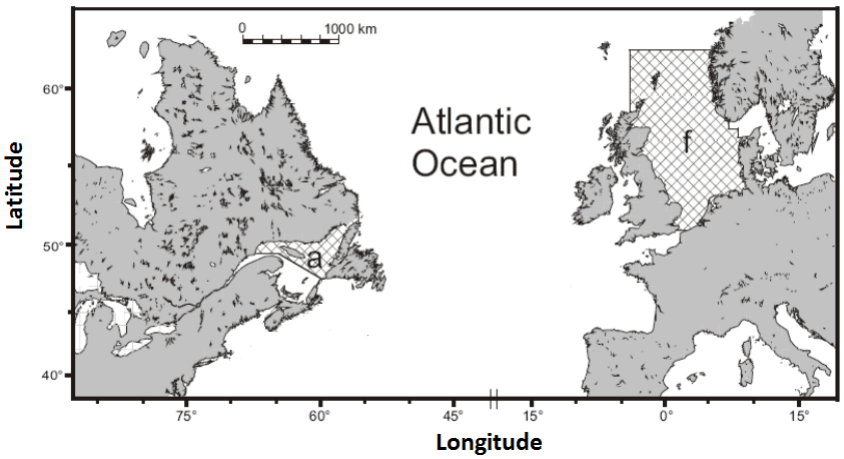


Figure 5.1: Regions (shaded area) of the Gulf of St Lawrence (a) and the North Sea (f).

5.2 Description of Spatio-temporal Survey Data to Model Functional Networks across Two Geographic Regions

5.2.1 Gulf of St Lawrence

We analysed data from the northern Gulf of St. Lawrence (48.00°N, 61.50°W, Fig.5.1a) groundfish and shrimp summer survey (1990-2013). The survey utilises a stratified random sampling design (Doubleday (1981)) with a standard tow using a benthic otter trawl. For each tow, all the fish were weighed and a sub-sample (200 individuals per species) was taken for computing length-frequency distributions. These length-frequency distributions were the basis of the data used here.

For the purpose of the analysis, the length-frequency distributions were used to calculate biomass over the same species and within the same year. Then, fish and invertebrate species were aggregated into the relevant trophic group by summing up the biomass (I-invertebrates, P- pelagics, SP- small piscivorous fish and LP- large piscivorous and top predators) (FishBase was used as a guidance point). This was performed for each of the 20 spatial clusters (explained in 5.3.1 *Introducing a Technique to Learn the Structure of Bayesian Network Models for the Gulf of St Lawrence and North Sea*) and for each year in the time window: 1990-2013, ending up with four variables for each spatial cluster across the time window. The data were standardised (sample mean removed from each

observation, which is then divided by the standard deviation) prior to conduction of the experiments.

5.2.2 North Sea

The analyses are based on the database of the International Bottom Trawl Survey (IBTS) for Quarter 1 (January to March), maintained by the International Council for the Exploration of the Sea (ICES) and conducted within ICES areas between 51-62° latitude (Fig.5.2, only areas 1 to 7 were considered in the study here due to limited quality and consistency of the data on the remaining spatial areas). These data are publicly available from the ICES Database of Trawl Surveys (DATRAS; www.ices.dk). The IBTS is a scientific fishing survey that follows a standard protocol: at each station, a GOV trawl is towed at 3 to 4 knots for a predefined duration. All species caught in relatively low numbers are counted and measured, while for very large catches, subsamples are taken and the resulting data scaled to the total catch. The data are recorded as length-frequencies by tow for each species and converted to catch per unit effort (CPUE; numbers per length class per hour) using tow durations.

In the study, CPUE was extracted for the time window: 1983-2010 and converted to biomass (kg per hour), using length-weight relationships and summing up over the same species and within the same year (www.fishbase.org). Next, fish species were aggregated by summing up the biomass into the relevant trophic group: pelagics (P), small piscivorous (SP) and large piscivorous and top predators (LP) (FishBase was used as a guidance point). Similarly to the Gulf, the nature of individual species summed into the trophic groups varied between the spatial areas but this was not of importance since they were always aggregated into the correct group. This was performed for each of the 7 areas and for each year in the time window. We also have available biomass data for different zooplankton species (source: ICES Working Group on Integrated Assessments of the North Sea- WGINOSE) but we decided to sum the biomass from selected copepod species to represent the zooplankton for the whole of North Sea in the time window 1983-2010.

Sea surface temperature (*temperature*) and net primary production (*Net PP*, refers

to the net production of carbon by primary level producers such as phytoplankton) data were model outputs (averages per area and year: 1983-2010) from the European Regional Seas Ecosystem Model (ERSEM; for more detail: Lenhart et al. (2010), van Leeuwen et al. (2013)) due to limitations in spatial and temporal resolution of these observations. Catch data (landings per area and year: 1983-2010), measured in tonnes live weight, were also obtained from the ICES database (ICES Historical Catch Statistics 1950-2010) but data was spatially combined and assigned to the northern (areas 1 and 3), central (areas 2, 4 and 7) and southern (areas 5 and 6) North Sea due to historical changes in collection and compilation of the landings data. This resulted in 6 observed (and continuous) variables: *catch*, *temperature*, *Net PP*, *P*, *SP* and *LP* for each spatial area and across the time window. Similarly to the Gulf, the data for NS was also standardised.

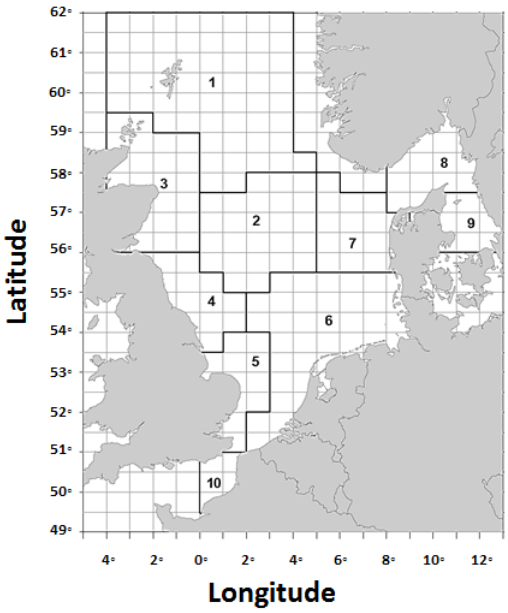


Figure 5.2: ICES statistical rectangles within the North Sea (areas 1 to 7 were used in this study). Source: ICES, Manual for the International Bottom Trawl Surveys.

5.3 Methods

5.3.1 Introducing a Technique to Learn the Structure of Bayesian Network Models for the Gulf of St Lawrence and North Sea

We learn the structure of static BNs from temporal data for each of the spatial regions (20 clusters for the Gulf of St Lawrence and 7 areas for the North Sea) by applying a hill-climb optimization technique. We perform the hill-climb with random restart (n=*10*) (for more detail refer to Chapter 3: *Preliminaries*). In this way, we apply the search for some number of iterations (n=*500*) until we hit a local maximum. We used the Bayesian Information Criterion (*BIC*, Schwarz et al. (1978)) for scoring candidate networks. In addition, to learn the network structure for each year in the time window, the hill-climbing was conducted on a window of data (size of window=*10*). In this way, we would be able to capture any significant functional interactions over the previous 10 years. The learned BN links represent dependence, these are spatial relationships that are predictive in an informative, not causal aspect (Milns et al. (2010)). We define a *confidence threshold*- the minimum confidence (estimate of the probability of finding a link) for an edge (or a link) to be accepted as an edge in the learned network structure. We found the window size and number of iterations through experimentation to be optimum when dealing with limited time series.

Gulf of St Lawrence A library of simple BNs, representing species interactions or functional relationships, based on expertise knowledge (Table 5.1, I-invertebrates, P-pelagics, SP-small piscivorous and LP-large piscivorous and top predators) was created. K-means (Hartigan & Wong (1979)) was applied to cluster the number of sampling stations (originally over 200 sampling stations per year, Fig.5.4a) on the mean latitude and longitude, resulting in 20 spatial clusters (or sub-regions, Fig.5.4b). Each spatial cluster was individually analysed to identify how the known functional relationships vary across time, but also to discover relationships between trophic groups, producing structures for 20 static BN models, equivalent to each one of the sub-regions in the Gulf of St. Lawrence oceanic area. Note that we detect the equivalence classes of each functional relationship and score the confidence of each relationship being in the

network over space and time. Adopting random restarts was preferably chosen compared to conditional independence tests for example as we wanted to learn the confidence of each functional relationship being in the network and not just examine the dependency relationships. We defined functional relationships of high confidence as those in which we have the greatest confidence of being in the network (*threshold* ≥ 0.3).

Table 5.1: Pre-defined Functional Relationships

Pre-defined Functional Relationships and Descriptions	
1. $I- > SP < -P$	Competition
2. $P < -I- > SP$	Predation
3. $P < -I- > SP, I- > LP$	Predation
4. $P < -I- > SP, P- > LP$	Predation
5. $P < -I- > SP- > LP, P- > LP$	Predation
6. $P < -I- > SP, LP < -SP- > P$	Intraguild Predation
7. $LP < -I- > P- > SP- > LP$	Omnivory
8. $P < -I- > SP- > LP$	Predation

North Sea For the North Sea, we first conducted an Autoregressive Hidden Markov model (ARHMM) and used the learned hidden variable from the ARHMM to incorporate it into the relevant spatial dataset (to recall 7 spatial areas) prior to the hill-climb to help us identify trophic interactions of high confidence between species groups and with their environment. Here, another hill-climb was also conducted to identify functional relationships between groups of species, without the influence of stressors. The trophic aggregated species matrix (*21*x*32*, 7 areas with three species groups each and longer time series) was conducted through the hill-climb for *4000* iterations. We defined interactions of high confidence in time as those in which we have the greatest mean confidence of being in the generated network (*threshold*$\geq$0.2). During this hill-climb, we generate a dynamic model, which we will refer to as global hidden DBN: **GHDBN**, the structure for which was imposed by only incorporating the resulting data learned species group interactions during the same hill-climb.

5.3.2 Spatial Autocorrelation

Spatial autocorrelation, the phenomenon that observations at spatially closer locations are more similar than observations at more distant observations, is nearly ubiquitous in ecology and can have a strong impact on statistical inference (Aderhold et al. (2012)). To incorporate potential spatial autocorrelation in our modelling approach, we connect each node in the network to an enforced parent node that represents the average biomass from the spatial neighbourhood (the three or four nearest neighbours) of the current geographic location (or area). Specifically, for the Gulf of St Lawrence, we produced two variants of our BN model: one that excludes a spatial node and one including the spatial node to compare the predictive performance between the two model variants.
In the North Sea, we enforce three parent nodes that represent the average biomass of the relevant trophic group (P, SP or LP) from the spatial neighbourhood of the current area. The same hill-climb procedure was used to learn the connections of each P, SP and LP spatial node in the relevant spatial network.

5.3.3 Generating Dynamic Bayesian Network Models to Predict Biomass in the Gulf of St Lawrence and North Sea

The experiments involved prediction of aggregated species biomass by inferring DBNs for each spatial area within the two separate geographic regions. The method of predicting the biomass was identical to the approach used in Chapter 4, (sub-section 4.3.2 *Hidden Variable Models to Predict Regime Shifts and Species Biomass*) and similarly, a non-parametric bootstrap (Friedman et al. (1999)) was applied *250* times for each variant of the model. Model performance, in terms of sum of squared error (SSE), was assessed for each model.

Gulf of St Lawrence For the Gulf, once the structure of the BN model was learned from the hill-climb, a DBN was inferred and two sets of experiments were conducted: one that excludes the spatial node (we will refer to this model variant as non-spatial DBN) and in the other, spatial node was included in the model (spatial DBN) to see if the node improves prediction.

North Sea For the North Sea, the experiments involved prediction of aggregated species biomass by inferring different dynamic Bayesian architectures: ARHMM, DBN, DBN with spatial nodes included (SDBN), hidden DBN (HDBN) and a novel modelling approach that we developed for this thesis: hidden DBN with spatial nodes included (HSDBN) (Table 5.2). The North Sea is genuinely a more complex marine region (compared to the Gulf) with dynamic climate-ocean interactions and ecological associations. That is why here, we evaluate the usefulness of different dynamic BN approaches with varying complexity and parameters, that might potentially reflect the different hypothesis of the North Sea system.

The network structure for ARHMM was fixed for all areas, whilst the network architectures for the other models were imposed by using structure learning from data in combination with external expertise on the North Sea zooplankton dynamics in the case of HDBN and HSDBN. The network architecture also varied within each spatial area.

Table 5.2: Summary of the applied dynamic BN models on the North Sea. Models are ordered based upon complexity level in regards to number of hidden variables (HVs) and spatial nodes.

Model	Name	HVs	Spatial nodes	Comments
1. ARHMM	Autoregressive Hidden Markov Model	1	None	Fixed structure
2. DBN	Dynamic Bayesian Network	1	None	Flexible structure but possible information loss due to having only a single HV
3. GHDBN	Global Hidden Dynamic Bayesian Network	1	None	Highlights trophic dynamics at a wider scale
4. SDBN	Spatial Dynamic Bayesian Network	1	3	Highlights the spatial effect
5. HDBN	Hidden Dynamic Bayesian Network	2	None	Highlights heterogeneity of the driving factors
6. HSDBN	Hidden Spatial Dynamic Bayesian Network	2	3	Highlights heterogeneity of the driving factors and the spatial effect

Here, we are also interested in predicting a pre-selected variable (here species dynamics, represented by the hidden variable) from each modelling approach based on the values of other variables (here biomass of species groups). Computation of the hidden variable has already been discussed in Chapter 4 and similarly, it was parameterised using the Expectation Maximization algorithm.

The hidden variable from every model was statistically tested for the presence of an increasing or decreasing trend using the Mann Kendall test (Jennings et al. (2002)) and the distribution of the hidden variable was compared to the observed variable it might represent (e.g. group of species biomass or zooplankton biomass) using the Mann-Whitney U test. The null hypothesis, that is being tested by Mann-Whitney U, is that the distributions of two groups (here, hidden variable and biomass) are identical. Note, if the p value is small ($p \leq .05$), we reject the null hypothesis and conclude that the two distributions are distinct, however if the p value is large ($p \geq .05$), the data do not give us any reason to reject the null hypothesis which does not necessarily mean that the two distributions are identical but there is no compelling evidence that they differ significantly (Cheung & Klotz (1997)). All statistical tests were reported at 5% significance level.

5.3.4 DBN with Two Hidden Variables and Spatial Nodes: HSDBN

We designed a dynamic BN model in which we incorporated two hidden variables: one discrete to model the general trophic dynamics (*general HV*) and another continuous specific hidden variable (*specific HV*) to see if we can learn the trophic level of zooplankton, which is missing due to limited spatial resolution but will be validated against the measured zooplankton for the whole of North Sea. This approach discourages the appearance of implicit interactions by including the unobserved hidden variables. This is the hidden DBN: **HDBN**, for which we propose a balanced architecture between structure learning from data and experts' knowledge on zooplankton dynamics, represented in this model by the *specific HV*. The HDBN is a functional network model in which nodes represent species groups and edges (connections between nodes) represent potential interactions of species groups with other groups, with abiotic factors

(e.g. temperature) and anthropogenic factors (e.g. commercial catch) that influence the species groups biomass. The strengths of such influences vary geographically and temporally.

Next, we further improved the HDBN by allowing for spatial autocorrelation and learning another dynamic BN model: **HSDBN**. Specifically, we enforced three parent nodes: *P sp.*, *SP sp.* and *LP sp.*, that represent the spatial values (of the equivalent *P*, *SP* and *LP* variables) to account for the effect of spatial autocorrelation. In the HSDBN modelling approach, we account for the heterogeneous nature of the driving factors (both biotic and abiotic) within each area and their changes over time. Hence, we can explore multiple species associations, model their dynamics with interactions from external stressors such as temperature and fisheries exploitation and compare the predictive performance of the model with HDBN and other probabilistic models already being used in the literature. The HSDBN structure for each area varies but the general form is presented in Fig.5.3. We incorporated the same unobserved hidden variables from the HDBN: *general HV* and *specific HV*. The *specific HV* was always connected to *P* and *Net PP* and *temperature* to *Net PP* (according to expert). In this way, we can inspect the state of the two HVs to see if they reflect any spatial and temporal changes in the trophic dynamics or capture spatial unmeasured effects which are not purely found within the constrained model structure.

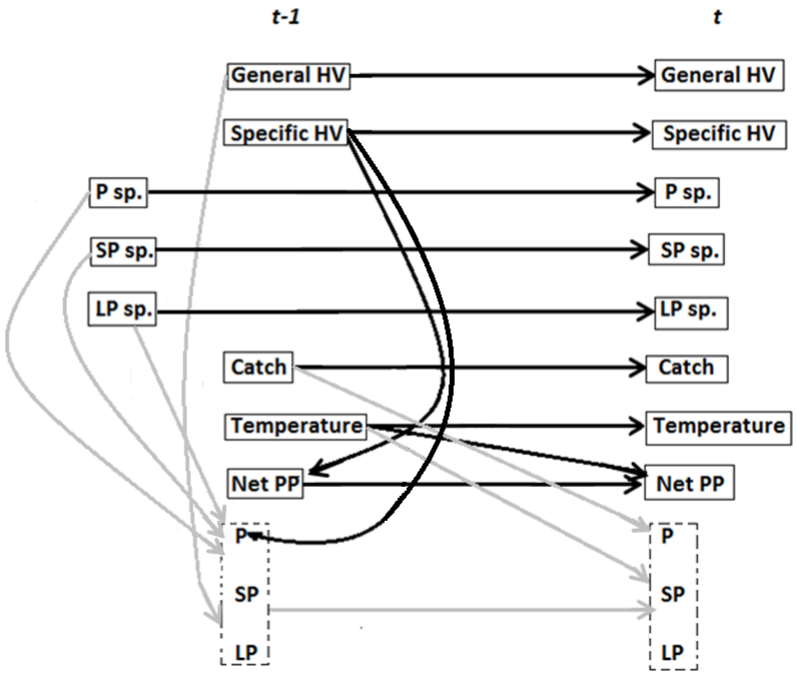


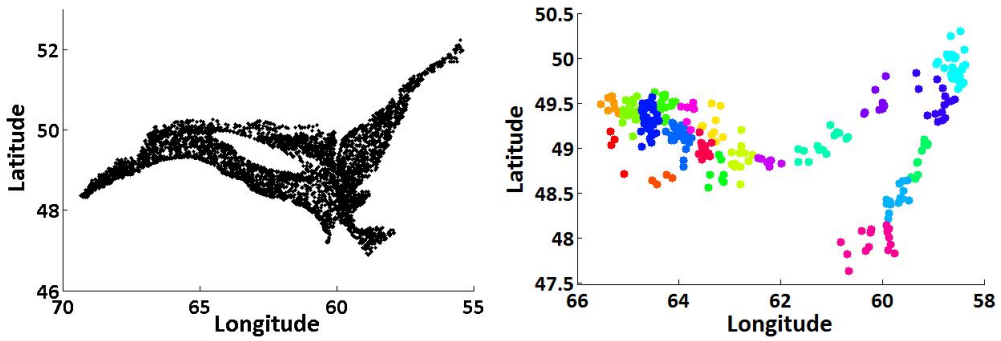
Figure 5.3: General structural form of the HSDBN model. Solid line represents fixed edges across areas. The three spatial nodes (*P sp.*, *SP sp.*, *LP sp.*), *general HV*, *catch* and *temperature* are individually linked to either *P*, *SP* or *LP* (represented by the dotted surrounding), depending on the spatial area (grey line). Connectivity between *P*, *SP* and *LP* also differs spatially. Edges between nodes (or variables) represent dependence relationships.

5.4 Results for the Gulf of St Lawrence

To recall, K-means was applied to cluster the number of sampling stations (Fig.5.4a), resulting in 20 spatial clusters (or sub-regions) within the region of St Lawrence (Fig.5.4b). Note that differences in density of the clustered stations could explain the slight spatial contrast between Fig.5.4a and Fig.5.4b.

5.4.1 Identified Functional Relationships within Spatio-temporal Scales

We now examine how learned by the hill-climb relationships amongst trophic groups of species vary across time and space. The relationship between invertebrates and pelag-



(a) Sampling stations before clustering (b) Sampling stations after clustering

Figure 5.4: Locations of the sampling stations before clustering (a) and after clustering in the region of St Lawrence(b).

ics (*I-P*) was found to be strongly significant (range: 0.3-1) and consistent in time and space (Fig.5.5a,b). Cluster 7 was the only cluster in which the relationship was found throughout the entire time series and in cluster 5 the relationship was found to be with highest confidence throughout time. Temporally, the confidence for the *I-P* relationship in majority of the clusters was found to be generally increasing with a small decline over recent years.

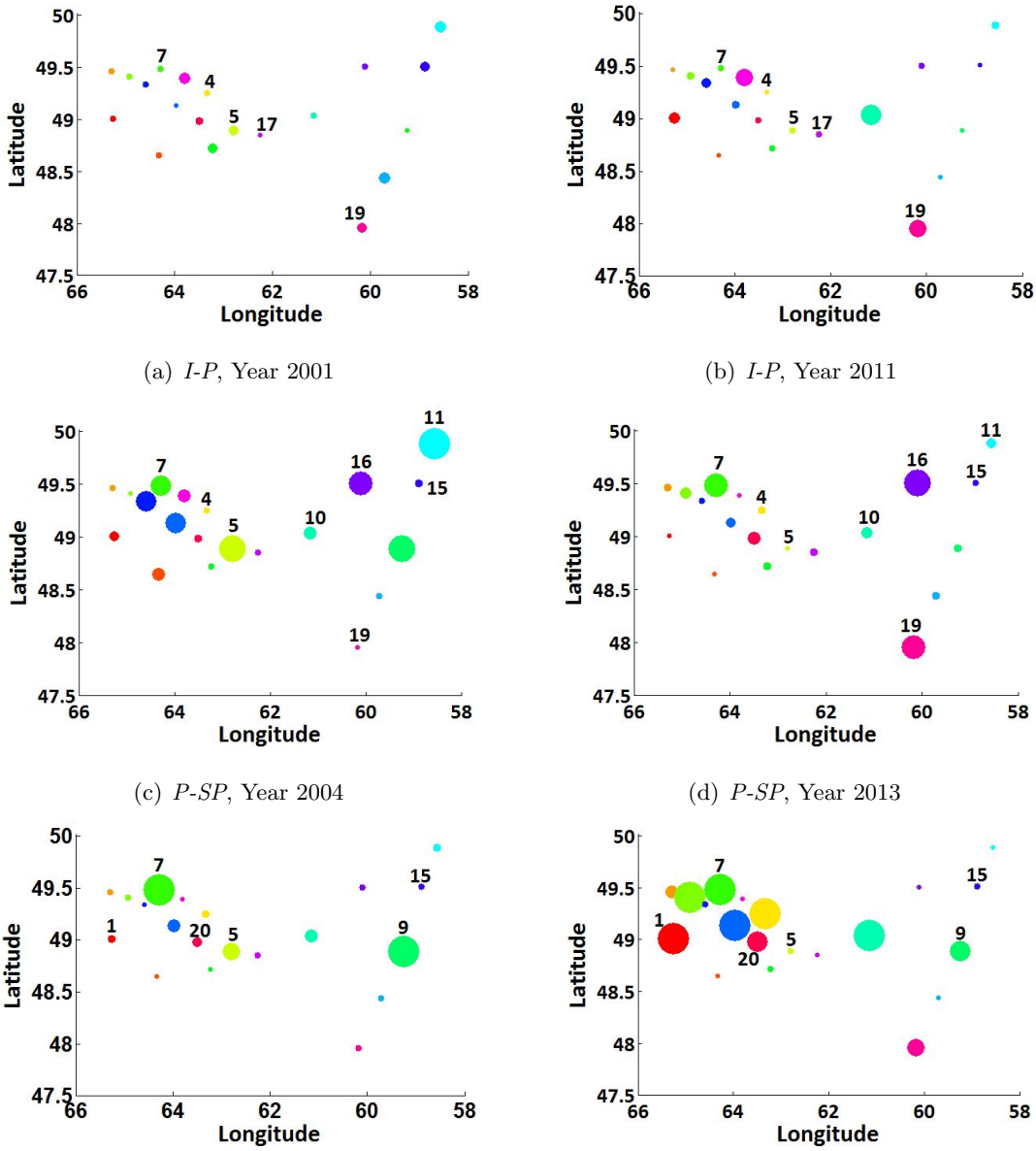
The invertebrates- small piscivorous fish (*I-SP*) relationship had the highest confidence throughout time in cluster 4. The relationship was relatively consistent in time for individual clusters, although we notice that different clusters were characterised with some specific temporal trends. For example, cluster 5 had a distinguished increase in early to mid-2000, whilst the opposite was found for cluster 19 (Fig.5.6a).

For cluster 19, the invertebrates- large predators (*I-LP*) relationship was identified throughout the entire time series but the most highly significant confidence was found for cluster 17 (range: 0.3-0.8). Temporally, both relationships: *I-SP* and *I-LP*, for majority of clusters were relatively stable but with declining trend at end of the time series.

We now consider the pelagics- small piscivorous fish (*P-SP*) relationship (Fig.5.5c,d). As with the *I-P* relationship, here *P-SP* was also the most highly confident for cluster 5 (range: 0.3-1). This *P-SP* relationship was highly consistent in time for clusters 4 and 16, in which the relationship was found throughout the entire time series. Compared

to *P-SP*, for the pelagics- large predators (*P-LP*) relationship, cluster 10 was the one in which the relationship was highly confident (range: 0.3-1). However, cluster 5 was the one in which the relationship was consistent throughout time. Across time, both relationships varied for the different clusters and it was difficult to find any temporal trends. However, some clusters declined around 2007 to 2010 (for example 7, Fig.5.6b) whilst clusters 5 and 19 (Fig.5.6b) increased around the same time and in most recent years. The increase around early to mid-2000, that we saw for *I-SP*, was also found for cluster 7.

The most highly confident small piscivorous- large predators (*SP-LP*) relationship (Fig.5.5e,f) in time that was also consistent in the series was found for cluster 20 (range: 0.3-1). The relationship was also consistent in time for cluster 9. Across time, similarly to the previous relationship, some clusters were relatively stable but some decline occurred around 2007 onwards (cluster 5, Fig.5.6c), whilst in other clusters increase in confidence was found for more recent years (for example cluster 7, Fig.5.6c). Cluster 19 was characterised with relatively low confidence throughout time (Fig.5.6c).



(a) *I-P*, Year 2001

(b) *I-P*, Year 2011

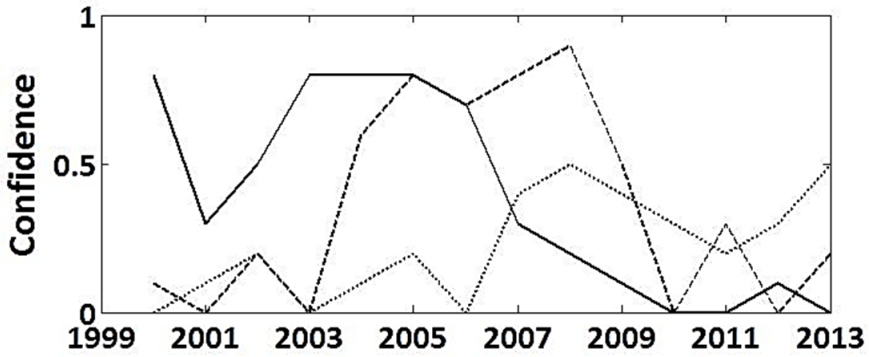
(c) *P-SP*, Year 2004

(d) *P-SP*, Year 2013

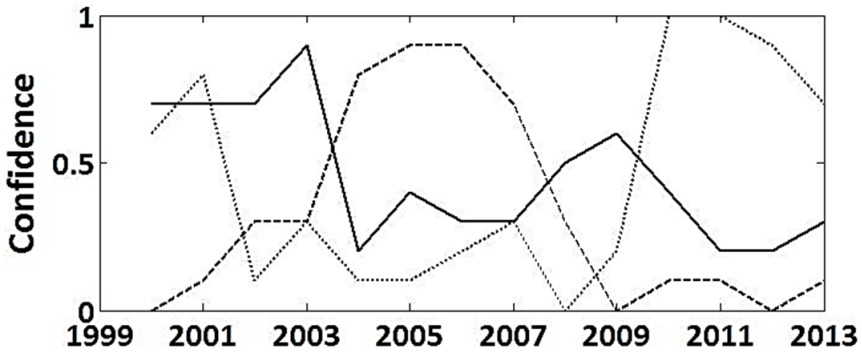
(e) *SP-LP*, Year 2008

(f) *SP-LP*, Year 2013

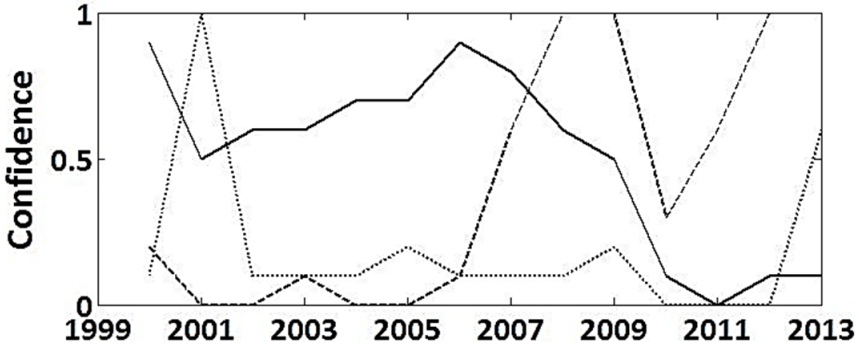
Figure 5.5: The learned *I-P*, *P-SP* and *SP-LP* relationships for all 20 spatial clusters (size of scattered bubbles is equivalent to the estimated confidence by the hill-climb). The clusters mentioned in 5.4.1 are numbered.



(a) *I-SP*



(b) *P-LP*



(c) *SP-LP*

Figure 5.6: The learned *I-SP*, *P-LP* and *SP-LP* relationships for clusters 5, 7 and 19 (represented by solid, dash and dot line respectively) for the time window: 2000-2013.

Next, we consider the variation of the pre-defined known functional relationships (Table 5.1) temporally and spatially. First, *function 1* and *2* were identified in all clusters. However, the significance of both functions varied across time with some

consistency in terms of spatial clusters. We find the emergence of "characteristic scales" of functional relationships, identified at spatially-specific geographic scales. Temporally, there was some decline in the significance of *function 1* and *function 2*, specifically in more recent years: 2010 to 2013 in all clusters. At the same time clusters like 9, 5 and 20 were found to be with relatively strong significance throughout time, outlining the importance of habitat quality at specific locations implying that in some regions prey are more affected by predators than in others. *Function 3* and *4*, *5*, *6*, *7* and *8* were not identified for all clusters and were only found in some years. However, again there was some spatial consistency in terms of different functions identified outlining only some clusters, highlighting the fact that relationships are scale dependant but also the importance of functional relationships for the local food web dynamics and structure. Other possible explanations include species abundance and distributional changes but in either case fishing could have had an important role.

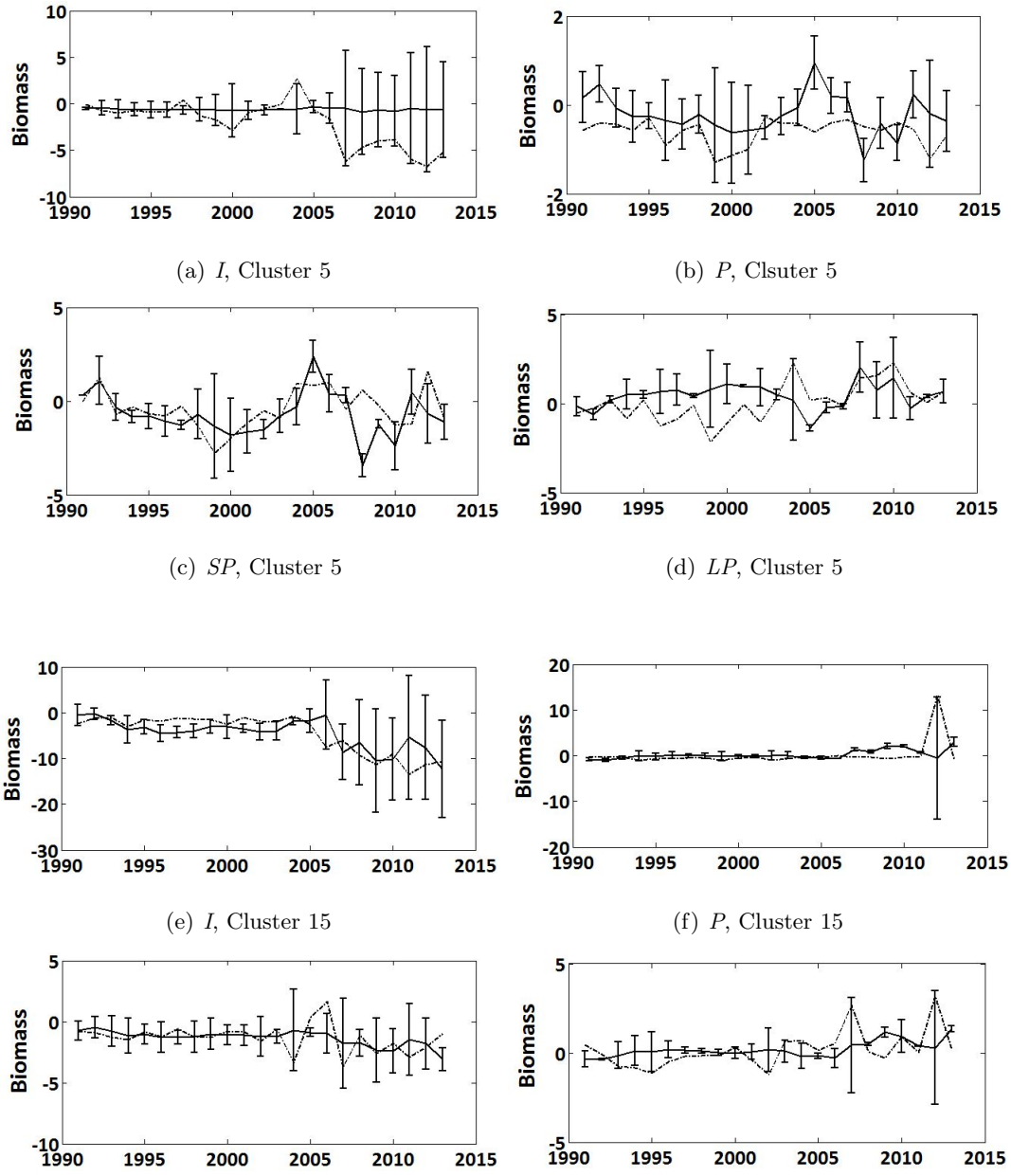
5.4.2 Prediction Performance of Dynamic and Spatial Dynamic Models

We now turn to the generated predictions by the DBNs for each spatial cluster. To recall, two variants of each model were produced: non-spatial DBN, excluding the spatial node and spatial DBN in which the spatial node was enforced and connected to each one of the other variables. Predictive performance between the two model variants was compared (Table 5.3). In general, predictive accuracy was improved once the spatial node was included in the model. Only for some clusters (6, 11, 17 and 18), better predictions were reported by the non-spatial DBN. In some clusters (for example 5 and 15), the predictive accuracy was significantly improved by the spatial DBN. Some clusters were genuinely easy to predict by both model variants, for example clusters 2 and 17. Conversely, some clusters were characterised by both models with lower predictive accuracy, for example clusters 9 and 10. The discovered spatial heterogeneity here in terms of the varying spatially predictive accuracy is a reflection of some of the mechanisms involved in shaping the local population dynamics. Such results can potentially provide us with insights on the ecological stability and resilience of the species dynamics and structure.

Table 5.3: SSE of DBN and spatial DBN. 95% confidence intervals reported in brackets

DBN	**spatial DBN**	**DBN**	**spatial DBN**
1. 5.58 (±9.29)	**1.** 4.38 (±7.08)	**11.** 12.44 (±20.56)	**11.** 16.54 (±34.34)
2. 0.24 (±0.36)	**2.** 0.14 (±0.12)	**12.** 69.90 (±308.02)	**12.** 30.55 (±64.30)
3. 16.20 (±29.92)	**3.** 10.76 (±17.16)	**13.** 12.68 (±16.63)	**13.** 12.06 (±9.63)
4. 10.09 (±14.70)	**4.** 9.68 (±12.58)	**14.** 196.11 (±271.68)	**14.** 109.37 (±102.42)
5. 44.20 (±51.17)	**5.** 11.27 (±12.47)	**15.** 77.45 (±605.26)	**15.** 23.62 (±47.80)
6. 20.20 (±40.42)	**6.** 20.22 (±34.29)	**16.** 17.15 (±18.40)	**16.** 14.86 (±13.46)
7. 25.29 (±55.38)	**7.** 19.47 (±26.86)	**17.** 5.88 (±8.78)	**17.** 6.12 (±6.67)
8. 38.72 (±46.22)	**8.** 19.59 (±11.78)	**18.** 2.68 (±3.94)	**18.** 3.43 (±3.72)
9. 125.19 (±240.49)	**9.** 92.14 (±111.67)	**19**. 80.32 (±112.90)	**19.** 77.19 (±72.43)
10. 104.31 (±167.02)	**10.** 60.62 (±62.38)	**20.** 13.20 (±22.08)	**20.** 10.70 (±13.64)

Now, let us look at some generated predictions for *I*, *P*, *SP* and *LP* throughout time for some of the clusters (e.g. 5 and 15). We notice that the trophic group of the invertebrates- *I*, was generally harder to predict (Fig.5.7a,e), comparing to the other functional groups, which could be nature of the aggregated species and specifically, their small size and high mortality rate compared to species from the groups of *SP* and *LP*. We also notice that it was harder to distinguish any temporal trends but rather the species groups were characterised with temporal variability and strong individual year effects.



(a) *I*, Cluster 5

(b) *P*, Clsuter 5

(c) *SP*, Cluster 5

(d) *LP*, Cluster 5

(e) *I*, Cluster 15

(f) *P*, Cluster 15

(g) *SP*, Cluster 15

(h) *LP*, Cluster 15

Figure 5.7: Biomass predictions generated by the spatial DBN for clusters 5 and 15 for the four trophic groups: *I*, *P*, *SP* and *LP*. Solid line indicates predictions and dash-dot line indicates standardised observed biomass. 95% confidence intervals report bootstrap predictions' mean and standard deviation.

5.4.3 Summary for the Gulf of St Lawrence

In this work, we have exploited the use of BNs with *spatial nodes* in order to identify patterns of functional relationships which proved significant in terms of predictive accuracy of our models, further concluding the spatial heterogeneity in this oceanic region. Our results showed *spatially* and *temporally* differentiated ecological networks. These networks indicate spatial relationship of species and habitat with the particular mechanisms varying from facilitation through trophic interactions. While the precise explanation behind the varying spatio-temporal confidence of some of the discovered relationships is not known, we expect them to be reflective of the underlying interactions within the community, thus suggesting similarity to the majority of the weak and some strong interactions expected of stable systems (Milns et al. (2010)).

Overall, the identified functional relationships were found to be consistently confident in time however we notice the spatially-specific differentiation. Such spatial heterogeneity could result from habitat fragmentation leading to decreased dispersal or the optimal habitat being located in a more restricted area, leading to increased aggregation (Frisk et al. (2011)). Individual year effects are very strong for this area as time increases, as already suggested by Duplisea & Castonguay (2006), which makes it difficult to determine temporal trends. However, some of the clusters' temporal increase in early to mid-2000 (for example, cluster 5 and 7 from Fig.5.6) could be owed to the fisheries moratorium in the area placed in 1994. In addition, our findings of recent temporal decline for some of the clusters' relationships (*P-LP* (cluster 5), *SP-LP* (cluster 19), Fig.5.6) we suggest to be due to predation release of small abundant species by the selective fishing of larger predators (Frisk et al. (2011)). Note again, the temporal variation of the systems was set apart in geographically-specific order, possibly due to site-specific fisheries exploitation targeting particular species.

Overall, we found high spatial heterogeneity in terms of the models' predictive accuracy, which we suggest to be due to the mechanisms involved in shaping the underlying species dynamics. For example resource availability, habitat selection, processes like dispersal and metapopulation effects (Frisk et al. (2011)) but also commercial fishing could have influence on the local community stability and structure, resulting in our

modelling approach identifying spatially-specific differences. The fact that our models managed to predict some clusters more accurately compared to others suggests that there are differences in the local population dynamics in terms of resilience to fisheries exploitation and climate. This could potentially provide managers with information on where commercial effort should be directed and what species to be targeted. Most importantly, we managed to show that once the spatial node is included in the model, predictive accuracy is improved and this highlights the general knowledge that one cluster's dynamics is influencing another, thus highlighting the need to account for spatial connectivity.

5.5 Results for the North Sea

5.5.1 Comparative Evaluation of Biomass Predictions

To recall, the predictive capability of the applied models is measured by how accurate are the modelled biomass predictions (sum of squared error estimation) comparing to the observed standardised biomass.

The results on biomass predictions of species groups showed great variability in their predictive accuracy from the applied probabilistic models: ARHMM (Table 5.4a), DBN (Table 5.4b), SDBN (Table 5.4c), HDBN (Table 5.4d) and HSDBN (Table 5.4e). Comparison of the predictive performance across all model types indicates varying spatially predictive accuracy which is a reflection of the models' features in combination with the spatially specific characteristics of each area and species aggregation.

When comparing the overall biomass (least SSE per area), the Hidden Spatial Dynamic Bayesian Network model (HSDBN) (Table 5.4e) managed to perform most accurately in certain spatial areas (1, 3, 4 and 6), compared to the other tested modelling approaches, which is reassuring that the inference scheme can handle the increased model complexity. HSDBN reported predictions with highest accuracy for most of the individual species groups (look at * in Table 5.4e). Interestingly, *P* and *SP* species were predicted more accurately compared to *LP*, highlighting the importance of species-specific effects in their habitat following external disturbances. For the remaining areas (2, 5 and 7), the Spatial Dynamic Bayesian Network model (SDBN, Table 5.4c) produced

better overall predictions. Although the general improvement in predictive accuracy of the HSDBN model over the competing probabilistic models, we notice the similar level of accuracy (least SSE difference: $\leq$5.0, between the generated overall predictions of two models) for some of the areas. For example, the Hidden Dynamic Bayesian Network (HDBN) and Dynamic Bayesian Network models (DBN) performed respectively with a similar level of accuracy, following the HSDBN for areas 3 and 4 (Table 5.4b, 5.4d, 5.4e).

Finally, the biomass predictions generated during the BN structure learning for the whole of North Sea by the Global Hidden Dynamic Bayesian Network model (GHDBN, Table 5.4f) were overall less accurate compared to HSDBN. Interestingly, the GHDBN performed with similar level of accuracy to the SDBN for areas 2 and 7, confirming the significance of the spatial relationship between these areas and their neighbours. The GHDBN will not be further addressed as a competing model in the discussion, we simply wanted to state the overall predictive accuracy during the learning process of a static BN.

Table 5.4: Sum of Squared Error (SSE) of *P*, *SP* and *LP* biomass predictions generated by ARHMM (a), DBN (b), SDBN (c), HDBN (d) and HSDBN (e). The * indicates most accurate predictions for individual species groups.

(a) ARHMM

Area	**P**	**SP**	**LP**
1.	24.79	24.71	27.29
2.	20.57	31.02	20.05
3.	21.87	20.16	25.68
4.	33.14	19.54*	33.07
5.	27.38	21.52	26.68
6.	30.12	25.52	12.79*
7.	24.53	22.49	27.27

(b) DBN

Area	**P**	**SP**	**LP**
1.	22.88*	20.31	33.86
2.	32.82	33.95	26.55
3.	21.51	15.35	22.57
4.	25.80	22.65	26.71
5.	27.02	32.28	32.38
6.	26.82	30.95	20.88
7.	27.76	25.77	24.35*

(c) SDBN

Area	P	SP	LP
1.	34.26	25.66	27.17
2.	13.13*	30.38	17.04*
3.	25.16	19.76	30.38
4.	26.83	26.25	27.48
5.	27.31	20.77	26.86
6.	29.41	20.21	14.72
7.	23.75*	21.09*	27.36

(d) HDBN

Area	P	SP	LP
1.	23.52	24.33	33.54
2.	20.92	31.99	24.20
3.	19.15	19.40	20.52*
4.	26.46	23.08	25.02
5.	31.05	22.93	25.14*
6.	27.09	26.52	21.15
7.	29.21	28.50	32.71

(e) HSDBN

Area	P	SP	LP
1.	23.46	20.11*	25.46*
2.	20.38	30.05*	25.03
3.	18.32*	14.90*	21.59
4.	25.65*	21.35	23.69*
5.	26.75*	20.76*	34.16
6.	24.85*	19.48*	14.67
7.	30.86	22.53	30.14

(f) GHDBN

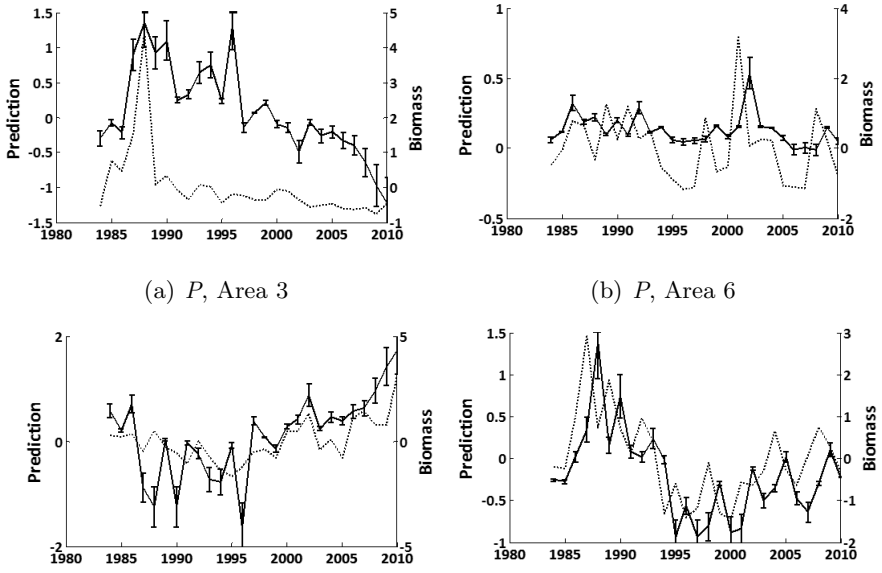
Area	P	SP	LP
1.	26.84	24.03	27.69
2.	14.65	29.20	19.90
3.	27.07	29.33	20.99
4.	26.08	23.72	28.56
5.	27.66	24.50	31.17
6.	26.19	21.31	15.14
7.	27.92	16.16	29.74

We now look at the groups of species: pelagics (P), small piscivorous (SP) and large piscivorous (LP) and their biomass predictions in time for some of the areas on which the HSDBN performed most accurately. The imposed HSDBN network structures for these areas are shown in the Appendix. We also illustrate our model's predictive accuracy through the use of "what if" type descriptions of the network structures in these areas in response to actual data changes (ICES DATRAS database and ERSEM model outputs).

The HSDBN managed to capture the species biomass variations in time for **area 3** (Fig.5.8a,c,e), specifically reflecting on the general decline of the P group, although,

failing to pick up some of the outlier observations for P and LP. Our HSDBN model predicted the general P and late 1990s SP biomass decline (Fig.5.8a,c) when the *fisheries catch* was high, however this started to change in more recent years in response to *fisheries catch* becoming low, which led to some increase only in the SP biomass. Interestingly, we notice some similarity in the reproduced individual year effects between SP and LP (Fig.5.8c,e), which coincided with increase in the SP biomass and surrounding *spatial* P biomass in recent years.

For **area 6**, even some of the outliers, that the HSDBN did not detect perfectly, we can see from Fig.5.8b,d,f, that the model reflected on the temporal biomass variations of P and explicit decline of the SP and LP groups. Similarly to area 1, our model was able to re-create the decline in the SP and LP biomass when *fisheries catch* started to decline in late 1990s to early 2000 and the surrounding *spatial* P biomass started to increase. Again similarly to area 1, the reproduced individual year effects for the P biomass were relatively strong, although we see some increase in early 2000 when the surrounding *spatial SP* biomass started to increase but the *spatial LP* biomass continued to decline.

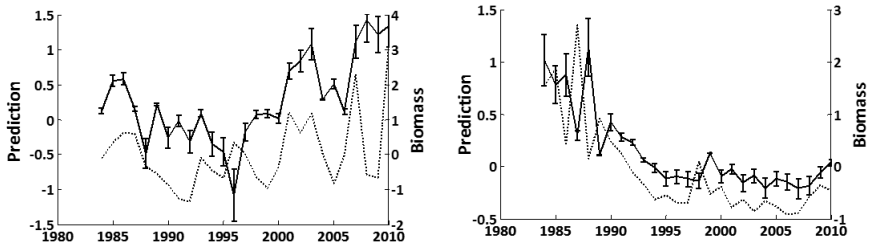


(a) P, Area 3

(b) P, Area 6

(c) SP, Area 3

(d) SP, Area 6



(e) *LP*, Area 3 (f) *LP*, Area 6

Figure 5.8: Biomass predictions of *P*, *SP* and *LP* generated by the HSDBN for areas 3 (a,c,e) and 6 (b,d,f). Solid line indicates predictions and dotted line indicates standardised observed biomass. 95% confidence intervals report bootstrap predictions' mean and standard deviation. Note the negative scale is due to standardisation.

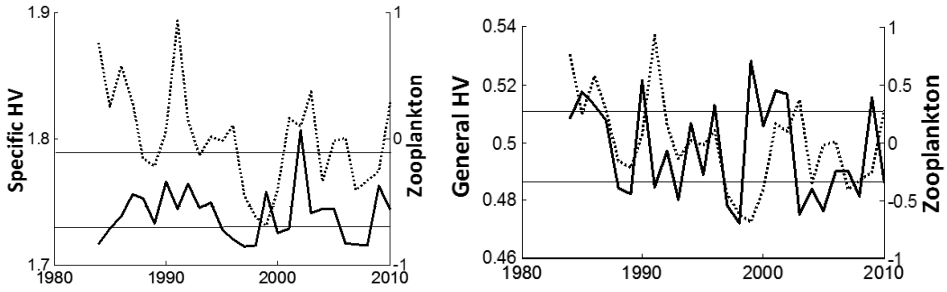
5.5.2 Analysis of the General and Specific Hidden Variables within Spatio-temporal Scales

To recall, in our HSDBN approach, we inspect the states of the two HVs- one *general* to model the general trophic dynamics and a *specific HV* to learn the spatial effect of zooplankton as it was missing in the model (due to limitation in spatial resolution) but is here validated against the measured zooplankton for the whole of North Sea. We show examples of learned hidden variables for only two of the areas.

The *general HV* for area 6 (figure not shown) captured some of the species biomass characteristics and it successfully managed to reflect on a temporal decline (trend identified, $p<.05$) in the series, coinciding with decline in the *LP* biomass for area 6 ($Z=0.4$, d.f.=26, $p=0.34$). The *specific HV* for area 6 is capturing the zooplankton dynamics with high similarity: ($Z=1.02$, d.f.=26, $p=0.31$, Fig.5.9a).

Results from both HDBN and HSDBN for area 5 showed that the zooplankton was actually modelled by the *general HV*, rather than the *specific HV*, the opposite of what we were aiming to find. Although the *general HV* managed to reflect on some zooplankton variations in time (Fig.5.9b), no temporal trends were found in either HV and no statistical similarity was found with either zooplankton or other species group biomass. Following the SDBN (failed to identify any statistical similarity or trend by

the *general HV*), it was the ARHMM that produced most accurate biomass predictions for 5 and the HVs generated by this model were characterised by a temporal decline (trend identified, $p<.05$), coinciding with decline in the P biomass for area 5 (Z=-0.71, d.f.=26, p=0.48).



(a) Area 6 (b) Area 5

Figure 5.9: The expected value of the specific HV (solid line) for area 6 (a) and general HV for area 5 (b) generated by HSDBN with the observed standardised biomass for the zooplankton (dotted line). The solid line indicates upper and lower 95% confidence intervals, obtained from bootstrap predictions' overall mean and standard deviation. Note the negative scale is due to standardisation.

To summarise, the *general HV* managed to capture changes in the variance of species groups biomass but that varied with the HSDBN predictive accuracy in different spatial areas. These results outline the importance of the spatially-specific driving factors on species dynamics and provide insights on spatial patterns in terms of ecological stability and resilience. The *specific HV* from our HSDBN managed to learn the zooplankton biomass variations in **all areas** (*general HV* for **area 5**), that provides us with more accuracy on the structure and functioning of the underlying ecological networks but also suggest the existence of strong regional differences for some areas.

5.5.3 Analysis of the Discovered Interactions between Species and their Environment

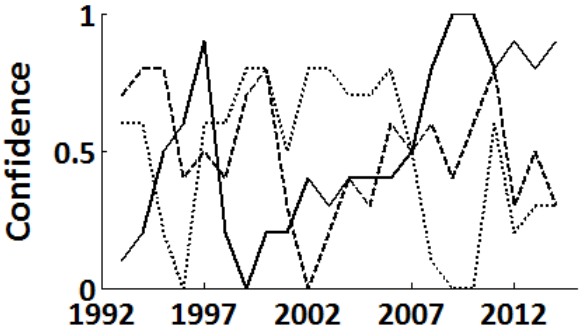
Discovered Interactions Within a Spatial Area We now investigate trophic associations and interactions with key stressors (*temperature* and *commercial catch*) identi-

fied *within each area*. Results from Table 5.5 showed that interactions of species groups with both anthropogenic and environmental factors, are of key importance when determining the local trophic dynamics and functional networks. In particular, there were high confident links identified between *catch* and *all groups of species* in areas **2**, **3** and **7**. In areas **1**, **4** and **6** interaction with *catch* was found only for some of the species groups. Conversely, **area 5** was the only area in which there were no confident links found with either one stressor. In this area, a single trophic interaction was found. Interestingly, there were high confident links identified between *temperature* and some of the *higher trophic level species groups* rather than *Net PP* on which *temperature* has a direct influence, suggesting the extensive linkages of some species following potential bottom-up effects.

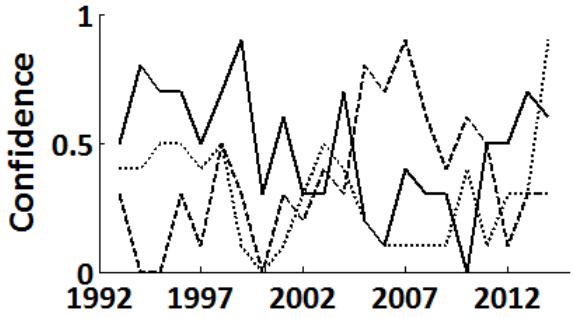
Table 5.5: Learned trophic associations and interactions with key stressors (*catch* and *temperature*) for each of the 7 spatial areas (the estimated mean confidence of each interaction, learned by the hill-climb for the time window: 1993-2010, is reported in brackets). The time window starts from 1993 due to the windowing required during the hill-climb.

Areas	Catch	Temperature	Trophic
1	P (0.26)	SP (0.46)	SP-LP (0.55)
2	P (0.51), SP (0.39), LP (0.41)	SP (0.54)	P-SP (0.70) Net PP-SP (0.44)
3	P (0.21), SP (0.27), LP (0.28)	None	P-SP (0.35), SP-LP (0.36), Net PP-P (0.22)
4	Net PP (0.42), LP (0.44)	P (0.22)	Net PP-LP (0.24), P-SP (0.22), P-LP (0.42), SP-LP (0.25)
5	None	None	P-LP (0.31)
6	Net PP (0.21), LP (0.23), SP (0.44)	SP (0.44)	P-SP (0.56)
7	Net PP (0.32), P (0.45), SP (0.46), LP (0.48)	SP (0.54)	Net PP-P (0.38), P-SP (0.64), P-LP (0.66)

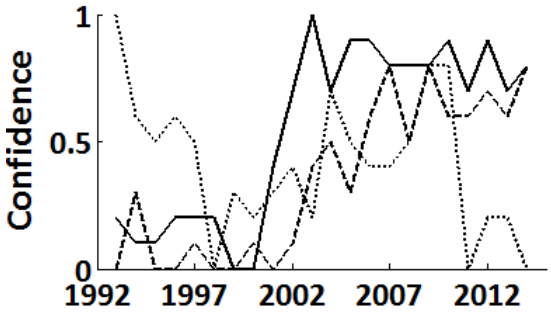
Discovered Interactions Between Spatial Areas We now look at the discovered interactions *across the whole of North Sea* for each type of functional relationship: *P-SP*, *P-LP* and *SP-LP* (Fig.A.5 shown in Appendix). The majority of confident links were identified for *P-SP*, followed by *P-LP*. Our results showed that links of high confidence were discovered between areas **1-3** (*P-SP* and *P-LP*), **2-5** (*P-SP* and *SP-LP*), areas located relatively spatially closer to each other. However at the same time we found high confident links between areas **1-6** (all functional relationships) and **1-7** (*P-SP* and *P-LP*), located at further distances. Next, we present the temporal variations of the identified relationships between some of the areas. Individual year effects were very strong for all relationships, however we notice the spatially-specific differentiation. The *P-SP* (Fig.5.10a) relationship was generally characterised with some temporal decline around late 1990s to early 2000 followed by increase over recent years but that was more evident between areas **1-3** and **1-7**, rather than **2-5**, which was generally of higher confidence ($\geq$0.5) in time, except for some decline occurring in more recent years. Similarly, *P-LP* (Fig.5.10b) was characterised by a declining trend again between areas **1-3**, whilst **2-3** was consistently increasing in time. The *SP-LP* (Fig.5.10c) relationship had a general trend of temporal increase from early 2000 which was evident between all areas.



(a) *P-SP*



(b) *P-LP*



(c) *SP-LP*

Figure 5.10: Estimated confidence by the hill-climb of each functional relationship occurring across the whole of North Sea: (a) *P-SP* between areas 1-3 (solid line), 1-7 (dashed) and 2-5 (dotted); (b) *P-LP* between areas 1-3 (solid), 2-3 (dashed) and 4-5 (dotted); (c) *SP-LP* between areas 1-6 (solid), 3-4 (dashed) and 4-6 (dotted). Note the time window starts from 1993 due to the windowing required during the hill-climb.

5.5.4 Summary for the North Sea

To summarise, results from our experiments showed that there was great spatial and temporal variation of the trophic dynamics in this ecosystem, however the high predictive accuracy of our hidden spatial dynamic BN (HSDBN) and its successful hidden variable characteristics in modelling the biomass changes and spatial unmeasured effect help us define this model as a *final* choice, out of the presented models, that we propose for further use by experts when looking at trophic dynamics in different ecosystems. Here, the HSDBN is the most thorough and comprehensive model choice which incorporates the effect of spatial autocorrelation but also the impact of fishing and

environmental stressors when modelling the spatial and temporal food web dynamics within the North Sea. The difference in predictive accuracy is to be expected as the Autoregressive Hidden Markov model (ARHMM) undertakes relatively simple modelling assumptions and fixed structure, that are unable to describe the species dynamics as accurately as our HSDBN. The general HSDBN improvement over the dynamic BN model (DBN) underlines the negative effect of information loss when only incorporating a single hidden variable and the similar performance of the hidden dynamic BN (HDBN) to HSDBN for some of the areas was due to structural similarity and increased modelling complexity but also due to less prominent spatial effects. The successful performance of the spatial dynamic BN (SDBN) in some areas highlights the uniform nature of the local trophic dynamics, because the importance of the driving factors is understated by the significance of the spatial relationship between neighbouring areas. Such results allowed us to identify patterns in the different spatial areas in terms of importance of the mechanisms involved in modelling the trophic dynamics. For example, we showed that the *general HV* for area **6** managed to reflect on the temporal decline, specifically for the *LP* group, whilst the *general HV* for **area 5** modelled the zooplankton dynamics (whilst *specific HV* modelled the *P* species group), highlighting the importance of fisheries for the ecological stability in some areas and that some species groups are more important for the functioning of the local ecological networks, compared to others. In addition, our results of predictive accuracy in terms of "what if" type descriptions further outline the heterogeneous nature of the species dynamics in area **6**, following the mutual influence of external disturbances and trophic interactions and their importance for the structure and stability of the local functional network. Interestingly, better biomass predictions for **area 5** were produced by the SDBN, suggesting the stronger spatial relationship that the neighbouring areas might have with the species dynamics in **area 5** but also its potential importance for habitat suitability, which could be further investigated by experts in terms of management schemes.

Our results show highly confident but *spatially* and *temporally* differentiated ecological networks that indicate spatial relationship of species and habitat with the particular mechanisms varying from facilitation through trophic interactions. Revealed trophic associations and interactions of species groups with their environment were consider-

ably better than random but we note that perfect reconstruction is unlikely due to the noisy data and complex ecological process involved (Faisal et al. (2010)). However, our findings complement traditional methods and have extended our knowledge into the complexity of North Sea dynamics and its ecological structure and stability.

In the next chapter, we modify our model to predict species dynamics in the North Sea further into the future in combination with fisheries and temperature scenarios to give strategic advice on short-term system response to pressure. The modelling work was extended to potentially benefit fisheries management in terms of assessment of choices on "optimum" levels of fisheries catch for different commercially important fish species.

Chapter 6

A Dynamic Bayesian Network Model to Predict Trends of Ecosystem Change in Response to Fisheries Catch, Temperature and Productivity Scenarios

6.1 Introduction

In this chapter, we use a modified version of a dynamic Bayesian network model to predict the response of different ecosystem components to change in anthropogenic and environmental factors within the North Sea. Through the development of fisheries catch, temperature and productivity scenarios, we explore the future of different fish and zooplankton species and examine what trends of fisheries exploitation and environmental change are potentially beneficial in terms of ecological stability and resilience. In the context of ecosystem-based approach to fisheries management, by using a multispecies ecosystem model that acknowledges fisheries are part of the environment, the species effects can be quantified across space and over time, under different fisheries exploita-

tion and climate scenarios.

The North Sea has been exploited for centuries by the surrounding countries and the state of its environment has been altered greatly by human activities (Jennings & Kaiser (1998)). However, in late 1990s the EU began a fleet reduction scheme and most recently, the European Unions Common Fisheries Policy (CFP) introduced significant changes to how fisheries are to be managed, including a landings obligation and management plans that take account of biological and technical interactions (Regulation (2013)). There is a growing recognition of the complexity of North Sea trophic dynamics and functional ecological networks, and together with fisheries legislative changes and climate warming trends taking place, there is an increasing demand for tools with which to explore alternative hypotheses about ecosystem function and response to change and address applied questions in the field of fisheries management (Mackinson & Daskalov (2007)). In the present work, we are specifically interested in the characteristics of Bayesian networks (BNs) to demonstrate the effects of various assumptions on the forward projections of variables of interest. Specifically, a dynamic BN model was applied to investigate the consequences of specific fisheries catch, temperature and productivity scenarios on different fish and zooplankton species. Through the developed scenarios, we explore the trends of ecosystem components and examine potential trade-offs and sensitivities. The approach we are using is a modified version of a dynamic hidden spatial BN model, which uses the functional network approach to predict species dynamics, accounting for species associations and interactions with external stressors and unmeasured hidden effects at spatial and temporal scale. Now, we extend this approach to model species dynamics further into the future by developing a set of scenarios, accounting for spatial heterogeneity and ecological complexity, which is important because species interactions can increase or reduce future changes at different scales, influencing the emergence of winners and losers (Barange et al. (2014)). Hence, we aim at predicting species year-to-year variations and understanding their dynamics, which is essential to give strategic advice on potential response of the system to pressure. We are interested in assessing choices on "optimum" levels of fisheries catch and potentially provide advice on short-term predictions to support fisheries management.

6.2 Description of the Spatio-temporal Data to Model Natural and Anthropogenic Scenarios

The analyses are based on the same dataset from Chapter 5 for the North Sea, that is why we only briefly mention some of the key points of the dataset.

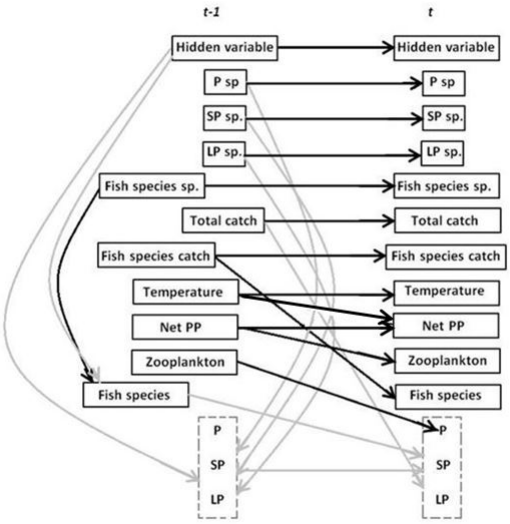
We only consider data collected within ICES (International Council for Exploration of the Seas) areas 1 to 7 (Fig.5.2 from Chapter 5, sub-section 5.2.2 *North Sea*) due to limited quality and consistency of the data on the remaining spatial areas. Individual fish species were aggregated by summing up the data into the relevant trophic group: pelagics (P), small piscivorous (SP) and large piscivorous and top predators (LP). However, here the main focus from the modelled fisheries catch scenarios will be on the individual species biomass: cod (*Gadus morhua*), haddock (*Melanogrammus aeglefinus*), herring (*Clupea harengus*), European plaice (*Pleuronectes platessa*), sole (*Solea solea*), saithe (*Pollachius virens*) and whiting (*Merlangius merlangus*). The fish survey data covers the period: 1983-2015. We also have available biomass data (overall for the North Sea) for different zooplankton species: *Calanus finmarchicus*, *Calanus helgolandicus* and small copepods: *Acartia*, *Temora*, *Para-pseudocalanus* for the time window 1983-2010. The fish Sea surface temperature (*temperature*), net primary production (*Net PP*) and catch data cover the time window: 1983-2010. The data was standardised (sample mean removed from each observation, which is then divided by the standard deviation) prior to conduction of the experiments but for the purpose of visualising the results, we reversed the standardisation of the modelled values.

6.3 Methods

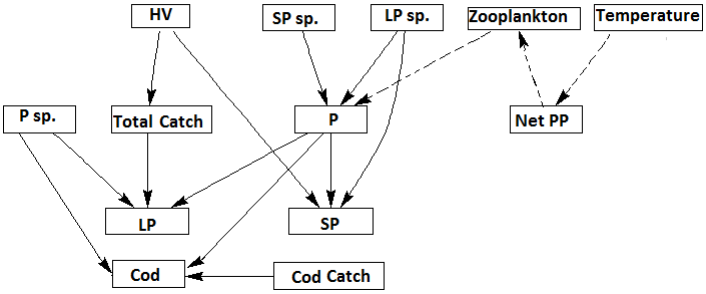
6.3.1 Description of the Model to Predict Species Trends in Response to Natural and Anthropogenic Scenarios

We used a modelling approach that integrates the functional network approach with a dynamic Bayesian network model to predict trends of different fish and zooplankton species from specific *fisheries catch*, *temperature* and *Net PP* scenarios. This approach is a modified version of the hidden spatial dynamic Bayesian network model developed

in Chapter 5 (we still refer to the model as Hidden Spatial Dynamic Bayesian Network model or HSDBN). This model is an extension of the model from the previous chapter in terms of predicting species data further into the future and modelling individual fish species dynamics under different effects from biotic and abiotic scenarios. We incorporate only one hidden variable (HV) and instead of a second HV, as originally in the previous chapter, we incorporate the observed zooplankton biomass for the North Sea. In addition to the three spatial nodes: *P sp.*, *SP sp.* and *LP sp.*, we add an additional spatial node as a parent node to the fish species variable, to account for the effect of spatial autocorrelation (refer back to Chapter 5, sub-section 5.3.2 *Spatial Autocorrelation*). The observed variables in the model include *total catch*, *fish species catch*, *temperature*, *Net PP*, total *zooplankton* biomass, *fish species biomass* and three aggregated trophic species groups: *P*, *SP* and *LP* and the equivalent spatial nodes from above. This totals 13 observed variables per area. The HSDBN structure varies but the general form is presented in Fig.6.1a, with example for one of the areas in Fig.6.1b. Hence, we can explore multiple species associations and model their future dynamics with interactions from external stressors and under specific scenario conditions. Using a recognised model structure, we can compare the modelled scenario outputs across spatial and temporal scales, accounting for the spatial heterogeneity and ecological complexity.



(a) HSDBN model structure



(b) Cod, Area 4

Figure 6.1: General structural form of the HSDBN model (a). Solid line represents fixed edges across areas. The spatial nodes (*P sp.*, *SP sp.*, *LP sp.*, *Fish species sp.*), *HV*, *catch* and *temperature* are individually linked to either *P*, *SP* or *LP* (represented by the dotted surrounding), depending on the spatial area (grey line). Connectivity between *P*, *SP* and *LP* and with the *fish species* also differs spatially. Network structure for area 4 (b) that models the dynamics of cod. The edges shown by a dotted line are defined by the expert.

6.3.2 Modelling Species Trends in Response to Fisheries Catch, Temperature and Productivity Scenarios

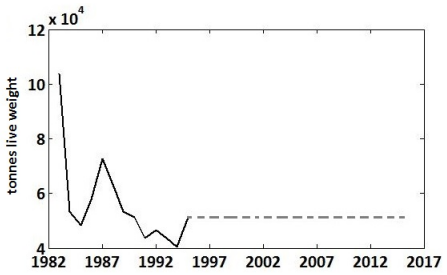
The experiments involved prediction of survey data under scenarios of fisheries catch, temperature and Net PP. Given the probability distribution over $\mathbf{Y}[t]$ where $\mathbf{Y}$=$Y_1...Y_n$ are the n variables observed along time t, to predict the biomass of each species and/or trophic group, we inferred the biomass at time *t+1...t+5* by using the observed evidence (or available data) from *t-1* and *t*. The choice of 2020 as the horizon for this study was chosen to limit uncertainty but most importantly, we wanted to provide some practical advice in terms of management objectives in line with the need to restore fish stocks. We used an exact inference method: the junction tree algorithm (Murphy (1998)). The hidden variable is specified as a discrete node which is parameterised using the Expectation Maximization algorithm in a maximum likelihood sense and assumes a discrete distribution. Similarly to Chapter 5, a non-parametric bootstrap (Friedman et al. (1999)) was applied 250 times for each modelling scenario.

First, we predict the survey data for each area using historical observations, we refer to this model output as *Historical*. Then, we use different fixed year levels from each individual fish species catch data to design our fisheries catch scenarios. We aim at having scenarios at varying levels of fisheries catch: *low*, *medium* and *high* (these to be referred from now on as scenarios of *L.FC.*, *M.FC.* and *H.FC.* respectively). We choose from the fisheries catch data three years equivalent to these levels and keep each level fixed from the chosen "scenario" year until the year 2015. We keep the other measured variables unchanged. For example, in order to model the dynamics of cod in area 4 in response to change in fisheries catch, we chose from the cod catch data the year 1995 to represent the year from which the scenario of *M.FC.* starts. Fig.6.2a illustrates the data input assuming this scenario and the generated output. Note, that the data input for testing the model prior to the scenario year includes all of the observed variables (and one HV) up to 1995 and after 1995 to 2015, the input is only the fixed value of the total fisheries catch and cod catch ($5x10^4$ tonnes live weight) (Fig.6.2b). In this way, we rule out the simple idea that observed values after the "scenario" year are causing the results to stabilize.

We perform this for each individual fish species and across each area, according to the originally published model structure. For example, in area 4 *catch* is a direct parent to *LP*, so in this area, we would investigate fisheries catch scenarios for individual *LP* fish species such as cod (Fig.6.1b). In area 1, *catch* is a direct parent to *P*, so in area 1 we examine the influence of fisheries catch on pelagic species only, such as herring. In this way, we can keep the historically driven interactions between variables and examine their modelled long-term biomass trends under potential changes in stressors such as fisheries catch. At the same time, we predict the trophic biomass (*P*, *SP* and *LP*) for each area and from the same fisheries catch scenarios that were applied to the individual fish species. Hence, we can examine how different ecosystem components respond to varying levels of fisheries catch, accounting for the heterogeneous nature of the modelled variables and driving factors within each area and their changes over time.

Input	Input during Model Testing		Output
	Prior to **1995**	After **1995**	
1. []	1. []	1. []	1. *HV* (1989-2020)
2. *P sp.* (1983-2015)	2. *P sp.* (1983-**1995**)	2. []	2. *P sp.* (1989-2020)
3. *SP sp.* (1983-2015)	3. *SP sp.* (1983-**1995**)	3. []	3. *SP sp.* (1989-2020)
4. *LP sp.* (1983-2015)	4. *LP sp.* (1983-**1995**)	4. []	4. *LP sp* (1989-2020)
5. *Cod sp.* (1983-2015)	5. *Cod sp.* (1983-**1995**)	5.[]	5. *Cod sp.* (1989-2020)
6. *Total Scenario Catch* (1983-2015)	6. *Total Scenario Catch* (1983-**1995**)	6. *Total Scenario Catch* (**1995** -2015)	6. *Total Scenario Catch* (1989-2020)
7. *Cod Scenario Catch* (1983-2015)	7. *Cod Scenario Catch* (1983-**1995**)	7. *Cod Scenario Catch* (**1995** -2015)	7. *Cod Scenario Catch* (1989-2020)
8. *T°* (1983-2010)	8. *T°* (1983-**1995**)	8. []	8. *T°* (1989-2020)
9. *Net PP* (1983-2010)	9. *Net PP* (1983-**1995**)	9. []	9. *Net PP* (1989-2020)
10. *Zooplankton* (1983-2010)	10. *Z* (1983-**1995**)	10. []	10. *Z* (1989-2020)
11. *P* (1983-2015)	11. *P* (1983-**1995**)	11. []	11. *P* (1989-2020)
12. *SP* (1983-2015)	12. *SP* (1983-**1995**)	12. []	12. *SP* (1989-2020)
13. *LP* (1983-2015)	13. *LP* (1983-**1995**)	13. []	13. *LP* (1989-2020)
14. *Cod* (1983-2015)	14. *Cod* (1983-**1995**)	14. []	14. *Cod* (1989-2020)

(a) Data input and output for Medium Fisheries Catch scenario for cod, area 4



(b) Medium Fisheries Catch scenario for cod, area 4

Figure 6.2: An example matrix from a *Medium Fisheries Catch* scenario model with initial input used in model definition, the input during model testing and the generated output (a). The time window for each variable is shown in brackets. Note, the time window for the output starts from 1989. "[]" represents variables for which no evidence is introduced and which are predicted. "Z" stands for zooplankton. The observed cod catch prior to the scenario year of 1995 (solid line) and fixed catch level for the Medium Fisheries Catch scenario (dashed line) is shown in (b).

We generate a 10% increase temperature scenario (*T.I.*) and Net PP scenarios: 30% increase and 30% decline (referred from now on as: *Net.I.* and *Net.D*) to understand the effects of temperature on primary production and its potential knock-on effects on different zooplankton species and trophic biomass higher up the food chain. We did consider a scenario of temperature decline but we only present the results following a potential increase in temperature. We used the year 1990 as a "divergent year", from which we would start the scenario conditions by manipulating the temperature or *Net PP* data to either increase or decline but keeping the rest of the observed data unchanged, e.g. if the average sea surface temperature for 1990 is 9°C, then for 1991 it would 9.9°C. For these two types of scenarios, the number of observed variables in the experimental set-up is 12 (*total catch*, *temperature*, *Net PP*, *Calanus finmarchicus*, *Calanus helgolandicus*, small copepods, *P sp.*, *SP sp.*, *LP sp.*, *P*, *SP* and *LP*).

6.4 Results

In the following, we describe the outputs from the modelled fisheries catch, temperature and Net PP scenarios by examining future trends of individual fish and zooplankton species at spatial and temporal scales. We explain the results from the scenarios in terms of how much worse (or better) the ecosystem components are predicted to become. Our results demonstrate some variability in the future trends of different species, which we explain through the use of "what if" type descriptions of the other variables in the model structure.

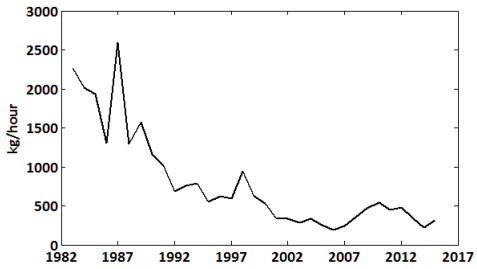
6.4.1 Fisheries Catch Scenarios

Cod

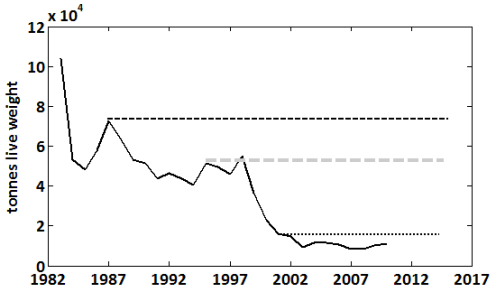
First, looking at the *Historical* output (Fig.6.3c), the model managed to capture the cod variations throughout time and predicted some increase in recent years which were then followed by some decline. These results highlight that there are some signs of recovery in recent future years possibly due to strict management regulations placed since the Millennium (Horwood et al. (2006)), which if continued, will hopefully give the stock a chance to rebuild completely in this area where the cod was formerly abundant (Engelhard et al. (2014)). However, at the same time, the cod interactions with other species groups and its spatial values (biomass in neighbouring areas) that we describe below also need to be accounted for in terms of short-term management.

Second, looking at the scenario outputs, as we would expect, the scenario of High Fisheries Catch (*H.FC.*) (Fig.6.3d) resulted in the lowest modelled cod survey data in areas 4 and 6 (thus, addressing in detail only area 4). We notice a sudden decline in early 1990s (as a result from the high scenario catch level), but then the modelled values were characterised by some fluctuating trend, that was higher than the observed data. This does not mean that if cod could continue to be fished at the highest recorded level the stock would still be ok but more likely when the cod survey (and spatial, Fig.6.3a) values are low and catch is high (Fig.6.3b), another species might increase and a year later that would cause the cod to increase. For example, in this area, cod is influenced by the dynamics of species group P (Fig.6.1b), which were predicted to be relatively stable

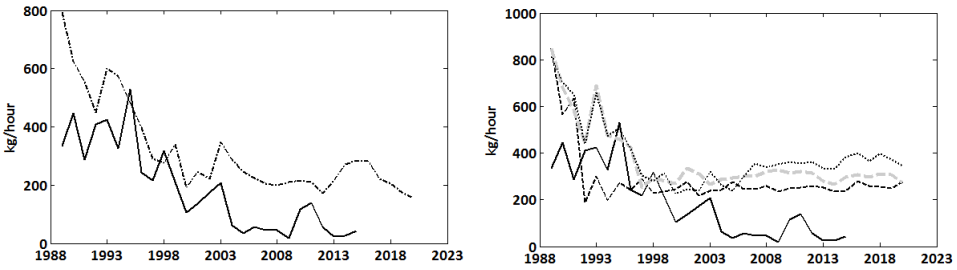
with an increasing trend in the near future, partly explaining the modelled cod results here. Under the scenario of Medium Fisheries Catch (*M.FC.*), the modelled survey data seemed to be genuinely stable throughout time that was higher than the scenario of *H.FC.* However, we notice that these two scenarios seem to converge in most recent years, highlighting the similarity in species response to contrasting levels of fisheries catch, thus still having the need to identify a potential "optimum" level of fisheries catch. The scenario of Low Fisheries Catch (*L.FC.*) resulted in the highest modelled cod survey data, highlighting the importance of fisheries catch on this species dynamics and identifying a potential "optimum" level of fisheries exploitation comparing to the medium and high levels from above.



(a) Cod spatial data, area 4



(b) Cod catch, area 4



(c) Cod survey data and *Historical* output, area 4

(d) Cod survey data and modelled scenario cod, area 4

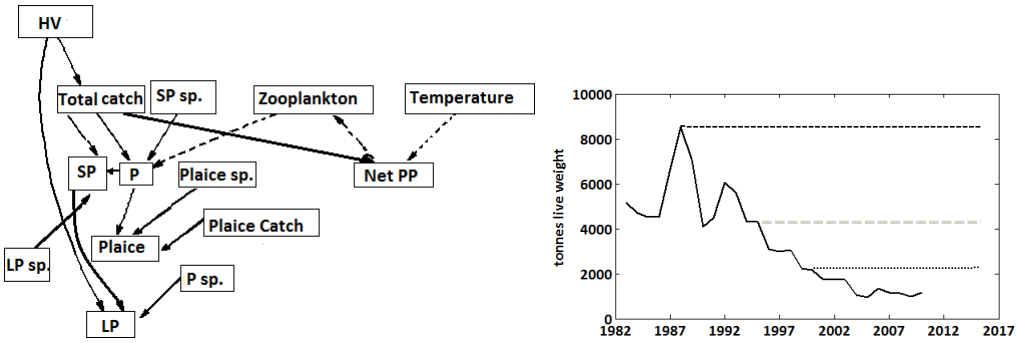
Figure 6.3: Recorded spatial cod data is shown in (a). The observed cod catch (live weight in tonnes) with the three fixed year levels of fisheries catch scenarios for the time window 1983-2015 is shown in (b). Recorded survey cod data (solid line) with the generated output by the *Historical* model (dotted line) for the time window 1989-2020 for area 4 (c). Recorded survey cod (solid line) with the modelled cod is shown in (d) under fisheries catch scenarios of high (black dashed line), medium (grey dashed line) and low (black dotted line) levels for the time window 1989-2020.

Plaice

Results from the *Historical* model (Fig.6.4c) projected relatively high trend of increase in the near future, which is suggested to be due to shifting distribution of plaice from southern to more northern areas (Engelhard et al. (2011)). We notice that the *Historical* model did not manage to predict the scale of the two recent peaks in the survey data, which could reflect a combination of year to year variation of recruitment into the area and variability in the survey data. Even, the genuine declining trend in the recorded plaice catch (Fig.6.4b) could further support the modelled increase, the future trend of the plaice is better explained by climate and spatial values that need to be taken into consideration when designing management measures that are more "adaptive" in future years (Engelhard et al. (2011)).

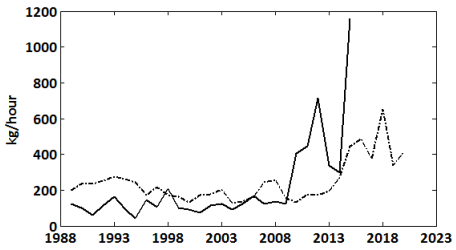
Now, let us look at the outputs from the fisheries scenarios: for the modelled plaice survey data in area 3, the scenario of *L.FC.* did result in the highest projected trend of plaice (Fig.6.4d). Interestingly, the two contrasting levels of scenarios of *H.FC.* (over

8000 tonnes) and *M.FC.* (just over 4000 tonnes) projected similar trends of plaice values in area 3. Assuming medium to high level of fisheries exploitation (and according to the model structure, Fig.6.4a), the projected plaice might be relatively stable with higher values than the recorded survey data when the dynamics of P species are relatively low (e.g. the influence of herring predation on plaice recruitment has been discussed before: Daan et al. (1985)) and plaice spatial values are high. These results highlight the variability in terms of "optimum" levels of fisheries exploitation (due to species interactions) and consequently species-specific response to ecosystem change.

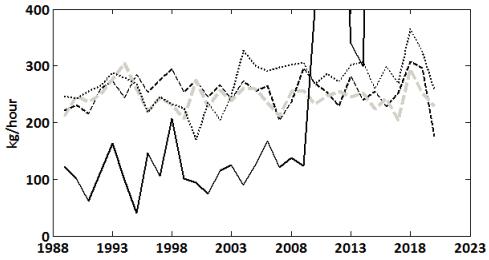


(a) Area 3

(b) Plaice catch, area 3



(c) Plaice survey data and *Historical* output, area 3



(d) Plaice survey data and modelled plaice, area 3

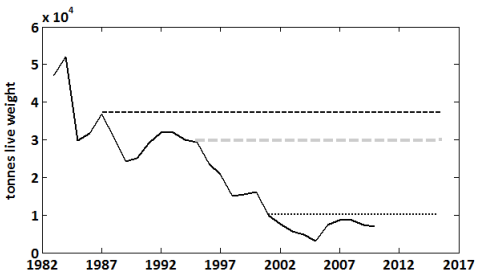
Figure 6.4: The model structure for area 3 that models the plaice dynamics is shown in (a). The dotted edges are defined by the expert. The observed plaice catch (live weight in tonnes) with the three fixed year levels of fisheries catch scenarios for the time window 1983-2015 is shown in (b). Recorded survey plaice data (solid line) with the generated output by the *Historical* model (dotted line) for the time window 1989-2020 is shown in (c). Note, the y-axis was cut off only for visual purposes. Recorded survey (solid line) with the modelled plaice data under fisheries catch scenarios of high (black dashed line), medium (grey dashed line) and low (black dotted line) levels for the time window 1989-2020 is shown in (d).

Whiting

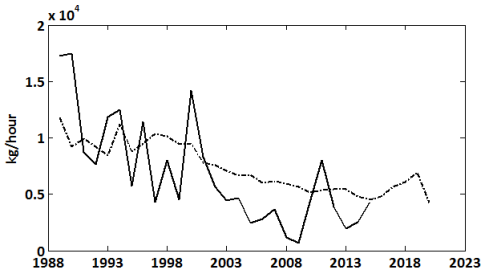
The *Historical* model (Fig.6.5b) managed to reflect on the declining trend of whiting throughout time and predicted some rising trends in the near future which were then followed by some decline. There is some obscurity regarding the status of whiting stock, which has caused further implications in terms of management but it has been discussed that whiting does not exhibit strong responses to climate change (Kerby et al. (2013)). Thus, controlling for the level of fisheries exploitation and considering trophic interactions and spatial values as we show below are of high significance in terms of short-term management.

We found the opposite of what we were expecting from the fisheries catch scenarios for whiting in area 3: a scenario of *L.FC.* produced whiting predictions, that were characterised with the lowest trend throughout time (Fig.6.5c). This suggests that potential trade-offs between some competitive species such as whiting and higher trophic level

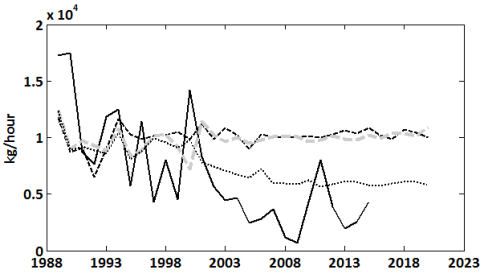
species (e.g. cod) are expected which have been discussed before (Mackinson & Daskalov (2007)). The surrounding predictions of the whiting spatial node were also characterised by a declining trend, which in combination with the medium to high catch from *M.FC.* and *H.FC.* and relatively low values of the P species group (the network from Fig.6.4a is the same but we would replace plaice with whiting) might allow for another species to increase (e.g. larger predator), which in turn would cause the projected whiting values here. We note that the predicted trends from the *M.FC.* and *H.FC.* scenarios were relatively similar which might be due to similarity in the level of fisheries catch but also due to the fact that trophic interactions are potentially more important for controlling the whiting dynamics compared to fisheries, as discussed in Trifonova et al. (2015) for this area. Interestingly, the hidden variable (HV) captured some of the expected "correct" characteristics: the scenario of *L.FC.* projected a strongly increasing trend of the HV, that was much higher than the HV from the *Historical* model. The HV is linked to the LP species group (which includes cod), so it is capturing changes in the variance of their survey data, potentially due to species associations and interactions (LP is influenced by SP and P *sp.*) and consequent trade-offs between species, that were not easily detected by the model predictions alone. Our results of modelling whiting in response to different fisheries levels and consequent rising trophic interactions and sensitivities that were captured by the HV, suggest that for effective management, reorganization of the fishing strategies in the mixed-fisheries context will be required to ensure that the right species are targeted and harvested sustainably. Thus, still having the need to identify a potential "optimum" level of fisheries catch to account for the effect of trade-offs between species.



(a) Whiting catch, area 3



(b) Whiting survey data and *Historical* output, area 3



(c) Whiting survey data and modelled whiting, area 3

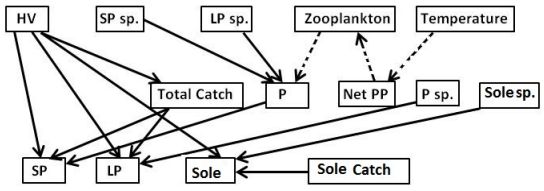
Figure 6.5: The observed whiting catch (live weight in tonnes) with the three fixed year levels of fisheries catch scenarios for the time window 1983-2015 is shown in (a). Recorded survey whiting data (solid line) with the generated output by the *Historical* model (dotted line) for the time window 1989-2020 is shown in (b). Recorded survey (solid line) with the modelled whiting data under fisheries catch scenarios of high (black dashed line), medium (grey dashed line) and low (black dotted line) levels for the time window 1989-2020 is shown in (c).

Sole

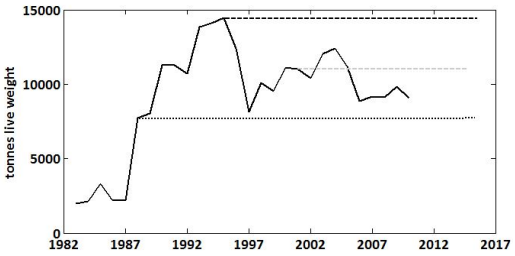
Our results from the *Historical* model captured the strong fluctuating trend of the sole survey data (Fig.6.6c), which is most likely due to the strong influence from temperature and environmental variability on this species (Engelhard et al. (2011)), and projected some stability in near future years. In terms of management implications, the influence from fisheries catch and climate change needs to be taken into consideration and similarly to other species, reorganization of the fishing strategies will be required to ensure

that this species is targeted sustainably in the correct spatial areas, following changes in distribution as explained below.

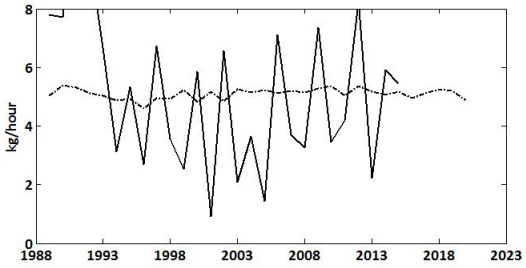
The scenario of *L.FC.* for area 6 produced the trend of highest magnitude throughout time (Fig.6.6d), although the trend was characterised with some fluctuation throughout time. That could be an indication of the relatively low catch levels in combination with the effect from environmental conditions on the recruitment and survival of sole (Perry et al. (2005)). The scenario of *H.FC.* did result in the lowest modelled trend throughout time, although there was some similarity with the predictions modelled by the scenario of *M.FC.*, which could be due to the more or less similarity in the catch levels between those two scenarios. However, we can see that the effect from fisheries catch on this species is potentially strong as it has been suggested by other studies (Engelhard et al. (2011)). Interestingly, it has been found for this species to be characterised by a distributional shift from northern to southern areas (opposite to the distributional shift of plaice), thus resulting in a potential increase of this species in areas such as area 6. This could further explain the similarity in the modelled predictions between the scenarios of *M.FC.*and *H.FC.*, thus the distributional shift is potentially masking the full effect from fisheries catch.



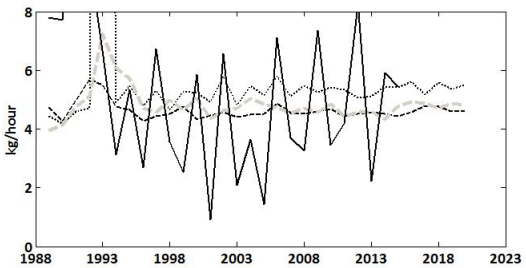
(a) Area 6



(b) Sole catch, area 6



(c) Sole survey data and *Historical* output, area 6



(d) Sole survey data and modelled sole, area 6

Figure 6.6: The network structure for area 6 that models the sole dynamics is shown in (a). The dotted edges are defined by the expert. The observed sole catch (live weight in tonnes) with the three fixed year levels of fisheries catch scenarios for the time window 1983-2015 is shown in (b). Recorded survey sole data (solid line) with the generated output by the *Historical* model (dotted line) for the time window 1989-2020 for area 6 is shown in (c). Recorded survey (solid line) with the modelled sole data under fisheries catch scenarios of high (black dashed line), medium (grey dashed line) and low (black dotted line) levels for the time window 1989-2020 is shown in (d). Note, the y-axis in (c) and (d) was cut off only for visualisation purposes.

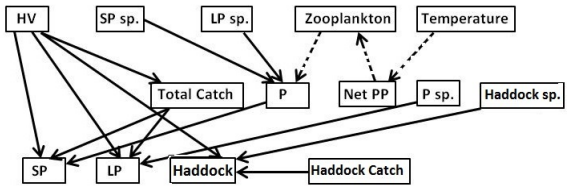
Haddock

Our *Historical* model projected relatively low trend of haddock data in the future (Fig.6.7c) which could be explained by the fact that studies have reported that haddock tends to move further north due to changes in climate (Mackinson & Daskalov (2007)), suggesting that keeping the catch to a potential "optimum" level is not the only key factor to the recovery of larger predators in this area. These results should be taken into consideration in terms of management as strategies should be re-directed towards other species in this area.

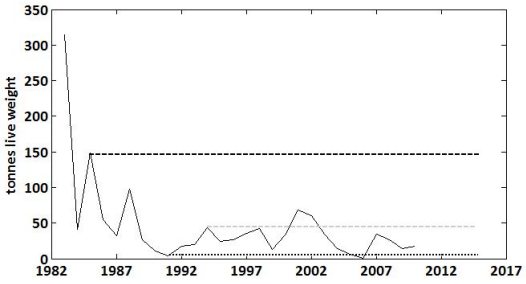
Interestingly, the opposite of what we were originally expecting was found for the haddock survey data from the different fisheries catch scenarios. The scenarios of *H.FC.* and *L.FC.* produced the highest and lowest trend of haddock survey data respectively (Fig.6.7d). We notice the strong individual year effects for this species dynamics, for

example in early 2000, that were evident throughout all catch scenarios. However, if we look at how the other ecosystem components are predicted to change in this area, we can see clearer results on the modelled future trends from the catch scenarios, which could potentially explain the unusual haddock predictions we found here.

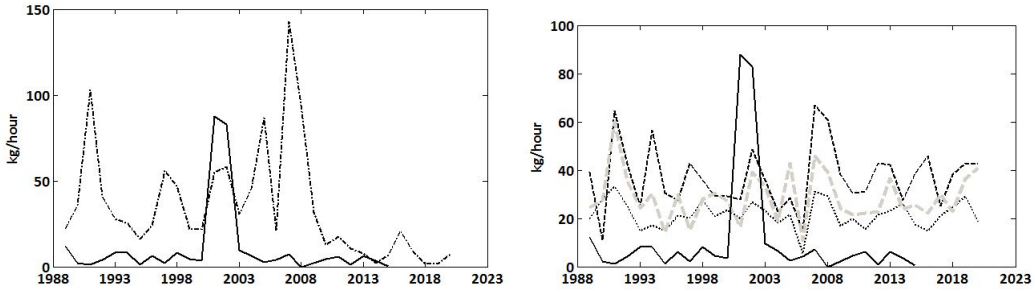
For example, the *SP* and *LP* species group data was predicted to be of the highest magnitude once the catch was kept to a minimum from the scenario of *L.FC.* (potentially due to decline in the catches of plaice and cod and subsequent increase in the biomass of these species). Interestingly, the scenario of *M.FC.* projected a clear increasing trend for only the hidden variable (HV), suggesting that the HV is potentially capturing changes in the variance of species biomass following species associations and interactions with *catch* (the HV is influenced by *catch*, *SP*, *LP* and *haddock*, Fig.6.7a). In addition, the spatial relationship between neighbouring areas could also explain the modelled haddock data: the surrounding spatial haddock values were characterised by a strong declining trend throughout time that was also projected to continue in the near future.



(a) Area 6



(b) Haddock catch, area 6



(c) Haddock survey data and *Historical* output, area 6

(d) Haddock survey data and modelled haddock, area 6

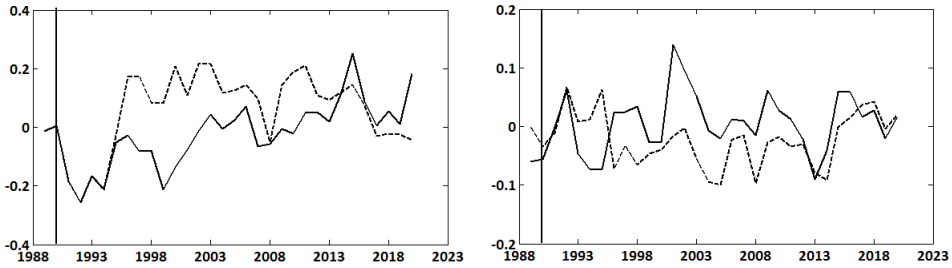
Figure 6.7: The network structure for area 6 that models the haddock dynamics is shown in (a). The dotted edges are defined by the expert. The observed haddock catch (live weight in tonnes) with the three fixed year levels of fisheries catch scenarios for the time window 1983-2015 is shown in (b). Recorded survey haddock data (solid line) with the generated output by the *Historical* model (dotted line) for the time window 1989-2020 for area 6 is shown in (c). Recorded survey (solid line) with the modelled haddock data under fisheries catch scenarios of high (black dashed line), medium (grey dashed line) and low (black dotted line) levels for the time window 1989-2020 is shown in (d).

6.4.2 Temperature and Net PP Scenarios

We are now looking at the potential influence of temperature on the future projections of productivity and consequently how the productivity will influence the future trends of different zooplankton species. We have chosen to present results only for areas 1, 3 and 6 due to the contrasting nature of the physical and bio-chemical characteristics of these areas.

For area 3 (and area 1), the scenario of *T.I.* resulted in an increasing trend of *Net PP* throughout time (Fig.6.8a) that was also higher than the *Historical* model. However, the *T.I.* scenario projected some *Net PP* decline in recent future years that was characterised by a converging trend with the projections of the *Historical* model, possibly indicating a drop in productivity, as it has been suggested by other studies (Blanchard et al. (2012)). Conversely, for area 6, the scenario of *T.I.* projected a trend

of lower *Net PP* values (Fig.6.8b) than the *Historical* model, potentially due to larger temperatures changes in southern areas. Similarly, there was a drop in productivity projected from 2017 onwards, that could be related to the overall future productivity conditions expected in the North Sea.



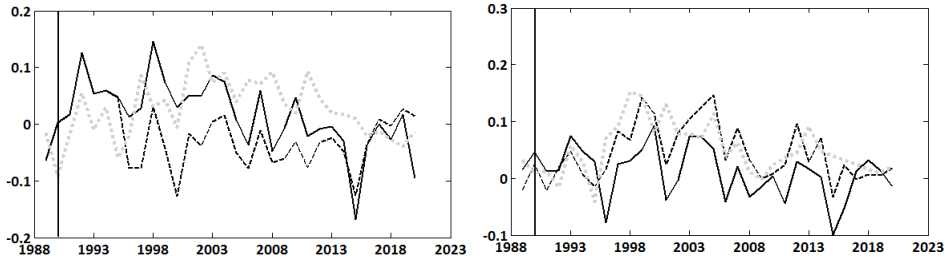
(a) *Net PP*, Area 3

(b) *Net PP*, Area 6

Figure 6.8: Modelled *Net PP* for areas 3 (a) and 6 (b), generated by the *Historical* model (solid line) and scenario of *T.I.* (black dashed line) for the time window: 1989-2020. Negative scale is due to standardisation. The vertical solid line indicates the year of divergence (1990), when we initiate the scenario conditions.

In area 1, it was difficult to interpret the future trend for *C.finmarchicus*, its predictions were slightly lower following a scenario of *T.I.*, whilst no significant change in the species trend was found from either of the *Net PP* scenarios, highlighting the importance of temperature. In area 3 (and area 6), results for this zooplankton species were clearer: a scenario of *T.I.* projected a lower trend throughout time for this species, comparing to the *Historical* model (Fig.6.9a). The opposite was found for *C.helgolandicus*: in all areas, the scenario of *T.I.* produced a trend of higher magnitude in time for this species (Fig.6.9b). In areas 3 and 6, it was also the scenario of *Net.D.* that let to higher values of both zooplankton species, as a consequence of temperature influence on productivity. We notice some future drop in the projected values of the *C.finmarchicus* species. In all three areas, similar dynamics were modelled for the small copepods, as were for the *C.finmarchicus*, a decline in the copepods' values was projected under a scenario of *T.I.* Here, we further confirm how different zooplankton species would potentially respond to change in temperature, highlighting that potential trade-offs will also emerge between lower trophic level species that might potentially cause further complications to species

higher up the food chain.

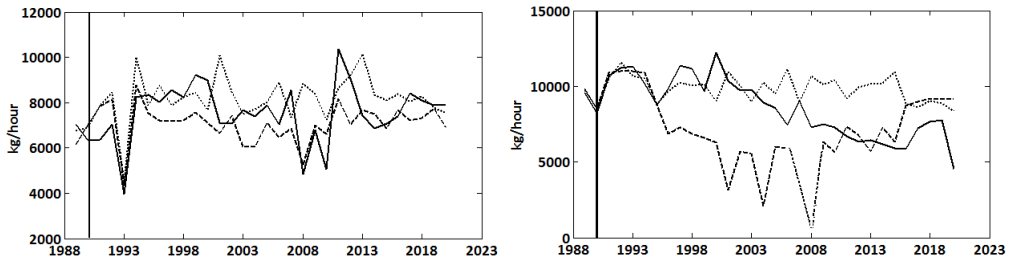


(a) *C.finmarchicus*, area 3 (b) *C.helgolandicus*, area 6

Figure 6.9: Modelled data by the *Historical* model (solid line) for *C.finmarchicus* for area 3 (a) and *C.helgolandicus* for area 6 (b) under scenarios of *T.I.* (black dashed line) and *Net.D.* (grey dotted line) for the time window: 1989-2020. Negative scale is due to standardisation. The vertical solid line indicates the year of divergence (1990), when we initiate the scenario conditions.

We were able to detect a knock-on effect on the future dynamics of the P species group survey data, following changes in temperature and productivity: in area 6, there was a clear increase in the future trend of the P species group following scenarios of *T.I.* and *Net.D.* (potentially beneficial to *C.helgolandicus* from above), outlining the temperature importance on species dynamics in some areas. In area 3 (and area 1), it was difficult to interpret the effect from temperature but the scenario of *Net PP* decline would consequently result to an increase in the trends of the herring and P species group survey data (Fig.6.10a,b). Here, the influence of productivity is likely to mask the effects from fisheries, as according to Fig.6.4a, the *catch* has a direct influence on *Net PP*, or cause a mixture of responses due to multiple causal mechanisms and stressors on the ecosystem (Halpern et al. (2008)).

To summarise, we found the modelled future zooplankton trends to be species-specific but there seems to be consistency in terms of their response to temperature change across the different areas, whilst more variability was found relating to productivity changes. In addition, we were able to confirm the potential influence from productivity and to some extent temperature (depending on the area) changes to species, higher up the food chain.



(a) Herring, area 3

(b) *P*, area 3

Figure 6.10: Modelled data by the *Historical* model (solid line) for herring (a) and *P* species group (b) for area 3 under scenarios from *Net PP* increase (black dashed line) and decline (grey dotted line) for the time window: 1989-2020. The vertical solid line indicates the year of divergence (1990), when we initiate the scenario conditions.

6.5 Summary

In this work, we explored the trends of ecosystem change in response to anthropogenic and environmental scenarios by modifying a dynamic data-driven functional network model, accounting for spatial heterogeneity and unmeasured spatial effects. It is important to note that we did not attempt to indicate levels of plausibility between these scenarios but rather explore the predictive results of species response to fisheries and environmental change. Our results highlighted that reducing fisheries catch will not necessarily lead to recovery of all commercially important fish species because fish consume one another, thus the total catch of one species will consequently affect that of others through knock-on effects in the food web. Overall, we found some spatial variability in terms of species response to different fisheries catch and productivity scenarios, highlighting the influence from factors such as trophic associations, spatial connectivity between areas and species interactions with their environment, that could potentially contribute towards the better understanding of ecological stability and resilience in a changing environment. However, at the same time, we found some universal species trends to changes in *catch* and *temperature*, that could provide some strategic advice on potential response of the system to such pressures.

In the scenarios modelled here, some trade-offs between species emerged in terms of

how they would respond to different levels of fisheries catch. Specifically, the potential recovery that we found for cod in recent future years could explain the modelled results for whiting because cod feeds on juvenile whiting (Mackinson & Daskalov (2007)). Similar results were found by Lynam & Mackinson (2015) and Lewy & Vinther (2004), suggesting a more dominant role of the cod in the food web after recovering from exploitation.

Although, we analyse the different scenarios in respect to the species of interest in the relevant area, we do acknowledge that one area's dynamics likely affect another by introducing the spatial nodes into the model structure. In this way, we also increase the confidence in the robustness of the approach and contribute to increased knowledge of model behaviour. One main issue encountered is the uncertainty in future trends, which is obviously inherent to any model linking external factors to species interactions. These linkages are of major importance for mixed-fisheries management (Ulrich et al. (2011)). However, the fact that we were able to recover genuine trends of species dynamics throughout space and time from the previous chapter and that we were able to identify similarity in our results here with other modelled species predictions (Lynam & Mackinson (2015); Lewy & Vinther (2004) and Vinther et al. (2004)) contributes to strengthening the confidence that our approach can provide some strategic advice on modelling species response to change.

Finally, in the modelled scenarios here, we found that some species appear more robust to changes in fisheries exploitation, compared to others, however changes in temperature and productivity might be more important in terms of the species long-term sustainability. Increase in temperature leads to an increase in lower trophic level species and consequently their predators, which we found true for some areas, whilst in others, the effect of temperature on fish was less evident due to interactions with productivity, which could be acting more strongly than the effect of fishing (Blanchard et al. (2010)). Such results confirm that species response to any future changes in temperature will be determined by their spatial habitat because temperature variations consequently lead to spatial variability in productivity, potentially causing further forcing on higher level trophic species and mixture of responses at spatial scales.

Our results allow dynamic assessment of choices, which should be able to provide strate-

gic advice on potential system response to pressure. In terms of management objectives and expectations, we support the idea that for a given area, reorganisation of the fishing fleet and management strategies will be required to ensure that the right species are targeted and harvested sustainably (Simpson et al. (2011)).

In the next chapter, we will look at the issue of generalising between different marine geographic regions (North Sea and Gulf of St Lawrence) in order to confirm key functional interactions. We are interested in examining the concept of functional equivalence across different datasets to determine level of similarity in the way ecosystems organise themselves functionally and respond to pressures.

Chapter 7

Models for Generalisation across Different Marine Ecosystems

7.1 Introduction

In this chapter, we generalise dynamic interactions between trophic groups of species from different marine ecosystems to confirm key functional relationships. Examining the nature of functional relationships that are ubiquitous over systems is important in order to understand the controlling mechanisms and study impacts of forces such as fishing and climate change. Different species may have similar functional roles within a system depending on the region. If we can model the type of interaction rather than the species itself, we can integrate processes into predictive models and study ecological controls and patterns in marine ecosystems that is essential for ecosystem-based approaches to fisheries management. Thus, being able to provide a general framework to model ecosystem dynamics is of growing importance for the protection of natural biodiversity and human resources. We predict species biomass by accounting for spatial autocorrelation and unmeasured effects to generalize the relative influence of forces such as fishing and temperature change on species dynamics. Although, different marine ecosystems might experience varying levels of fisheries catch and climate change, we believe that we can find similarity in the genuine species response to natural and anthropogenic pressures. We are specifically interested in modelling species functional

networks and trophic dynamics with a focus on the concept of functional equivalence across different marine ecosystems, thus allowing more robust models to be built and predictions to be made about future biomass.

In this chapter, we learn dynamic Bayesian networks for 7 geographically different sub-regions within the Gulf of St. Lawrence with varying trophic structure and evaluate these Bayesian models on 7 different areas within the North Sea. In addition, we incorporate hidden variables in the models to examine unmeasured effects in order to outline the importance of different mechanisms on the dynamics of exploited ecosystems in different regions. In particular, a focus is made on the development of models that identify functionally equivalent trophic interactions (as opposed to species like we did in Chapter 1) in different fish communities with the aim of predicting the influence from different pressures (as opposed to functional collapse like we did in Chapter 4), further highlighting key mechanisms and interactions that are significant to the population dynamics.

7.2 Description of Biomass Data to Model Generalisation across Two Geographic Regions

The analyses are based on the same dataset from Chapter 5 for the Gulf of St Lawrence and North Sea, that is why we only briefly mention some of the key points of the dataset. We analysed data from the northern Gulf of St. Lawrence (48.00°N, 61.50°W, Fig.7.1) groundfish and shrimp summer survey (1990-2010). K-means (Hartigan & Wong (1979)) was applied to limit the number of variables and cluster the number of sampling stations (originally over 200 sampling stations per year) on the mean latitude and longitude, resulting in 20 spatial clusters (or sub-regions). However, from these 20 clusters, seven were chosen based on their predictability (the 7 clusters that were most easily predictable from Chapter 5: clusters 1, 2, 4, 11, 13, 17 and 18, Fig.7.1) so that we use equivalent number of spatial areas from both regions.

For the North Sea, we only consider data collected within ICES (International Council for Exploration of the Seas) areas 1 to 7 (Fig.5.2 from Chapter 5, sub-section 5.2.2 *North Sea*) due to limited quality and consistency of the data on the remaining spatial

areas.

For both regions, fish species were aggregated into the relevant trophic group by summing up the biomass (P- pelagics, SP- small piscivorous fish and LP- large piscivorous and top predators) (FishBase was used as a guidance point). The nature of individual species summed into the trophic guilds varied between the spatial areas but this was not of importance since they were always aggregated into the correct trophic group. This was performed for each of the areas and for each year in the time window: 1990-2010, ending up with three variables for each spatial area (and region) across the time window. The data was standardised (sample mean removed from each observation, which is then divided by the standard deviation) prior to conduction of the experiments.

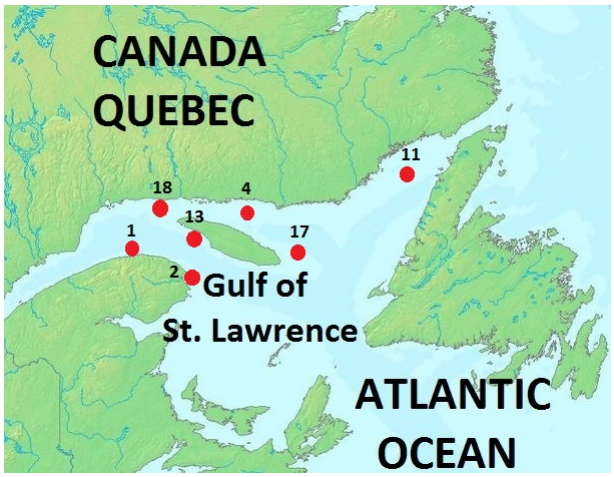


Figure 7.1: Locations of the spatial clusters for the Gulf of St Lawrence

7.3 Methods

7.3.1 Generalisation between Different Geographic Regions to Predict Species Biomass

We learn dynamic Bayesian networks (DBNs) for 7 geographically different sub-regions within the Gulf of St. Lawrence (each with its own network structure) and evaluate these Bayesian models on 7 different areas within the North Sea .

The experiments involved prediction of aggregated species biomass for pelagics (P),

small piscivorous (SP) and large piscivorous (LP) fish species. For detailed description of modelling time series and predicting species biomass using DBNs look at Chapter 4, sub-section 4.3.2 *Hidden Variable Models to Predict Regime Shifts and Species Biomass*. Fig.7.2 represents an example DBN for cluster 2 of the Gulf with a hidden variable and the three observed variables. The model structure for each one of the clusters of the Gulf is defined by the data-learning optimization technique that was applied and explained in Chapter 5, sub-section 5.3.1 *Introducing a Technique to Learn the Structure of Bayesian Network Models*, but only links with a *threshold* of ≥ 0.3 were kept.

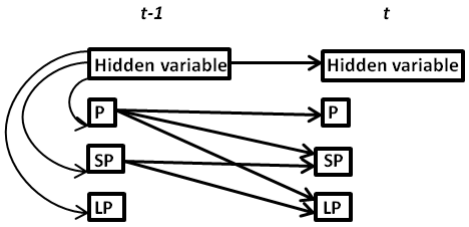


Figure 7.2: The dynamic Bayesian network structure for cluster 2 from the Gulf of St Lawrence that was used to model the dynamics for the areas within the North Sea.

Each dynamic model structure from the Gulf was evaluated on each area from the North Sea, thus resulting in 7 model variants for each area from the North Sea or 49 models altogether. From now on, we will refer to each of those model variants by non-spatial DBNs. In addition, to incorporate potential spatial autocorrelation in our models, we enforce a parent node that represents the average biomass of the combined trophic groups from the spatial neighbourhood (the three or four nearest neighbours) of the current area (model referred as spatial DBN). This adds up to another 49 model variants with a spatial node altogether for the North Sea. Fig.7.3 shows the model structure for cluster 13 from the Gulf with a spatial node added. We wanted to compare if the inclusion of spatial nodes results in improvement of predictive accuracy. Non-parametric bootstrap was applied 250 times for each model variant to identify statistical confidence in the predictions. Model performance was compared by the sum of squared error (SSE).

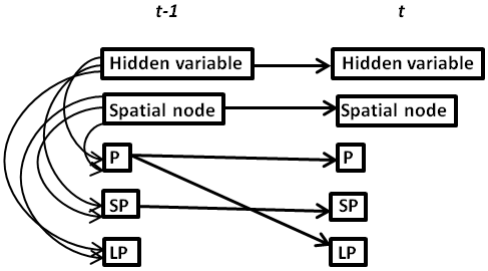


Figure 7.3: The spatial DBN structure for cluster 13 from the Gulf of St Lawrence that was used to model the dynamics for the areas within the North Sea.

7.3.2 Modelling Hidden Variables to Identify Controlling Mechanisms and Drivers of Change in the North Sea

The hidden variable was parameterised using the EM algorithm and then validated against the observed variable it might represent (either fisheries exploitation, temperature or zooplankton in the North Sea, which are all key drivers of change in this region). The distribution of the hidden variable was compared to the observed variable it might represent using the Mann-Whitney U test. The null hypothesis, that is being tested by Mann-Whitney U, is that the distributions of two groups (here, hidden variable and temperature for example) are identical. Note, if the p value is small ($p \leq .05$), we reject the null hypothesis and conclude that the two distributions are distinct, however if the p value is large ($p \geq .05$), the data do not give us any reason to reject the null hypothesis which does not necessarily mean that the two distributions are identical but there is no compelling evidence that they differ significantly (Cheung & Klotz (1997)). All statistical tests were reported at 5% significance level.

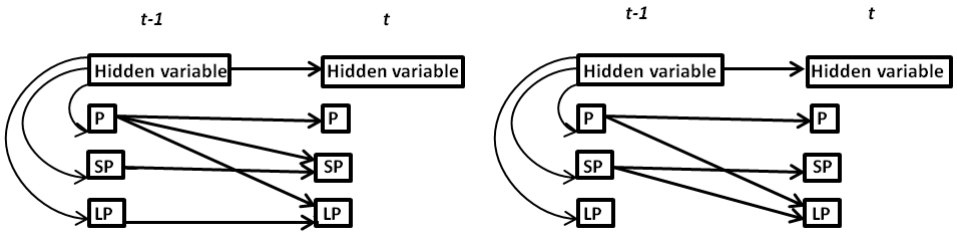
7.4 Results

7.4.1 Generating Biomass Predictions for the North Sea using Dynamic Models from the Gulf of St Lawrence

Results for the North Sea showed varying predictive accuracy with the different spatial clusters from the Gulf of St Lawrence, highlighting some sub-regions that consistently

perform better than others.

Some clusters from the Gulf were highlighted in terms of generating most accurate overall biomass predictions per area. For example, clusters 11 and 17 from the Gulf produced the most accurate predictions for majority of the areas. The structure for these clusters is shown in Fig.7.4. These results confirm the significance of interactions such as $P->LP$, highlighting the importance of species group P for the trophic dynamics of larger predators from the species group LP. In addition, the successful performance of these structures in modelling the species dynamics for more than one area within the North Sea, highlights the similarity between such areas in terms of ecological structure and importance of some trophic groups compared to others.



(a) Cluster 11 (b) Cluster 17

Figure 7.4: The non-spatial DBN structures for clusters 11 (a) and 17 (b) from the Gulf of St Lawrence that were used to model the dynamics of species in the areas of North Sea.

It was interesting to find out that even after the addition of the spatial node in the model structure for the Gulf, some clusters were outlined again in terms of their overall predictive performance within North Sea areas. In all areas, except for one- area 4, the same clusters (after the addition of a spatial node) and their relevant structures from the Gulf modelled the species dynamics in the North Sea with the highest accuracy. This strengthens the confidence in our methods in terms of detecting functionally equivalent species interactions, even the varying level from natural and anthropogenic sources on different marine ecosystems. Our results highlight the similarity of the underlying trophic dynamics and the relevant importance of different species groups and interactions for the underlying local ecological networks. The only exception from this-

area 4, we suggest to be due to the less influential spatial effect on this area's dynamics and potentially higher influence from local mechanisms such as recruitment, prey availability and migration.

In addition, we further confirm the importance of including spatial autocorrelation when building predictive ecological models. We found that in a majority of the areas the predictive performance was improved once the spatial node was included in the model structure (Table 7.1). There were some exceptions to that, for example in areas 2, 4 and 6, however the negative influence from adding the spatial node was insignificant as it only slightly lowered the accuracy of the predictive results.

Table 7.1: Sum of Squared Error (SSE) of the overall trophic group species predictions for the North Sea, generated by the non-spatial and spatial DBNs from the Gulf of St Lawrence (GSL). The cluster from the GSL that produced the most accurate overall biomass predictions is shown in brackets.

Area: North Sea	Non-spatial DBN	Spatial DBN
1	22.31 (GSL cluster 17)	19.16 (GSL cluster 17)
2	15.21 (GSL cluster 18)	16.96 (GSL cluster 18)
3	17.12 (GSL cluster 11)	15.85 (GSL cluster 11)
4	19.57 (GSL cluster 17)	20.29 (GSL cluster 11)
5	19.36 (GSL cluster 1)	19.04 (GSL cluster 1)
6	15.19 (GSL cluster 11)	15.52 (GSL cluster 11)
7	20.13 (GSL cluster 11)	19.82 (GSL cluster 11)

We now present the biomass predictions for the individual species groups to explore the models' predictive accuracy and spatial consistency. We can see from Table 7.2 that some clusters from the Gulf were highlighted in terms of producing the most accurate predictions for P, SP and LP in the North Sea. For example, the P species group was most accurately predicted by cluster 1, followed by clusters 11 and 4 from the Gulf. Genuinely, predictions for the P species group were most accurate followed by predictions for the LP species group, suggesting some of the common dynamics such species undergo in different ecosystems. Specifically, large predators have undergone periods of decline due to high commercial value and their consequent exploitation. Similarly to P species, cluster 1 modelled the most accurate SP predictions, followed by cluster 13. Finally, for the LP species group, cluster 17 from the Gulf modelled this

species group dynamics with accuracy for majority of the North Sea areas. In cluster 1, the dynamic model structure consists of trophic interactions $P->SP$ and $P->LP$, whilst for cluster 17 the structure was characterised by the relationships: $P->LP$ and $SP->LP$, outlining the significance of these interactions for the overall structure and functionality in different marine ecosystems.

Table 7.2: Sum of Squared Error (SSE) of P, SP and LP predictions for the North Sea, generated by the non-spatial DBN from the Gulf of St Lawrence (GSL). The cluster from the GSL that produced the most accurate biomass predictions is shown in brackets.

Area: North Sea	P	SP	LP
1	19.72 (GSL cluster 1)	19.10 (GSL cluster 13)	25.63 (GSL cluster 17)
2	10.57 (GSL cluster 4)	19.89 (GSL cluster 13)	10.59 (GSL cluster 18)
3	9.77 (GSL cluster 4)	17.11 (GSL cluster 1)	22.11 (GSL cluster 11)
4	21.28 (GSL cluster 11)	19.37 (GSL cluster 13)	17.02 (GSL cluster 17)
5	19.86 (GSL cluster 1)	15.69 (GSL cluster 11)	21.68 (GSL cluster 1)
6	20.03 (GSL cluster 1)	15.40 (GSL cluster 1)	7.92 (GSL cluster 11)
7	18.71 (GSL cluster 11)	17.76 (GSL cluster 1)	19.80 (GSL cluster 17)

Now, let us look at the individual species predictions after the addition of a spatial node to the model structure. Overall, the majority of the individual species predictions improved once the spatial node was included in the model. There were only very few cases that the opposite was seen but similarly to the overall spatial comparison, the negative influence was not severely drastic. Thus, we further confirm the importance of spatial autocorrelation and the fact that one area's dynamics likely influence another in terms of building ecological approaches for modelling trophic dynamics.

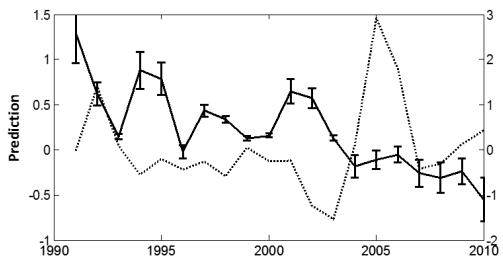
What was more interesting to see was that, the spatial structure for some of the clusters in the Gulf was also the structure (before adding the spatial node) to accurately predict the species dynamics in the same areas within the North Sea (Look at the * symbol in Table 7.3). That was the case for majority of the species groups SP and LP, whist for the P species group, the accurate performance for only one cluster remained unchanged after adding the spatial node. These results suggest the similarity of the underlying dynamics for medium to larger sized predators following natural and anthropogenic sources but also the significance of smaller pelagic species for the ecological functionality and stability. In addition, it was interesting to see that in the areas, for which the Gulf structure performed well in modelling the North Sea dynamics, were also the areas,

in which there was a slight negative influence from adding the spatial node. This suggests that in such areas the functional relationships between species and ecological mechanisms are much more influential comparing to spatial interactions.

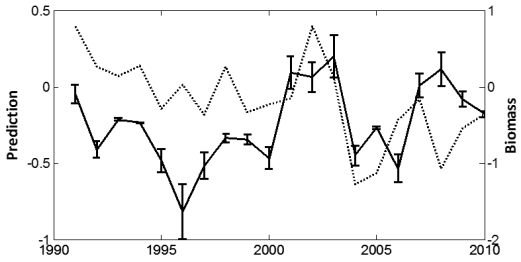
Table 7.3: Sum of Squared Error (SSE) of *P*, *SP* and *LP* predictions for the North Sea, generated by the spatial DBN from the Gulf of St Lawrence (GSL). The cluster from the GSL that produced the most accurate biomass predictions is shown in brackets. The * symbol indicates the clusters from the non-spatial DBN that also generated the most accurate predictions for the relevant species group within the North Sea.

Area: North Sea	P	SP	LP
1	8.90 (GSL cluster 4)	18.98 (GSL cluster 13*)	18.93 (GSL cluster 17*)
2	10.44 (GSL cluster 18)	19.96 (GSL cluster 13*)	10.23 (GSL cluster 1)
3	9.97 (GSL cluster 18)	13.60 (GSL cluster 18)	16.60 (GSL cluster 11*)
4	19.86 (GSL cluster 1)	19.22 (GSL cluster 1)	18.06 (GSL cluster 11)
5	19.96 (GSL cluster 1*)	15.83 (GSL cluster 11*)	21.32 (GSL cluster 1*)
6	20.65 (GSL cluster 13)	14.93 (GSL cluster 18)	8.75 (GSL cluster 11*)
7	15.97 (GSL cluster 4)	17.03 (GSL cluster 1*)	23.10 (GSL cluster 11)

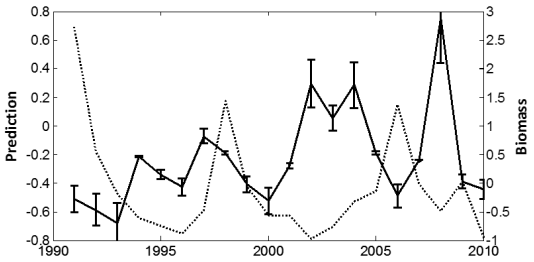
Now, we move onto the generated predictions of the species groups and specifically their characteristics throughout time. For example, Fig.7.5 shows the predictions for *P*, *SP* and *LP* for area 3 in the North Sea generated by the non-spatial DBN from cluster 11 in the Gulf. We can see that the model is genuinely capturing some of the temporal variations, specifically for *P* and *SP*. *SP* (Fig.7.5b) was characterised with some increase in early 2000, that was captured by the model. The network structure that was used to generate these predictions is shown in Fig.7.4a and we can see that the *SP* species group is influenced by the dynamics of species group *P*.



(a) *P*, Area 3



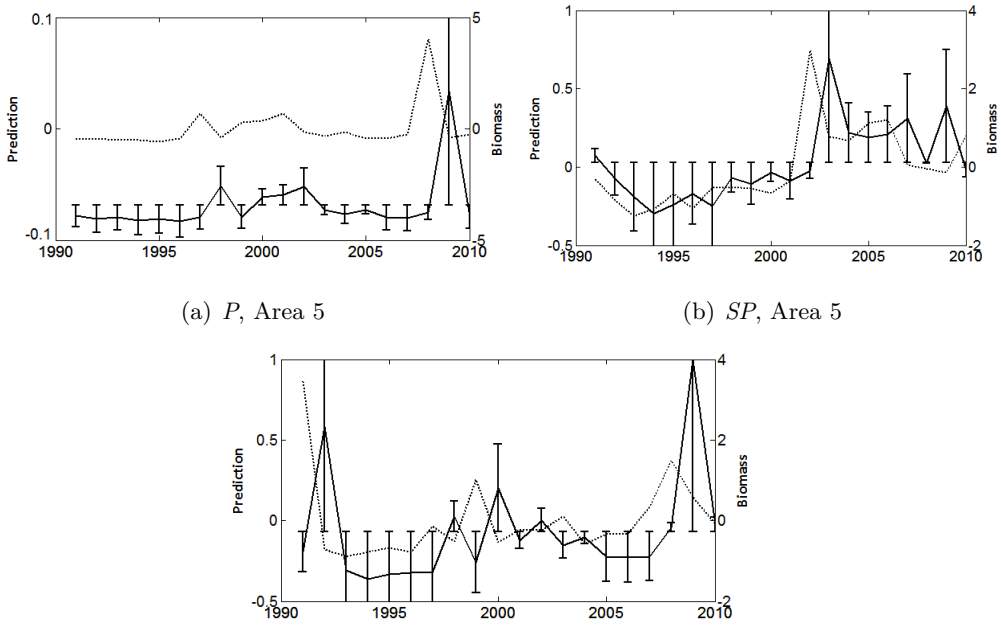
(b) *SP*, Area 3



(c) *LP*, Area 3

Figure 7.5: Generated biomass predictions for species groups *P* (a), *SP* (b) and *LP* (c) for area 3 in the North Sea by the non-spatial DBN model from cluster 11 in the Gulf. Predictions are shown by the solid line and standardised observed biomass by the dotted line.

Let us look at another example- cluster 1 from the Gulf modelling the species dynamics for area 5 in the North Sea. The *P* species group was characterised by a relatively stable trend throughout time with the exception of some year effects, which were captured by the model (Fig.7.6a). *SP* was characterised by a declining trend in the beginning of the time series but it started to increase in some recent future years, both of which were projected with high accuracy by our dynamic model (Fig.7.6b). Similarly to *P*, it was difficult to characterise *LP* with a trend, however our model managed to reflect on the strong fluctuation throughout time (Fig.7.6c)



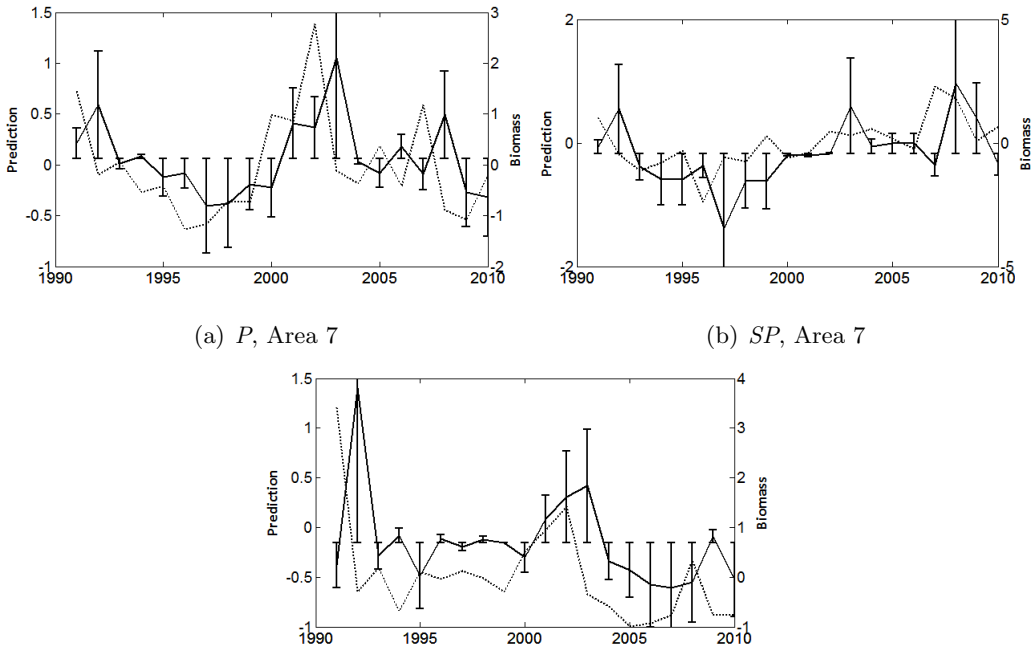
(a) *P*, Area 5

(b) *SP*, Area 5

(c) *LP*, Area 5

Figure 7.6: Generated biomass predictions for species groups *P* (a), *SP* (b) and *LP* (c) for area 5 in the North Sea by the non-spatial DBN model from cluster 1 in the Gulf. Predictions are shown by the solid line and standardised observed biomass by the dotted line.

Finally, let us look at the predictions by the non-spatial dynamic model for cluster 11 from the Gulf evaluated on area 7 in the North Sea. Here, our model reflected on the temporal variations for the *P* species group (Fig.7.7a) and captured the declining trend of the *LP* species group (Fig.7.7c). Predictions for the *SP* species group (Fig.7.7b) were probably least accurate. It was difficult to detect a temporal trend for this species group, there were some distinct individual year effects for example in late 1990s and early 2000, but our model managed to capture only some of these fluctuations.



(a) *P*, Area 7

(b) *SP*, Area 7

(c) *LP*, Area 7

Figure 7.7: Generated biomass predictions for species groups *P* (a), *SP* (b) and *LP* (c) for area 7 in the North Sea by the non-spatial DBN model from cluster 11 in the Gulf. Predictions are shown by the solid line and standardised observed biomass by the dotted line.

To summarise, we managed to show the success of our dynamic model in terms of capturing species variations throughout time, which to recall was trained on data from the Gulf of St Lawrence and evaluated on data from the North Sea. Predictions were not always perfect, which we suggest to be due to couple of things. First, there could be some effect from external factors such as temperature and climate or effect from other mechanisms such as prey availability or habitat degradation, that we did not account for in the model, which could have had some negative influence on the produced predictions here. Another thing, which is probably most likely, is that such ecological data is very noisy and complex characterised with strong variability throughout time and a more complex structure would be needed to perfectly capture the species dynamics, as we showed in Chapter 5 for the North Sea.

However, here the goal was not to find the "perfect" structure that models the trophic

dynamics but to generalise functional interactions between different marine ecosystems in order to find out what are the key relationships between species groups, which we showed by modelling species dynamics from different datasets. Also, we managed to generalise the influence from natural and anthropogenic forces in terms of capturing genuine temporal trends such as the decline in medium to larger-sized species that are of high commercial importance and have suffered severe declines throughout time and in different spatial ecosystems.

7.4.2 Identified Driving Factors for the Ecosystem Dynamics in the North Sea

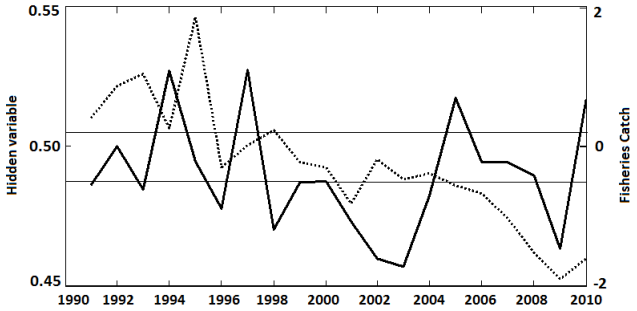
To recall, we include a hidden variable in our models to see if the hidden variable reflects any changes in the species variance or potentially captures the dynamics of other mechanisms such as temperature or fisheries catch that might be influencing the species groups. Then, the hidden variable was validated against the observed variable it might represent. To recall, we are comparing the distribution of the expected value of the hidden variable versus the distribution of the observed variable it might represent. The null hypothesis, that is being tested by Mann-Whitney U, is that the distributions of two groups (here, hidden variable and fisheries catch or temperature) are identical. Note, if the p value is small ($p \leq .05$), we reject the null hypothesis and conclude that the two distributions are distinct, however if the p value is large ($p \geq .05$), the data do not give us any reason to reject the null hypothesis which does not necessarily mean that the two distributions are identical but there is no compelling evidence that they differ significantly (Cheung & Klotz (1997)).

For majority of the areas, we found the hidden variable to be modelling either fisheries catch or zooplankton however for some areas, the hidden variable was modelling temperature, suggesting the importance of the relevant species group and/or external stressor for the structure and functionality of the local dynamics. Here, we are going to present the expected value for some of the hidden variables (versus the observed variables they might represent) from the clusters in the Gulf that managed to most accurately predict the species dynamics within the North Sea (from Table 7.1).

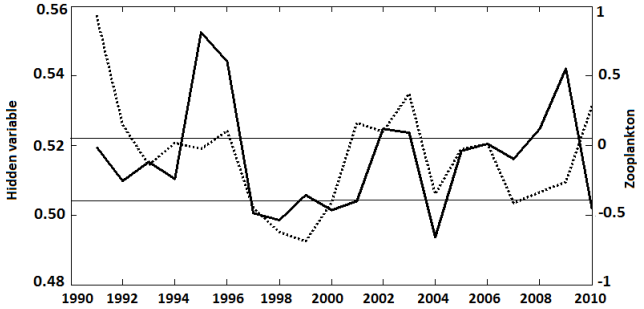
First, we show what the expected value of the hidden variable from the non-spatial DBN model is potentially modelling for each area for the North Sea (Table 7.4). It was interesting to see that in only area 5, the hidden variable was modelling temperature. As expected, in areas 1 (Fig.7.8a) and 2, the hidden variable was modelling the fisheries catch, as we know from earlier work (Chapter 5) these are key areas of fisheries exploitation characterised with complex and diverse dynamics. Specifically, the hidden variable in Fig.7.8a managed to reflect on the declining trend of the catch in early 2000. Also, it was interesting to see that in areas 3 and 4, the hidden variable was modelling the trend of zooplankton in the North Sea, as we have previously shown for these areas (from Chapter 5), the trophic dynamics and species associations are of bigger influence comparing to fisheries exploitation. In fig.7.8b, we can see that the learned hidden variable for area 4 is projecting a decline that coincided with the decline in zooplankton in late 1990s to early 2000, also capturing some of the individual year effects, for example from years 2002 and 2004.

Table 7.4: Summary of the expected value of the hidden variable generated by the non-spatial DBN model from the Gulf of St Lawrence (GSL) for areas 1 to 7 in the North Sea. The table shows what the hidden variable is expected to model and the relevant summary statistics from the Mann-Whitney U test.

Area: North Sea	Non-spatial DBN	Statistics (Mann-Whitney U)
1	Fisheries catch (GSL cluster 17)	Z= 1.59, d.f.=19, p=0.11
2	Fisheries catch (GSL cluster 18)	Z= 0.17, d.f.=19, p=0.34
3	Zooplankton (GSL cluster 11)	Z= 1.10, d.f.=19, p=0.86
4	Zooplankton (GSL cluster 17)	Z= 1.34, d.f.=19, p=0.37
5	Temperature (GSL cluster 1)	Z= 0.05, d.f.=19, p=0.89
6	Zooplankton (GSL cluster 11)	Z= 0.56, d.f.=19, p=0.99
7	Fisheries catch (GSL cluster 11)	Z= 1.12, d.f.=19, p=0.73



(a) Hidden variable, Area 1



(b) Hidden variable, Area 4

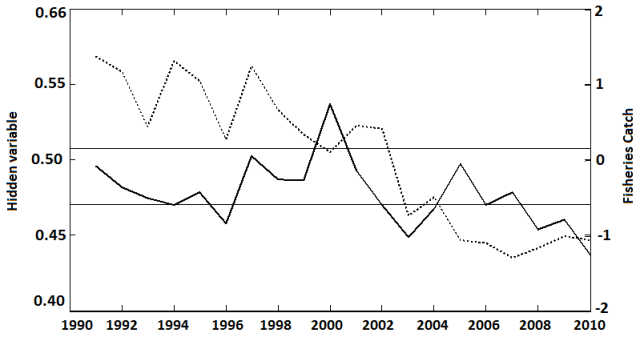
Figure 7.8: The expected value of the hidden variables (solid line) for areas 1 (a) and 4 (b), generated by the non-spatial DBN model from cluster 17 in the Gulf, with the standardised observed data (dotted line) for fisheries catch and zooplankton in the North Sea. The solid lines indicate upper and lower 95% confidence intervals.

Now, let us look at the expected values of the hidden variables produced by the spatial DBN models from the Gulf. As expected, we can see that the hidden variables from the spatial models are modelling the same observed variables for the different areas in the North Sea, as did the hidden variables from the non-spatial models, with the exception of two areas- 4 and 7 (Table 7.5). The fact that the same observed variables were modelled is due to the potential importance of these mechanisms on the underlying ecological structure and stability, even after adding the influence from species spatial relations. As for area 4, after the addition of the spatial node in the model, the hidden variable modelled the fisheries catch, rather than zooplanknton, which could

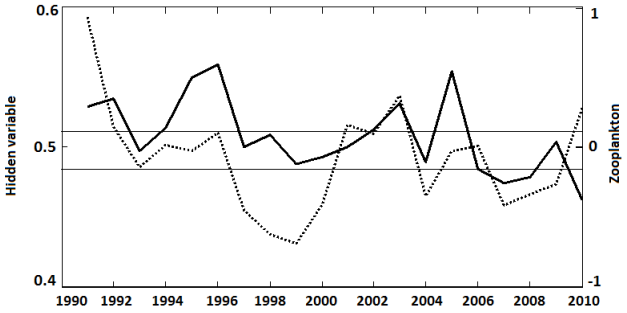
be due to the increased influence from neighbouring species and the importance of fisheries catch on the overall trophic dynamics. We can see that the hidden variable reflected on the declining trend of the catch, specifically in late 2000s (Fig.7.9a). The opposite was found for area 7- originally the hidden variable from the non-spatial model projected similarity with the trend of the fisheries catch but after the addition of the spatial node, the hidden variable from the spatial dynamic model projected similarity with the zooplankton. This suggests that accounting for spatial autocorrelation adds more accuracy to the importance of certain species groups in terms of modelling trophic dynamics and structure. From Fig.7.9b, we can see that the hidden variable managed to capture the zooplankton temporal variations, specifically some of the decline in late 1990s and late 2000s.

Table 7.5: Summary of the expected value of the hidden variable generated by the spatial DBN model from the Gulf of St Lawrence (GSL) for areas 1 to 7 in the North Sea. The table shows what the hidden variable is expected to model and the relevant summary statistics from the Mann-Whitney U test.

Area: North Sea	Spatial DBN	Statistics (Mann-Whitney U)
1	Fisheries Catch (GSL cluster 17)	Z= 0.77, d.f.=19, p=0.10
2	Fisheries Catch (GSL cluster 18)	Z= 0.10, d.f.=19, p=0.56
3	Zooplankton (GSL cluster 11)	Z= -0.61, d.f.=19, p=0.21
4	Fisheries Catch (GSL cluster 11)	Z= 0.12, d.f.=19, p=0.18
5	Temperature (GSL cluster 1)	Z= -0.59, d.f.=19, p=0.79
6	Fisheries Catch (GSL cluster 11)	Z= 1.13, d.f.=19, p=0.88
7	Zooplankton (GSL cluster 11)	Z= 0.75, d.f.=19, p=0.22



(a) Hidden variable, Area 4



(b) Hidden variable, Area 7

Figure 7.9: The expected value of the hidden variables (solid line) for areas 4 (a) and 7 (b), generated by the spatial DBN model from cluster 11 in the Gulf, with the standardised observed data (dotted line) for fisheries catch and zooplankton in the North Sea. The solid lines indicate upper and lower 95% confidence intervals.

To summarise, we managed to show that the hidden variables from our models captured the temporal variations and variance of different observed variables and external stressors, which allowed us to confirm key mechanisms in terms of defining ecological structure and stability across spatial scales. Although, the studied marine ecosystems exhibit different level of exploitation and climate variability, we managed to show what are the key factors defining trophic dynamics and species groups acting in an ecosystem. Highlighted mainly was the importance of fisheries exploitation, followed by zooplankton and finally temperature.

7.5 Summary

The use of dynamic Bayesian models in this work is unique because it exploits functional equivalence between different datasets in conjunction with hidden variables to predict biomass and trophic dynamics. The recognition of a hidden variable is important in fish community change studies of this nature because it allows causes of change which are not purely found within the constrained model structure.

Our findings demonstrate that although the effects of external stressors differ across ecosystems, general results on the trophic structure and functioning can be drawn. Our results show that different ecosystems can organize themselves functionally in a similar way and provide us with insights into why fished ecosystems collapse and why they sometimes do not recover. By adding spatial nodes to the models to correct for spatial autocorrelation, the predictive performance was preserved for some of the areas, confirming key functional interactions. Overall, the learned hidden variables were complex and varied throughout the NS areas, possibly due to severe climate-ocean interactions and fisheries exploitation. However, these hidden variables appeared to capture temporal changes of the fisheries catch and zooplankton, outlining the importance of these mechanisms on the trophic dynamics in different exploited systems.

Here, our Bayesian network models also facilitate the direct incorporation of data-driven interactions into the structures and parameters. The modelling approach also differs from other methods in how correlative structures, which are assumed to represent functional relationships, discovered in one system can be imposed upon another system. The components of the other system which best fit these structures can then be found in other systems. What our approach assumes is that there are only a few ways for similar ecosystems to organize themselves structurally (or functionally) even though the ecosystem components may have different qualities, thus our results suggest that ecosystems respond similarly to natural and anthropogenic forces. This can provide real insights into why fished ecosystems collapse and why they sometimes do not recover when a perturbation stops, consequently providing strategic advice to managers and policy makers.

The next chapter discusses the methods and the findings of this thesis, highlighting

contributions of the research, limitations and future works.

Chapter 8

Conclusions

This chapter discusses the conclusions reached, based on the research presented in this thesis. Firstly, the research contributions are outlined, followed by an analysis of the limitations. Finally, potential future work is explained, addressing both research limitations and extending the applicability of the work.

8.1 Thesis Contributions

8.1.1 Functional Network Models with Hidden Variables

Literature analysis showed that marine populations are being threatened by both natural and anthropogenic sources and in order to understand issues between sustainability and management, populations cannot be addressed in isolation but in relation to their interactions and associations with external factors. In this research, we introduce and fully define the concept of the functional network approach of modelling species dynamics. Functional network models are dynamic models that take into consideration species interactions (both direct and indirect) and their associations with climate and fisheries exploitation. The functional network approach is flexible and explorative, allowing for complex relationships to be recovered from collected field data, thus providing a metric for assessing community structure and resilience in response to natural and anthropogenic influences.

We exploited functional equivalence between different datasets in combination with

dynamic models that used hidden variables to predict functional collapse and species biomass. We explored the use of recognised statistical metrics in conjunction with hidden variables which proved useful (compared to those without) in terms of detecting early-warning signals (significantly rising variance and hidden variable fluctuations) of functional changes. The hidden variable managed to reflect some of the metrics' characteristics in terms of identifying regime shifts that are known to have occurred and it also appeared to capture changes in the variance of different species biomass. Therefore, the recognition of a hidden variable is important in fish population studies because it represents something external to the fish community and trophic dynamics, that is not purely constrained within the model structure.

We extended the concept of functional equivalence and showed that different ecosystems can organize themselves functionally similar. By applying dynamic models that essentially represent functional relationships between species we showed that trophic interactions discovered in one system can be imposed upon another system. Thus, our analysis suggested that even the varying levels of exploitation and climate variability across marine ecosystems and their species with different qualities, there are only few ways ecosystems can organise themselves functionally. This further implies that there will be similar ways of collapsing and forcing from external factors.

8.1.2 Hidden Spatial Dynamic Bayesian Network Model

We developed a modelling approach that analyses multiple associations between groups of species and their environment accounting for unmeasured effects and spatial autocorrelation. Our proposed model was able to produce novel insights on ecosystem's dynamics and ecological interactions mainly because we account for the heterogeneous nature of the driving factors within spatially differentiated areas and their changes over time. Our findings demonstrate that accounting for additional sources of variation, by combining structure learning from data and experts' knowledge in the model architecture, has the potential for gaining deeper insights into the structure and stability of ecosystems. We were able to find highly confident but spatially and temporally differentiated ecological networks that indicate spatial relationship of species and habitat

with the particular mechanisms varying from facilitation through trophic interactions. The model has further highlighted the importance of spatial heterogeneity in modelling trophic dynamics, the value of accounting for latent effects in learning biomass changes across space and time but also the need for further understanding species-specific effects in their habitat following anthropogenic disturbances. Expert knowledge is important, however allowing for data-driven interactions and associations when building predictive models also should be performed, as we have done in this work, which allowed us to identify patterns in the characteristics of different spatial areas and mechanisms involved in shaping the functional ecological networks.

The model was further modified and we extended its use in combination with fisheries catch, temperature and productivity scenarios to predict the future biomass of individual fish species and model their response to change in pressures. Our findings demonstrated that reducing fisheries exploitation will potentially lead to trade-offs between species in terms of uniform population recovery. In addition, other factors such as trophic interactions and the spatial relationship between neighbouring areas should be taken into consideration when designing management strategies. Fisheries management measures will contribute to improvements in the biodiversity of the fish community, but food web interactions and productivity changes will mediate changes. Our results allow dynamic assessment of choices, which should be able to provide strategic advice on potential system response to pressure.

8.1.3 Application to Several Datasets

The hidden spatial dynamic Bayesian network was applied to the North Sea, as this region has been a key centre of research in this work, due to its high biological productivity and valuable fisheries resources. The successful performance of our model highlighted the heterogeneous nature of the species dynamics, providing us with more accurate insights on the structure of the underlying local ecological networks. Thus, extending our knowledge into the complexity of North Sea dynamics and its ecological resilience and stability. For most fish species, we were able to predict a trend of increase or decline in response to change in fisheries catch. Therefore, we imply that for a given

area, reorganisation of the fishing fleet and management strategies will be required to ensure that the right species are targeted and harvested sustainably.

In addition, we analysed datasets from the regions of Georges Bank and East Scotian Shelf. Our results were able to show the reliability of our modelling approach in terms of detecting early-warning signals of functional change across different geographic regions. Specifically, the hidden variable alone (without regime metrics) managed to reflect the correct ecosystem dynamics, which was more evident in the East Scotian Shelf region, where a larger regime shift has occurred. We also showed that our dynamic models can be used to identify key functional species in the East Scotian Shelf, based on a regime shift in Georges Bank.

Finally, a dataset from the Gulf of St Lawrence was also analysed. We were able to discover meaningful ecological networks that were more precisely spatially-specific rather than temporally, which indicated spatial relationship of species and habitat. We were able to learn functional network models for this region and evaluate these models on different areas within the North Sea, being able to identify key functional interactions to generalise influence from natural and anthropogenic forces.

8.2 Limitations

The hidden spatial dynamic Bayesian network model was found to be the most thorough and comprehensive model choice to handle our ecological data. Although, we detect and highlight here its limitations.

- **Data Quality** In this work, we focus on ecological temporal datasets and our modelling approach is consequently adapted to the analysis of this type of data. Ecological data is genuinely noisy (relating to the complexity of the data) due to the multiple natural processes involved in generating such data and some sampling variation for the survey data (according to experts), which could have explained the sensitivity of the generated biomass predictions from our modelling approach.

- **Small Samples** Our functional network models are based on a relatively small set of samples, which if not combined together may influence the reliability of our

results.

- **Structural Uncertainties** The models described and compared in Chapter 5 use different criteria for parameters and structure learning. The species dynamics and relationships learned with one model may not be well supported in the training phase, or in the inference, using another modelling architecture. This issue affects the performance of any model, but also provides us with interesting insights on the nature of the different hypothesis that models reveal for an ecosystem. We note that perfect reconstruction of an ecosystem's structure is unlikely due to the noisy data and complex ecological process involved or issues related to selecting the correct type of variables.

- **Running Time** While the majority of the steps, involved in the construction of our model, managed to run in a reasonable amount of time, the step involving structure-learning (hill-climb) and the step involving the use of inference to calculate the hidden variable (Expectation Maximisation algorithm) is actually relatively long, which has resulted in the algorithm running overnight and sometimes up to two days.

8.3 Future Work

The following sections bring to the attention potential future work, based on the limitations discussed above and the extension of the methods presented in this thesis.

8.3.1 Application to Different Kind of Data

Our modelling framework was built to handle complex systems such as the North Sea, so consequently we assume there is a degree of complexity when modelling fisheries interactions. The assumptions are based on key processes within the environment, accounting for influence from external factors such as fisheries catch. One aspect of the underlying processes that could be further investigated includes fishermen behaviour or effort information to estimate catch potentials for distinct fleets. Few adjustments in the model could potentially allow it to handle technical interactions alongside biologi-

cal information in order to provide routine advice for mixed-fisheries management and quantify the socio-economic aspect when looking at future sustainability.

The model has the potential to be "transferred" across stakeholder groups in order to increase knowledge communication and awareness. Typically, results from Bayesian Networks are tested with experts and with stakeholders, to evaluate the reliability of the model outputs and the level of accessibility to the model. Therefore, future work will involve testing the model results with stakeholders and experts to improve vertical portability (moving from scientists to managers to stakeholders), and thus bridge the communication gap between fisheries scientists and managers. This will help to facilitate results communication to local stakeholders as well as informing managers on stakeholders fisheries use.

In addition, the model holds the ability to be "transferred" to different ecosystems and used to predict species dynamics and biomass, with only slight re-adjustments of the network structure. The methods provided in this thesis may be used to explore other datasets without the need for many adjustments, apart from setting up the parameters in the learning algorithms. Also, further datasets should be compared to the results presented in this work to explore generalisation across marine ecosystems at a much wider scale. Currently, the modelling framework is being extended to predict regime shifts and model species dynamics within the region of the Baltic Sea. Similarly to the North Sea, it is a diverse and complex region and the modelling approach developed in this thesis will be used to examine species trends in response to natural and anthropogenic scenarios.

Finally, the modelling framework presented in this thesis could be potentially applied to other ecological datasets like plants, mammals or birds. Spatial heterogeneity is an issue that relates to migrating mammals and birds but also there is a reasonable degree of uncertainty in terms of their use of space and interactions with their environment. Thus, applying our hidden spatial dynamic approach could be potentially beneficial in terms of providing insights on other species dynamics that would be beneficial to conservation and management.

8.3.2 Extension of Modelling Techniques

With regards to structural uncertainty and data combination, additional work on ensembles of classifiers could focus on the joined probability distribution and apply different prior knowledge to smooth the parameters. Potential work could involve informative priors based upon available expertise to create scenarios for an environmentally-oriented management. This prior notion can be incorporated into the model by adding a species dependent prior on the average biomass. Prior knowledge could be integrated separately into each individual functional network. This would allow further understanding of the nature of the combination between observed variables and may improve structural performances. The thesis follows the application of local optimization technique such as the hill-climb. Although, other structure learning algorithms could have been used, investigating the impact of other searching techniques was not the aim of this work, in general, some evolutionary algorithms may be efficient and provide good results.

Appendix A

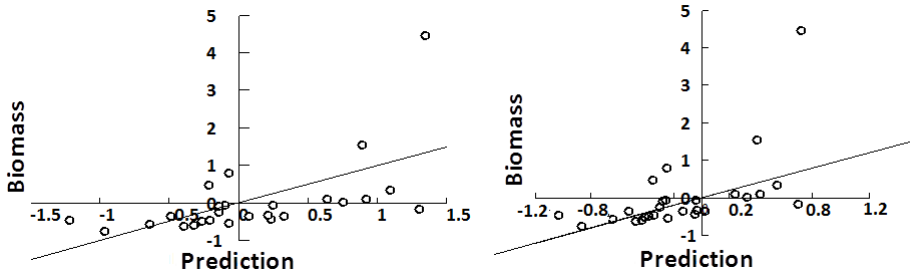
Additional Results

This appendix contains additional plots and results relating to Chapter 5 and Chapter 6.

A.1 Chapter 5 Additional Results

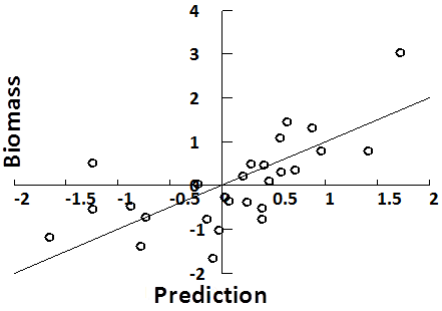
We show the groups of species: pelagics (P), small piscivorous (SP) and large piscivorous (LP) and their biomass predictions in time for some of the areas on which the hidden spatial dynamic Bayesian network model (HSDBN) performed most accurately and compare these with some of the inferior competing models (we focus on models with the least SSE difference: $\leq$5.0, between the overall biomass predictions). The imposed HSDBN network structures for these areas are shown in Fig.A.4

HSDBN outperformed the other probabilistic models in predicting species biomass for area 3 (Fig.A.1a,c,e) and was followed by the hidden dynamic Bayesian network model (HDBN) (Fig.A.1b,d,f) with relatively similar accuracy.

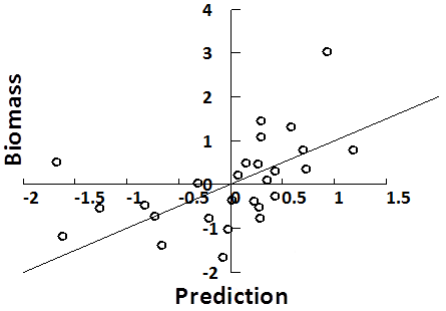


(a) *P*, HSDBN

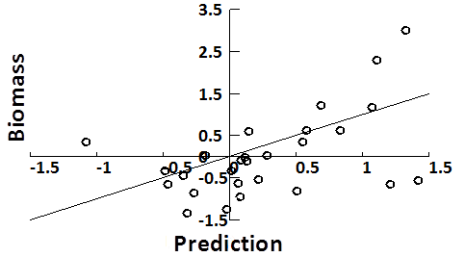
(b) *P*, HDBN



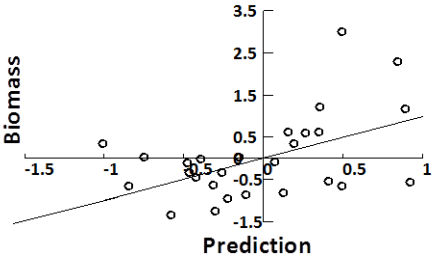
(c) *SP*, HSDBN



(d) *SP*, HDBN



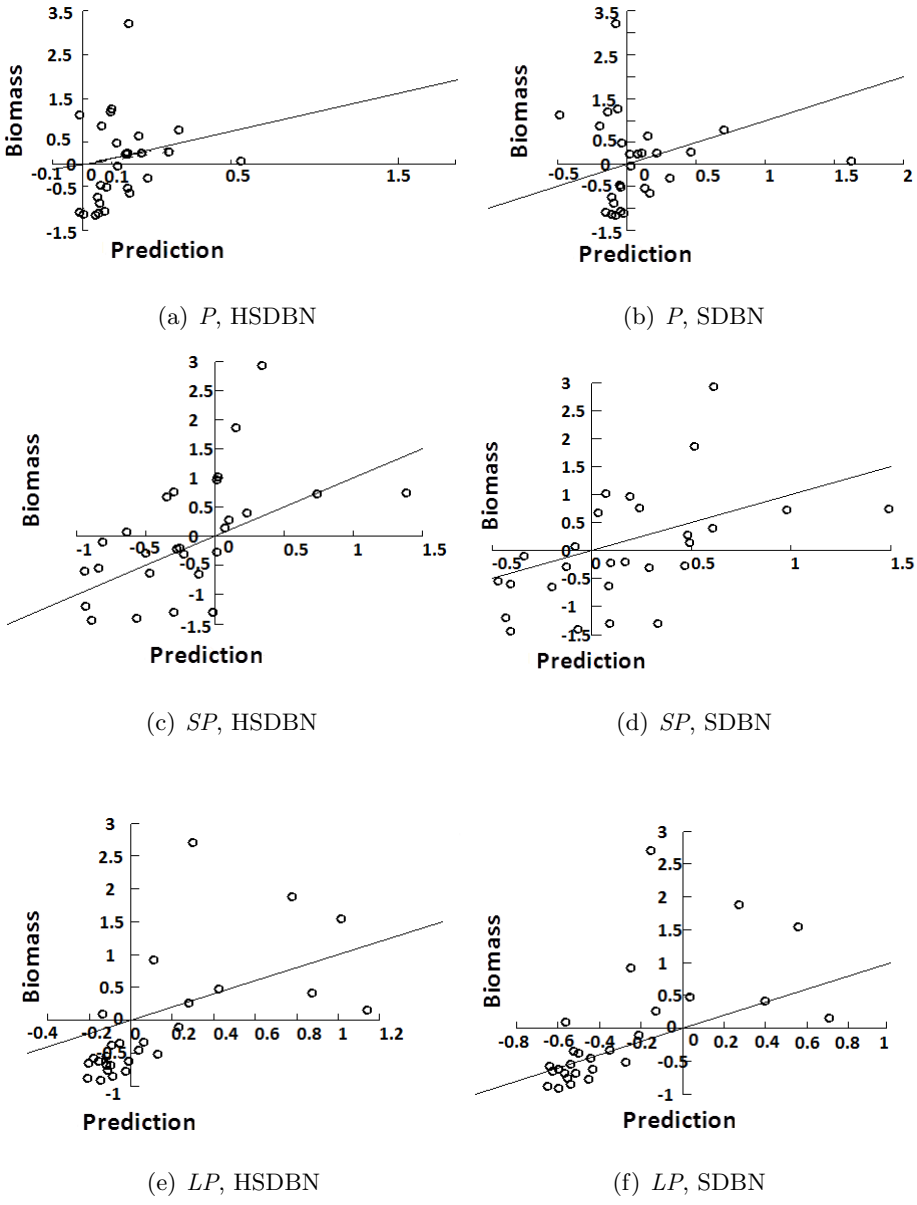
(e) *LP*, HSDBN



(f) *LP*, HDBN

Figure A.1: Observed standardised biomass of *P*, *SP* and *LP* and their predictions generated by HSDBN (a,c,e) and HDBN (b,d,f) for area 3. Note the negative scale is due to standardisation. Diagonal represents perfect prediction.

Our results showed most accurate predictions by the HSDBN for area 6 (Fig.A.2a,c,e) which was followed with similar performance by the spatial dynamic Bayesian network model (SDBN) (Fig.A.2b,d,f).



(a) *P*, HSDBN

(b) *P*, SDBN

(c) *SP*, HSDBN

(d) *SP*, SDBN

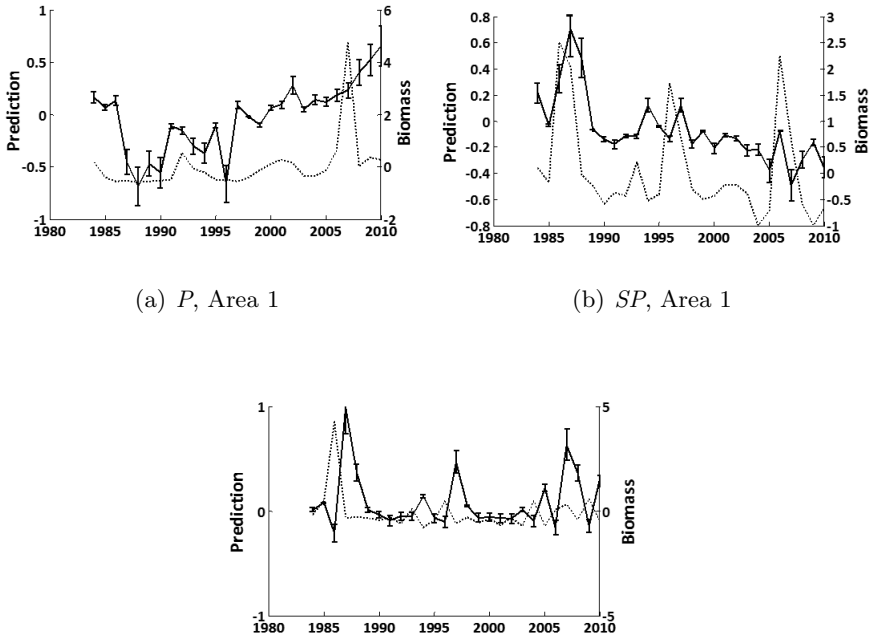
(e) *LP*, HSDBN

(f) *LP*, SDBN

Figure A.2: Observed standardised biomass of *P*, *SP* and *LP* and their predictions generated by HSDBN (a,c,e) and SDBN (b,d,f) for area 6. Note the negative scale is due to standardisation. Diagonal represents perfect prediction.

The HSDBN produced most accurate predictions also for area 1, reflecting on the declining trend specifically for *SP* and *LP* and generally strong individual year effects for *P* (Fig.A.3b,c). The model was able to reproduce the decline in the *SP* and *LP* biomass, when *temperature* was increasing and *fisheries catch* was declining in combination with

strongly varied year effects of the surrounding *spatial P* biomass.



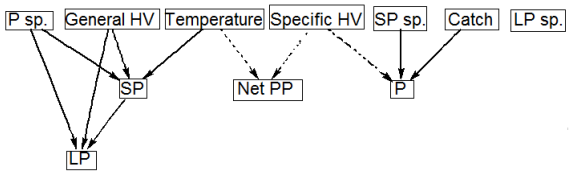
(a) *P*, Area 1

(b) *SP*, Area 1

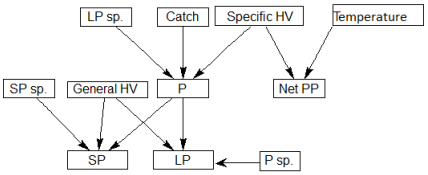
(c) *LP*, Area 1

Figure A.3: Biomass predictions of *P* (a), *SP* (b) and *LP* (c) generated by the HSDBN for area 1. Solid line indicates predictions and dotted line indicates standardised observed biomass. 95% confidence intervals report bootstrap predictions' mean and standard deviation. Note the negative scale is due to standardisation.

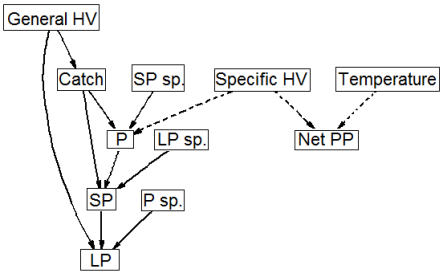
The HSDBN network structures used to generate species predictions and their dynamics for the species groups in the North Sea areas are shown below.



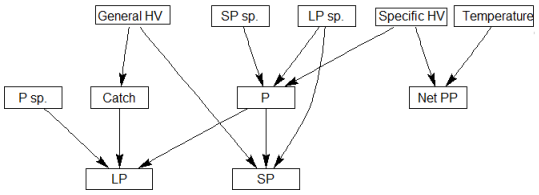
(a) Area 1



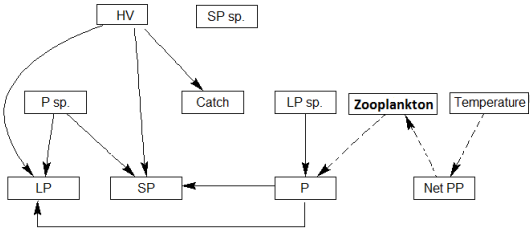
(b) Area 2



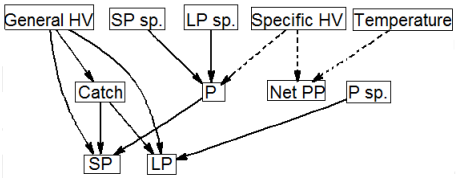
(c) Area 3



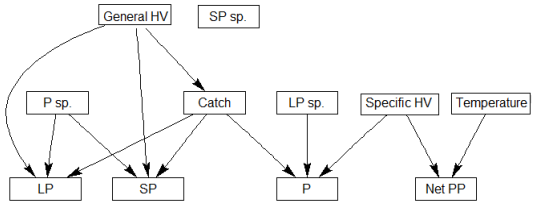
(d) Area 4



(e) Area 5

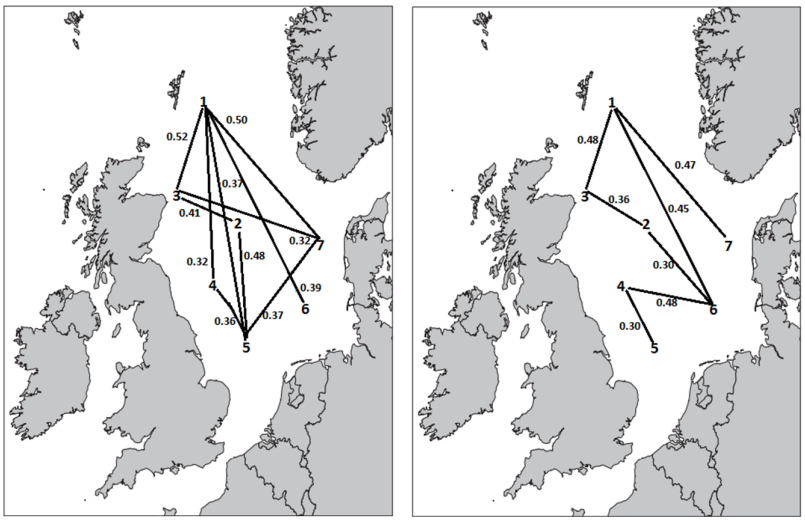


(f) Area 6

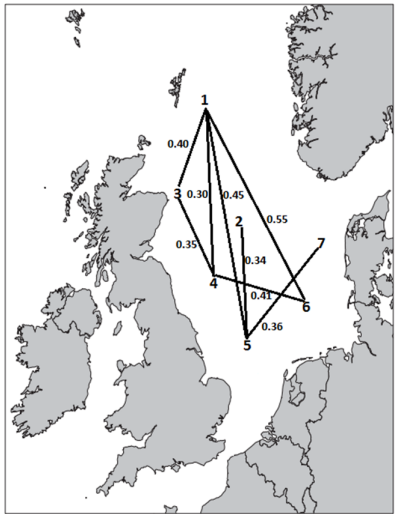


(g) Area 7

Figure A.4: HSDBN network structure for areas 1 to 7 in the North Sea. *HV* stands for hidden variable. Spatial nodes are abbreviated as *P sp.*, *SP sp.* and *LP sp.* Edges between nodes (or variables) represent dependence relationships, the edges shown by a dotted line are defined by the expert.



(a) *P-SP* (b) *P-LP*



(c) *SP-LP*

Figure A.5: Learned trophic interactions *P-SP* (a), *P-LP* (b) and *SP-LP* (c) for all 7 spatial areas (numbered in figure). The estimated mean confidence by the hill-climb, for the time window: 1993-2014 (to recall, longer time series here), is shown above or below the links. Note the links represent dependence, not causality. The time window starts from 1993 due to the windowing required during the hill-climb.

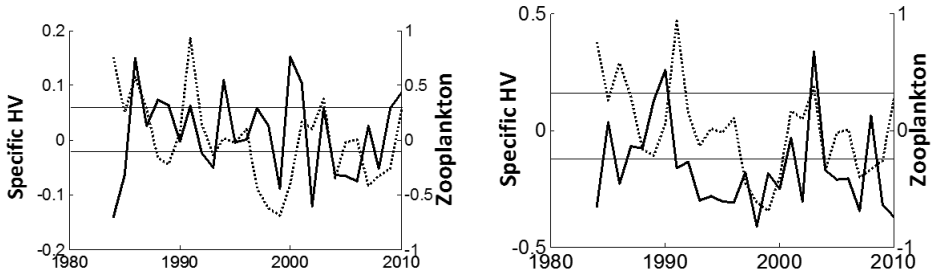
The *general HV* for **areas 1** (figure not shown) captured some of the species biomass

characteristics and it successfully managed to reflect on a temporal decline (trend identified, $p<.05$) in the series, coinciding with decline in the *SP* biomass for **area 1** ($Z=$ 1.59, d.f.=26, $p=0.11$). For **area 2** the HV was complex and much less explicit (absence of statistical trend, figure not shown). The *specific HV* from these areas is capturing the zooplankton dynamics with high similarity: **area 1** ($Z=0.17$, d.f.=26, $p=0.86$, Fig.A.6a) and **area 2** ($Z=$-0.88, d.f.=26, $p=0.38$). The zooplankton is characterised by a distinct decline until late 1990s which was captured by the *specific HV* from **area 2** (trend identified, $p<.05$), whilst no statistical trend was identified for **areas 1**.

Results from the HSDBN approach for **area 7** failed to identify any statistical similarity or trend by the *general HV* however the *specific HV* (Fig.A.6b) performed more accurately in terms of reflecting on the zooplankton biomass ($Z=0.02$, d.f.=26, $p=0.99$). It captured the declining trend ($p<0.05$) in time with the lowest values found around late 1990s.

Following the SDBN (failed to identify any statistical similarity or trend by the *general HV*), it was the ARHMM that produced most accurate biomass predictions for **areas 7** and the HV generated by this model was characterised by a temporal decline (trend identified, $p<.05$), coinciding with decline in the *P* biomass for **area 7** ($Z=0.56$, d.f.=26, $p=0.73$).

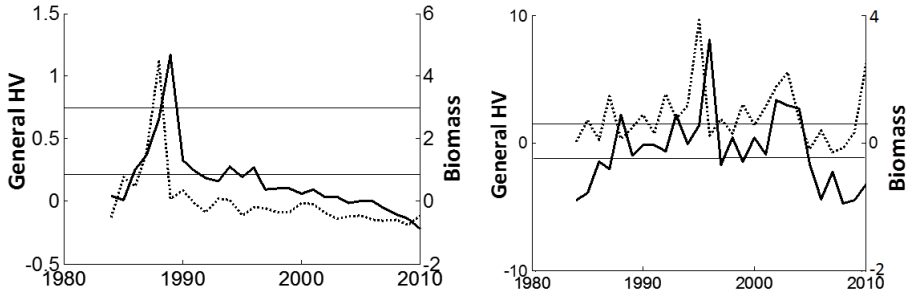
For **areas 3** and **4**, HSDBN performed well in terms of the *specific HV* learning the zooplankton dynamics (**area 3**: $Z=1.40$, d.f.=26, $p=0.16$ and **area 4**: $Z=0.10$, d.f.=26, $p=0.92$) whilst the *general HV* was much less explicit and clear, failing to identify statistical similarity with the observed trophic biomass.



(a) Area 1 (b) Area 7

Figure A.6: The expected value of the specific HV (solid line) for areas 1 (a) and 7 (b) generated by HSDBN with the observed standardised biomass for the zooplankton (dotted line). The solid line indicates upper and lower 95% confidence intervals, obtained from bootstrap predictions' overall mean and standard deviation. Note the negative scale is due to standardisation.

Results from the comparative evaluation of biomass predictions indicated that species groups biomass is predicted with relatively similar accuracy for **areas 3** and **4** by the DBN. The *general HV* generated by the DBN for **area 3** (Fig.A.7a) managed to capture some of the temporal dynamics of the area (trend identified, $p<.05$) and specifically, the HV identified similarity with the declining pattern in time of the P (Z=0.61, d.f.=26, p=0.54) and LP biomass (Z=1.12, d.f.=26, p=0.26). For **area 4**, the *general HV* from DBN, Fig.A.7b (absence of statistical trend) was characterised by initial temporal increase until late 1990s which coincides with increase in the P biomass (Z=-0.69, d.f.=26, p=0.49), followed by temporal decline in more recent years observed for both P and LP biomass (Z=-0.73, d.f.=26, p=0.47).



(a) Area 3 (b) Area 4

Figure A.7: The expected value of the general HV (solid line) for area 3 (a) and area 4 (b) generated by DBN with the observed standardised biomass for the P trophic group (dotted line). The solid line indicates upper and lower 95% confidence intervals, obtained from bootstrap predictions' overall mean and standard deviation. Note the negative scale is due to standardisation.

A.2 Chapter 6 Additional Results

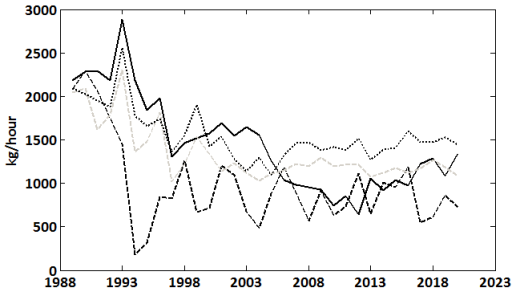
Table A.1: Table that shows the influence of catch on different species groups and the relevant species of interest from that group in each area.

Area	Fisheries Catch	Species of Interest
1	P	Herring
2	P	Herring
3	P, SP	Herring, Plaice, Sole, Whiting
4	LP	Cod, Haddock, Saithe
5	None	None
6	SP, LP	Plaice, Sole, Whiting, Cod, Haddock and Saithe
7	P, SP and LP	Herring, Plaice, Sole, Whiting, Cod, Haddock and Saithe

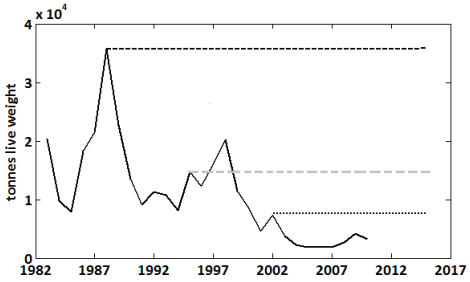
Table A.2: Table that shows the spatial neighbourhood of each area.

Area	Spatial Neighbourhood
1	2, 3 and 7
2	1, 3, 4, 6 and 7
3	1, 2 and 4
4	2, 3, 5 and 6
5	2, 4 and 6
6	2, 4, 5 and 7
7	1, 2 and 6

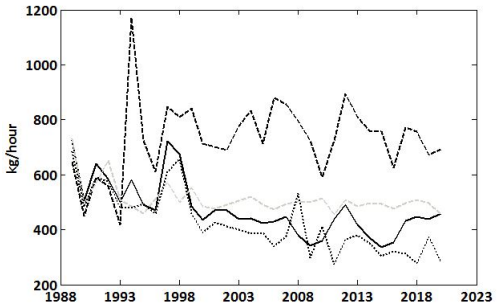
We show the modelled survey data for different fish species from the fisheries catch scenarios explained in Chapter 6.



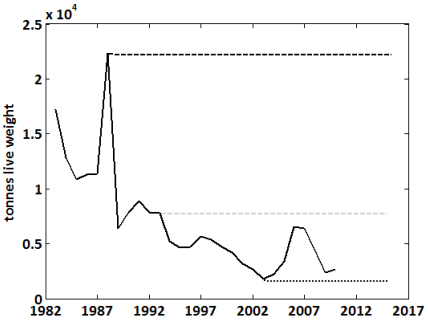
(a) Cod, area 6



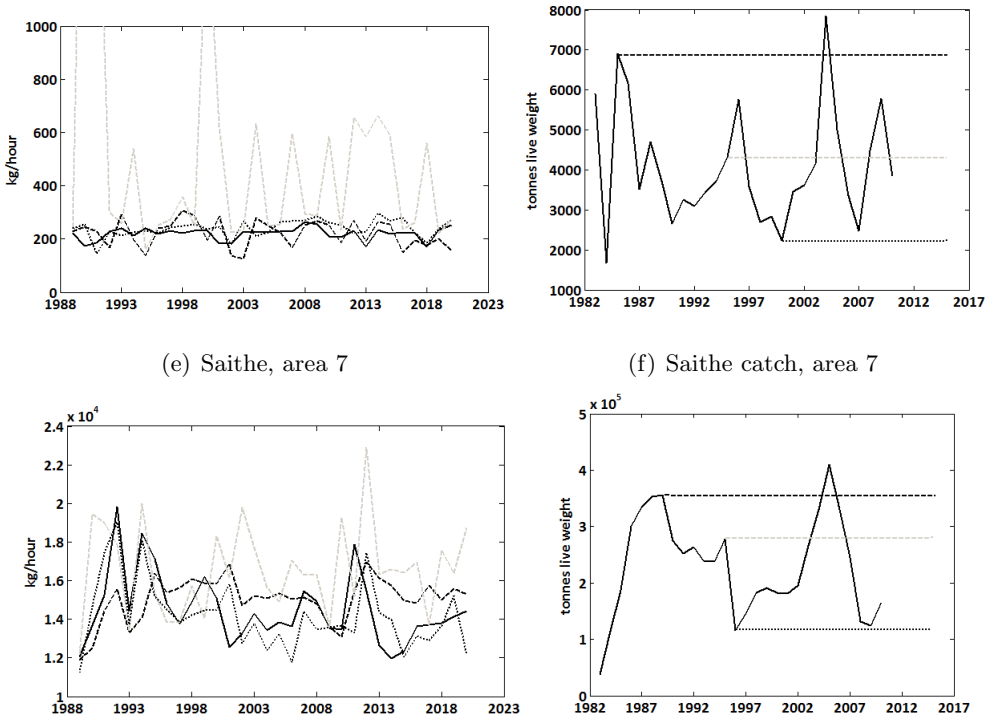
(b) Cod catch, area 6



(c) Whiting, area 7



(d) Whiting catch, area 7



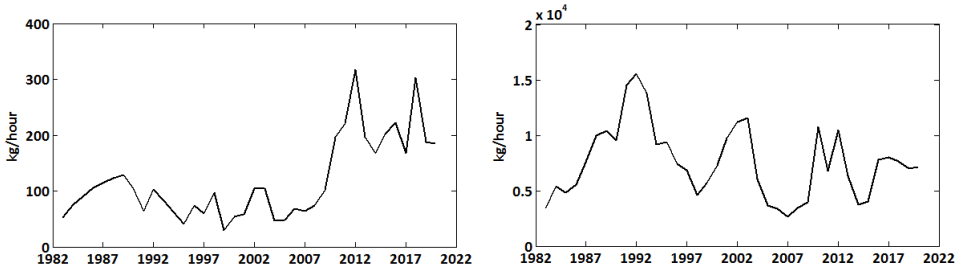
(e) Saithe, area 7

(f) Saithe catch, area 7

(g) Herring, area 1

(h) Herring catch, area 1

Figure A.8: The modelled survey data, generated by the *Historical* model (solid line) and scenarios of *H.FC.*(black dashed line), *M.FC.* (grey dotted line) and *L.FC.* (black dotted line) for the time window: 1989-2020. The observed catch (tonnes live weight is shown next to the modelled predictions.) Note, the y-axis in some of the figures was cut off only for visual purposes.



(a) Plaice spatial node, area 3

(b) Whiting spatial node, area 3

Figure A.9: The spatial values for plaice (a) and whiting (b) in area 3.

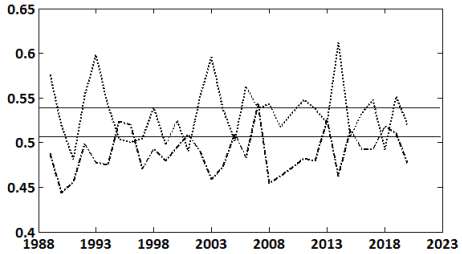


Figure A.10: The expected value of the hidden variable for area 3 generated by the *Historical* model for whiting (dashed line) and from the scenario of Low Fisheries Catch (dotted line). The solid line indicates upper and lower 95% confidence intervals, obtained from bootstrap predictions' overall mean and standard deviation.

References

Aderhold, A., Husmeier, D., Lennon, J. J., Beale, C. M. & Smith, V. A. (2012), 'Hierarchical bayesian models in ecology: Reconstructing species interaction networks from non-homogeneous species abundance data', *Ecological Informatics* **11**, 55–64.

Akaike, H. (1974), 'A new look at the statistical model identification', *Automatic Control, IEEE Transactions on* **19**(6), 716–723.

Al-ani, T. (2011), *Hidden Markov models in dynamic system modelling and diagnosis*, INTECH Open Access Publisher.

Andrewartha, H. & Browning, T. (1961), 'An analysis of the idea of resources in animal ecology', *Journal of theoretical biology* **1**(1), 83–97.

Astrom, K. (1965), 'Optimal control of markov processes with incomplete state information', *Journal of Mathematical Analysis and Applications* **10**(1), 174–205.

Barange, M., Merino, G., Blanchard, J., Scholtens, J., Harle, J., Allison, E., Allen, J., Holt, J. & Jennings, S. (2014), 'Impacts of climate change on marine ecosystem production in societies dependent on fisheries', *Nature Climate Change* **4**(3), 211–216.

Beaugrand, G., Brander, K. M., Lindley, J. A., Souissi, S. & Reid, P. C. (2003), 'Plankton effect on cod recruitment in the north sea', *Nature* **426**(6967), 661–664.

Begon, M., Harper, J. & Townsend, C. (1990), 'Life-history variation', *Ecology: Individuals, populations, and communities. 2nd edn. Boston, Massachusetts: Blackwell Scientific Publications* pp. 473–509.

Bengio, Y. & Frasconi, P. (1995), 'An input output hmm architecture', *Advances in neural information processing systems* pp. 427–434.

Berthold, M. R., Borgelt, C., Höppner, F. & Klawonn, F. (2010), *Guide to intelligent data analysis: how to intelligently make sense of real data*, Springer Science & Business Media.

Binder, J., Koller, D., Russell, S. & Kanazawa, K. (1997), 'Adaptive probabilistic networks with hidden variables', *Machine Learning* **29**(2-3), 213–244.

Bishop, C. (2006), *Pattern Recognition and Machine Learning*, Springer, New York.

Blanchard, J. L., Coll, M., Trenkel, V. M., Vergnon, R., Yemane, D., Jouffre, D., Link, J. S. & Shin, Y.-J. (2010), 'Trend analysis of indicators: a comparison of recent changes in the status of marine ecosystems around the world', *ICES Journal of Marine Science: Journal du Conseil* **67**(4), 732–744.

Blanchard, J. L., Jennings, S., Holmes, R., Harle, J., Merino, G., Allen, J. I., Holt, J., Dulvy, N. K. & Barange, M. (2012), 'Potential consequences of climate change for primary production and fish production in large marine ecosystems', *Philosophical Transactions of the Royal Society of London B: Biological Sciences* **367**(1605), 2979–2989.

Board, O. S. et al. (2006), *Dynamic Changes in Marine Ecosystems:: Fishing, Food Webs, and Future Options*, National Academies Press.

Bouckaert, G. (1995), *Improving performance measurement*, San Francisco: Jossey-Bass Publishers.

Carpenter, S. R., Brock, W. A., Cole, J. J. & Pace, M. L. (2014), 'A new approach for rapid detection of nearby thresholds in ecosystem time series', *Oikos* **123**(3), 290–297.

Chen, Q., Han, R., Ye, F. & Li, W. (2011), 'Spatio-temporal ecological models', *Ecological Informatics* **6**(1), 37–43.

Chen, S. H. & Pollino, C. A. (2012), 'Good practice in bayesian network modelling', *Environmental Modelling & Software* **37**, 134–145.

Cheung, Y. K. & Klotz, J. H. (1997), 'The mann whitney wilcoxon distribution using linked lists', *Statistica Sinica* **7**(3), 805–813.

Cooke, J. (1999), 'Improvement of fishery-management advice through simulation testing of harvest algorithms', *ICES Journal of Marine Science: Journal du Conseil* **56**(6), 797–810.

Cooper, G. F. & Herskovits, E. (1992), 'A bayesian method for the induction of probabilistic networks from data', *Machine learning* **9**(4), 309–347.

Cressie, N. & Wikle, C. K. (2015), *Statistics for spatio-temporal data*, John Wiley & Sons.

Cury, P. M., Mullon, C., Garcia, S. M. & Shannon, L. J. (2005), 'Viability theory for an ecosystem approach to fisheries', *ICES Journal of Marine Science: Journal du Conseil* **62**(3), 577–584.

Daan, N., Rijnsdorp, A. & Van Overbeeke, G. (1985), 'Predation by north sea herring clupea harengus on eggs of plaice pleuronectes platessa and cod gadus morhua', *Transactions of the American Fisheries Society* **114**(4), 499–506.

Dakos, V., Van Nes, E. H., D'Odorico, P. & Scheffer, M. (2012), 'Robustness of variance and autocorrelation as indicators of critical slowing down', *Ecology* **93**(2), 264–271.

de Santa Olalla, F. M., Dominguez, A., Ortega, F., Artigao, A. & Fabeiro, C. (2007), 'Bayesian networks in planning a large aquifer in eastern mancha, spain', *Environmental Modelling & Software* **22**(8), 1089–1100.

Dean, T. & Kanazawa, K. (1989), 'A model for reasoning about persistence and causation', *Computational intelligence* **5**(2), 142–150.

Dempster, A. P., Laird, N. M. & Rubin, D. B. (1977), 'Maximum likelihood from incomplete data via the em algorithm', *Journal of the royal statistical society. Series B (methodological)* pp. 1–38.

Doubleday, W. (1981), Manual on groundfish surveys in the Northwest Atlantic, Technical report, NAFO.

Dunne, J. A., Williams, R. J. & Martinez, N. D. (2002), 'Network structure and biodiversity loss in food webs: robustness increases with connectance', *Ecology letters* **5**(4), 558–567.

Duplisea, D. E. & Castonguay, M. (2006), 'Comparison and utility of different size-based metrics of fish communities for detecting fishery impacts', *Canadian Journal of Fisheries and Aquatic Sciences* **63**(4), 810–820.

Eddy, S. R. (1996), 'Hidden markov models', *Current opinion in structural biology* **6**(3), 361–365.

Efron, B. & Tibshirani, R. (1995), *Cross-validation and the bootstrap: Estimating the error rate of a prediction rule*, Division of Biostatistics, Stanford University.

Engelen, G., White, R. & Uljee, I. (1997), 'Integrating constrained cellular automata models, gis and decision support tools for urban planning and policy making', *Decision support systems in urban planning* pp. 125–155.

Engelhard, G. H., Pinnegar, J. K., Kell, L. T. & Rijnsdorp, A. D. (2011), 'Nine decades of north sea sole and plaice distribution', *ICES Journal of Marine Science: Journal du Conseil* **68**(6), 1090–1104.

Engelhard, G. H., Righton, D. A. & Pinnegar, J. K. (2014), 'Climate change and fishing: a century of shifting distribution in north sea cod', *Global change biology* **20**(8), 2473–2483.

Faisal, A., Dondelinger, F., Husmeier, D. & Beale, C. M. (2010), 'Inferring species interaction networks from species abundance data: A comparative evaluation of various statistical and machine learning methods', *Ecological Informatics* **5**(6), 451–464.

Fernandes, J. A., Cheung, W. W., Jennings, S., Butenschön, M., Mora, L., Frölicher, T. L., Barange, M. & Grant, A. (2013), 'Modelling the effects of climate change on the distribution and production of marine fishes: accounting for trophic interactions in a dynamic bioclimate envelope model', *Global change biology* **19**(8), 2596–2607.

Fernandes, J. A., Irigoien, X., Goikoetxea, N., Lozano, J. A., Inza, I., Pérez, A. & Bode, A. (2010), 'Fish recruitment prediction, using robust supervised classification methods', *Ecological Modelling* **221**(2), 338–352.

Fogarty, M. J. & Murawski, S. A. (1998), 'Large-scale disturbance and the structure of marine systems: fishery impacts on georges bank', *Ecological Applications* **8**(sp1), S6–S22.

Franco, C. (2014), Modelling the dynamics of CaCO3 budgets in changing environments using a Bayesian Belief Network approach, PhD thesis, University of Essex.

Friedman, N., Goldszmidt, M. & Wyner, A. (1999), Data analysis with bayesian networks: A bootstrap approach, *in* 'Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence', Morgan Kaufmann Publishers Inc., pp. 196–205.

Friedman, N., Linial, M., Nachman, I. & Pe'er, D. (2000), 'Using bayesian networks to analyze expression data', *Journal of computational biology* **7**(3-4), 601–620.

Friedman, N. et al. (1997), Learning belief networks in the presence of missing values and hidden variables, *in* 'Proceedings of the Fourteenth International Conference on Machine Learning (ICML 1997)', Vol. 97, pp. 125–133.

Frisk, M. G., Duplisea, D. E. & Trenkel, V. M. (2011), 'Exploring the abundance-occupancy relationships for the georges bank finfish and shellfish community from 1963 to 2006', *Ecological Applications* **21**(1), 227–240.

García-Carreras, B., Dolder, P., Engelhard, G. H., Lynam, C. P., Bayliss-Brown, G. A. & Mackinson, S. (2015), 'Recent experience with effort management in europe: Implications for mixed fisheries', *Fisheries Research* **169**, 52–59.

Gaston, K. J., Blackburn, T. M., Greenwood, J. J., Gregory, R. D., Quinn, R. M. & Lawton, J. H. (2000), 'Abundance–occupancy relationships', *Journal of Applied Ecology* **37**(s1), 39–59.

Ghahramani, Z. & Jordan, M. I. (1997), 'Factorial hidden markov models', *Machine learning* **29**(2-3), 245–273.

Gröger, J. P. & Fogarty, M. J. (2011), 'Broad-scale climate influences on cod (gadus morhua) recruitment on georges bank', *ICES Journal of Marine Science: Journal du Conseil* p. fsq196.

Gröger, J. P., Missong, M. & Rountree, R. A. (2011), 'Analyses of interventions and structural breaks in marine and fisheries time series: Detection of shifts using iterative methods', *Ecological Indicators* **11**(5), 1084–1092.

Haapasaari, P. & Karjalainen, T. P. (2010), 'Formalizing expert knowledge to compare alternative management plans: sociological perspective to the future management of baltic salmon stocks', *Marine Policy* **34**(3), 477–486.

Haapasaari, P., Michielsens, C., Karjalainen, T. P., Reinikainen, K. & Kuikka, S. (2007), 'Management measures and fishers' commitment to sustainable exploitation: a case study of atlantic salmon fisheries in the baltic sea', *ICES Journal of Marine Science: Journal du Conseil* **64**(4), 825–833.

Haeckel, E. (1869), 'Remarks on the protoplasm theory', *The Quarterly Journal of Microscopical Science* **10**, 223–229.

Haines-Young, R. (2011), 'Exploring ecosystem service issues across diverse knowledge domains using bayesian belief networks', *Progress in Physical Geography* **35**(5), 681–699.

Halpern, B. S., Walbridge, S., Selkoe, K. A., Kappel, C. V., Micheli, F., D'Agrosa, C., Bruno, J. F., Casey, K. S., Ebert, C., Fox, H. E. et al. (2008), 'A global map of human impact on marine ecosystems', *Science* **319**(5865), 948–952.

Hamilton, S. H., Pollino, C. A. & Jakeman, A. J. (2015), 'Habitat suitability modelling of rare species using bayesian networks: Model evaluation under limited data', *Ecological Modelling* **299**, 64–78.

Hammond, T. & O'Brien, C. (2001), 'An application of the bayesian approach to stock assessment model uncertainty', *ICES Journal of Marine Science: Journal du Conseil* **58**(3), 648–656.

Hartigan, J. A. & Wong, M. A. (1979), 'Algorithm as 136: A k-means clustering algorithm', *Applied statistics* pp. 100–108.

Heckerman, D., Geiger, D. & Chickering, D. M. (1995), 'Learning bayesian networks: The combination of knowledge and statistical data', *Machine learning* **20**(3), 197–243.

Hirsch, H.-G. (2001), 'Hmm adaptation for applications in telecommunication', *Speech Communication* **34**(1), 127–139.

Horwood, J., O'Brien, C. & Darby, C. (2006), 'North sea cod recovery?', *ICES Journal of Marine Science: Journal du Conseil* **63**(6), 961–968.

Inman, D., Blind, M., Ribarova, I., Krause, A., Roosenschoon, O., Kassahun, A., Scholten, H., Arampatzis, G., Abrami, G., McIntosh, B. et al. (2011), 'Perceived effectiveness of environmental decision support systems in participatory planning: Evidence from small groups of end-users', *Environmental Modelling & Software* **26**(3), 302–309.

Inza, I., Larrañaga, P., Blanco, R. & Cerrolaza, A. J. (2004), 'Filter versus wrapper gene selection approaches in dna microarray domains', *Artificial intelligence in medicine* **31**(2), 91–103.

Jennings, S., Alvsvåg, J., Cotter, A., Ehrich, S., Greenstreet, S., Jarre-Teichmann, A., Mergardt, N., Rijnsdorp, A. & Smedstad, O. (1999), 'Fishing effects in northeast atlantic shelf seas: patterns in fishing effort, diversity and community structure. iii. international trawling effort in the north sea: an analysis of spatial and temporal trends', *Fisheries Research* **40**(2), 125–134.

Jennings, S., Greenstreet, S., Hill, L., Piet, G., Pinnegar, J. & Warr, K. (2002), 'Long-term trends in the trophic structure of the north sea fish community: evidence from stable-isotope analysis, size-spectra and community metrics', *Marine Biology* **141**(6), 1085–1097.

Jennings, S. & Kaiser, M. J. (1998), 'The effects of fishing on marine ecosystems', *Advances in marine biology* **34**, 201–352.

Jennings, S., Pinnegar, J. K., Polunin, N. V. & Boon, T. W. (2001), 'Weak cross-species relationships between body size and trophic level belie powerful size-based trophic structuring in fish communities', *Journal of Animal Ecology* **70**(6), 934–944.

Jensen, F. V. (2001), 'Bayesian networks and decision graphs. statistics for engineering and information science', *Springer* **32**, 34.

Jiao, Y. (2009), 'Regime shift in marine ecosystems and implications for fisheries management, a review', *Reviews in Fish Biology and Fisheries* **19**(2), 177–191.

Juang, B. H. & Rabiner, L. R. (1991), 'Hidden markov models for speech recognition', *Technometrics* **33**(3), 251–272.

Kane, J. (2007), 'Zooplankton abundance trends on georges bank, 1977–2004', *ICES Journal of Marine Science: Journal du Conseil* **64**(5), 909–919.

Kenny, A. J., Skjoldal, H. R., Engelhard, G. H., Kershaw, P. J. & Reid, J. B. (2009), 'An integrated approach for assessing the relative significance of human pressures and environmental forcing on the status of large marine ecosystems', *Progress in Oceanography* **81**(1), 132–148.

Kerby, T. K., Cheung, W. W., van Oosterhout, C. & Engelhard, G. H. (2013), 'Wondering about wandering whiting: distribution of north sea whiting between the 1920s and 2000s', *Fisheries Research* **145**, 54–65.

Kirkpatrick, S., Jr., D. G. & Vecchi, M. P. (1983), 'Optimization by simulated annealing', *science* **220**(4598), 671–680.

Kleisner, K. M., Coll, M., Lynam, C. P., Bundy, A., Shannon, L., Shin, Y.-J., Boldt, J. L., Diallo, I., Fox, C., Gascuel, D. et al. (2015), 'Evaluating changes in marine communities that provide ecosystem services through comparative assessments of community indicators', *Ecosystem Services* **16**, 413–429.

Koller, D. & Friedman, N. (2009), *Probabilistic graphical models: principles and techniques*, MIT press.

Krebs, C. J. C. J. (1994), *Ecologythe experimental analysis of distribution and abundance*, number 574.5028 K74/1994.

Krivtsov, V. (2004), 'Investigations of indirect relationships in ecology and environmental sciences: a review and the implications for comparative theoretical ecosystem analysis', *Ecological Modelling* **174**(1), 37–54.

Krogh, A., Brown, M., Mian, I. S., Sjölander, K. & Haussler, D. (1994), 'Hidden markov models in computational biology: Applications to protein modeling', *Journal of molecular biology* **235**(5), 1501–1531.

Krueger, T., Page, T., Hubacek, K., Smith, L. & Hiscock, K. (2012), 'The role of expert opinion in environmental modelling', *Environmental Modelling & Software* **36**, 4–18.

Kuikka, S., Hildén, M., Gislason, H., Hansson, S., Sparholt, H. & Varis, O. (1999), 'Modeling environmentally driven uncertainties in baltic cod (gadus morhua) management by bayesian influence diagrams', *Canadian Journal of Fisheries and Aquatic Sciences* **56**(4), 629–641.

Kuikka, S. & Varis, O. (1997), 'Uncertainties of climatic change impacts in finnish watersheds: a bayesian network analysis of expert knowledge', *Boreal Environment Research* **2**(1), 109–128.

Kwoh, C.-K. & Gillies, D. F. (1996), 'Using hidden nodes in bayesian networks', *Artificial intelligence* **88**(1), 1–38.

Larranaga, P., Sierra, B., Gallego, M. J., Michelena, M. J. & Picaza, J. M. (1997), Learning bayesian networks by genetic algorithms: a case study in the prediction of survival in malignant skin melanoma, *in* 'Artificial Intelligence in Medicine', Springer, pp. 261–272.

Lee, D. C. & Rieman, B. E. (1997), 'Population viability assessment of salmonids by using probabilistic networks', *North American Journal of Fisheries Management* **17**(4), 1144–1157.

Lenhart, H.-J., Mills, D. K., Baretta-Bekker, H., Van Leeuwen, S. M., der Molen, J. v., Baretta, J. W., Blaas, M., Desmit, X., Kühn, W., Lacroix, G. et al. (2010), 'Predicting the consequences of nutrient reduction on the eutrophication status of the north sea', *Journal of Marine Systems* **81**(1), 148–170.

Levontin, P., Kulmala, S., Haapasaari, P. & Kuikka, S. (2011), 'Integration of biological, economic, and sociological knowledge by bayesian belief networks: the interdisciplinary evaluation of potential management plans for baltic salmon', *ICES Journal of Marine Science: Journal du Conseil* **68**(3), 632–638.

Lewy, P. & Vinther, M. (2004), 'A stochastic age-length-structured multispecies model applied to north sea stocks', *ICES CM* p. 33.

Liang, S., Fuhrman, S., Somogyi, R. et al. (1998), Reveal, a general reverse engineering algorithm for inference of genetic network architectures, *in* 'Pacific symposium on biocomputing', Vol. 3, p. 2.

Lynam, C. P. & Mackinson, S. (2015), 'How will fisheries management measures contribute towards the attainment of good environmental status for the north sea ecosystem?', *Global Ecology and Conservation* **4**, 160–175.

Lynam, T., Drewry, J., Higham, W. & Mitchell, C. (2010), 'Adaptive modelling for adaptive water quality management in the great barrier reef region, australia', *Environmental Modelling & Software* **25**(11), 1291–1301.

Mackinson, S. & Daskalov, G. (2007), 'An ecosystem model of the north sea to support an ecosystem approach to fisheries management: description and parameterisation', *Cefas Science Series Technical Report* **142**, 195.

Marcot, B. G., Holthausen, R. S., Raphael, M. G., Rowland, M. M. & Wisdom, M. J. (2001), 'Using bayesian belief networks to evaluate fish and wildlife population viability under land management alternatives from an environmental impact statement', *Forest ecology and management* **153**(1), 29–42.

McCann, R. K., Marcot, B. G. & Ellis, R. (2006), 'Bayesian belief networks: applications

in ecology and natural resource management', *Canadian Journal of Forest Research* **36**(12), 3053–3062.

McCulloch, C. E. & Neuhaus, J. M. (2001), *Generalized linear mixed models*, Wiley Online Library.

McGrory, C. A. & Titterington, D. (2009), 'Variational bayesian analysis for hidden markov models', *Australian & New Zealand Journal of Statistics* **51**(2), 227–244.

Mihajlovic, V. & Petkovic, M. (2001), 'Dynamic bayesian networks: A state of the art'.

Milns, I., Beale, C. M. & Smith, V. A. (2010), 'Revealing ecological networks using bayesian network inference algorithms', *Ecology* **91**(7), 1892–1899.

Morgan, M. G. & Henrion, M. (1990), 'Uncertainty: a guide to dealing with uncertainty in quantitative risk and policy analysis cambridge university press', *New York, New York, USA* .

Murphy, K. (1998), 'A brief introduction to graphical models and bayesian networks'.

Murphy, K. (2001), An introduction to graphical models, Technical report.

Murphy, K., Mian, S. et al. (1999), Modelling gene expression data using dynamic bayesian networks, Technical report, Technical report, Computer Science Division, University of California, Berkeley, CA.

Murphy, K. P. (2002), Dynamic bayesian networks: representation, inference and learning, PhD thesis, University of California, Berkeley.

Murphy, K. et al. (2001), 'The bayes net toolbox for matlab', *Computing science and statistics* **33**(2), 1024–1034.

Nielsen, T. D. & Jensen, F. V. (2009), *Bayesian networks and decision graphs*, Springer Science & Business Media.

Olff, H., Alonso, D., Berg, M. P., Eriksson, B. K., Loreau, M., Piersma, T. & Rooney, N. (2009), 'Parallel ecological networks in ecosystems', *Philosophical Transactions of the Royal Society of London B: Biological Sciences* **364**(1524), 1755–1779.

Pavlović, V., Rehg, J. M., Cham, T.-J. & Murphy, K. P. (1999), A dynamic bayesian network approach to figure tracking using learned dynamic models, *in* 'Computer Vision, 1999. The Proceedings of the Seventh IEEE International Conference on', Vol. 1, IEEE, pp. 94–101.

Pearl, J. (1988), *Probabilistic reasoning in intelligent systems: Networks of plausible inference*, San Mateo, CA: Morgan Kaufmann.

Pede, V. O., Florax, R. J. & Lambert, D. M. (2014), 'Spatial econometric star models: Lagrange multiplier tests, monte carlo simulations and an empirical application', *Regional Science and Urban Economics* **49**, 118–128.

Perry, A. L., Low, P. J., Ellis, J. R. & Reynolds, J. D. (2005), 'Climate change and distribution shifts in marine fishes', *science* **308**(5730), 1912–1915.

Planque, B., Fromentin, J.-M., Cury, P., Drinkwater, K. F., Jennings, S., Perry, R. I. & Kifani, S. (2010), 'How does fishing alter marine populations and ecosystems sensitivity to climate?', *Journal of Marine Systems* **79**(3), 403–417.

Rabiner, L. R. (1989), 'A tutorial on hidden markov models and selected applications in speech recognition', *Proceedings of the IEEE* **77**(2), 257–286.

Regulation, E. (2013), 'No. 1380/2013 of the european parliament and of the council of 11 december 2013 on the common fisheries policy, amending council regulations (ec) no 1954/2003 and (ec) no 1224/2009 and repealing council regulations (ec) no 2371/2002 and (ec) no 639/2004 and council decision 2004/585/ec', *Official Journal of the European Union. L* **354**, 22–61.

Richards, R., Sanó, M., Roiko, A., Carter, R., Bussey, M., Matthews, J. & Smith, T. F. (2013), 'Bayesian belief modeling of climate change impacts for informing regional adaptation options', *Environmental modelling & software* **44**, 113–121.

Rothschild, B. J. & Shannon, L. (2004), 'Regime shifts and fishery management', *Progress in Oceanography* **60**(2), 397–402.

Russell, S., Binder, J., Koller, D. & Kanazawa, K. (1995), Local learning in probabilistic networks with hidden variables, *in* 'IJCAI', Vol. 95, Citeseer, pp. 1146–1152.

Scheffer, M., Carpenter, S., Foley, J. A., Folke, C. & Walker, B. (2001), 'Catastrophic shifts in ecosystems', *Nature* **413**(6856), 591–596.

Schwarz, G. et al. (1978), 'Estimating the dimension of a model', *The annals of statistics* **6**(2), 461–464.

Simpson, S. D., Jennings, S., Johnson, M. P., Blanchard, J. L., Schön, P.-J., Sims, D. W. & Genner, M. J. (2011), 'Continental shelf-wide response of a fish assemblage to rapid warming of the sea', *Current Biology* **21**(18), 1565–1570.

Smith, C. S., Howes, A. L., Price, B. & McAlpine, C. A. (2007), 'Using a bayesian belief network to predict suitable habitat of an endangered mammal–the julia creek dunnart (sminthopsis douglasi)', *Biological Conservation* **139**(3), 333–347.

Smith, V. A., Yu, J., Smulders, T. V., Hartemink, A. J. & Jarvis, E. D. (2006), 'Computational inference of neural information flow networks', *PLoS computational biology* **2**(11), e161.

Smyth, P. (1994), 'Hidden markov models for fault detection in dynamic systems', *Pattern recognition* **27**(1), 149–164.

Stanculescu, I., Williams, C. K. & Freer, Y. (2014), 'Autoregressive hidden markov models for the early detection of neonatal sepsis', *Biomedical and Health Informatics, IEEE Journal of* **18**(5), 1560–1570.

Tattari, S., Schultz, T. & Kuussaari, M. (2003), 'Use of belief network modelling to assess the impact of buffer zones on water protection and biodiversity', *Agriculture, ecosystems & environment* **96**(1), 119–132.

Ticehurst, J. L., Newham, L. T., Rissik, D., Letcher, R. A. & Jakeman, A. J. (2007), 'A bayesian network approach for assessing the sustainability of coastal lakes in new south wales, australia', *Environmental Modelling & Software* **22**(8), 1129–1139.

Trifonova, N., Kenny, A., Maxwell, D., Duplisea, D., Fernandes, J. & Tucker, A. (2015), 'Spatio-temporal bayesian network models with latent variables for revealing trophic dynamics and functional networks in fisheries ecology', *Ecological Informatics* **30**, 142–158.

Tucker, A. & Duplisea, D. (2012), 'Bioinformatics tools in predictive ecology: applications to fisheries', *Philosophical Transactions of the Royal Society B: Biological Sciences* **367**(1586), 279–290.

Tyrrell, M., Link, J. & Moustahfid, H. (2011), 'The importance of including predation in fish population models: implications for biological reference points', *Fisheries Research* **108**(1), 1–8.

Ulrich, C., Reeves, S. A., Vermard, Y., Holmes, S. J. & Vanhee, W. (2011), 'Reconciling single-species tacs in the north sea demersal fisheries using the fcube mixed-fisheries advice framework', *ICES Journal of Marine Science: Journal du Conseil* **68**(7), 1535–1547.

Uusitalo, L. (2007), 'Advantages and challenges of bayesian networks in environmental modelling', *Ecological modelling* **203**(3), 312–318.

Uusitalo, L., Kuikka, S., Kauppila, P., Söderkultalahti, P. & Bäck, S. (2012), 'Assessing the roles of environmental factors in coastal fish production in the northern baltic sea: A bayesian network application', *Integrated environmental assessment and management* **8**(3), 445–455.

Uusitalo, L., Kuikka, S. & Romakkaniemi, A. (2005), 'Estimation of atlantic salmon smolt carrying capacity of rivers using expert knowledge', *ICES Journal of Marine Science: Journal du Conseil* **62**(4), 708–722.

van Leeuwen, S. M., van der Molen, J., Ruardij, P., Fernand, L. & Jickells, T. (2013), 'Modelling the contribution of deep chlorophyll maxima to annual primary production in the north sea', *Biogeochemistry* **113**(1-3), 137–152.

Varis, O., Kettunen, J. & Sirviö, H. (1990), 'Bayesian influence diagram approach to

complex environmental management including observational design', *Computational Statistics & Data Analysis* **9**(1), 77–91.

Vaseghi, S. V. (2008), *Advanced digital signal processing and noise reduction*, John Wiley & Sons.

Vilizzi, L. & Copp, G. H. (2013), 'Application of fisk, an invasiveness screening tool for non-native freshwater fishes, in the murray-darling basin (southeastern australia)', *Risk Analysis* **33**(8), 1432–1440.

Vinther, M., Reeves, S. A. & Patterson, K. R. (2004), 'From single-species advice to mixed-species management: taking the next step', *ICES Journal of Marine Science: Journal du Conseil* **61**(8), 1398–1409.

Wikle, C. K. (2003), 'Hierarchical bayesian models for predicting the spread of ecological processes', *Ecology* **84**(6), 1382–1394.

Wikle, C. K., Berliner, L. M. & Cressie, N. (1998), 'Hierarchical bayesian space-time models', *Environmental and Ecological Statistics* **5**(2), 117–154.

Wooldridge, J. M. (2005), 'Simple solutions to the initial conditions problem in dynamic, nonlinear panel data models with unobserved heterogeneity', *Journal of applied econometrics* **20**(1), 39–54.

Wooldridge, S. & Done, T. (2004), 'Learning to predict large-scale coral bleaching from past events: A bayesian approach using remotely sensed data, in-situ data, and environmental proxies', *Coral Reefs* **23**(1), 96–108.

Xuan, T. (2004), Autoregressive hidden Markov model with application in an El Nino study, PhD thesis, University of Saskatchewan Saskatoon.