Incorporating Regime Metrics into Latent Variable Dynamic Models to Detect Early-Warning Signals of Functional Changes in Fisheries Ecology

Neda Trifonova[1], Daniel Duplisea[2], Andrew Kenny[3], David Maxwell[3], Allan Tucker[1]

[1] Department of Computer Science, Brunel University, London, UK[1]
[2] Fisheries and Oceans, Canada[2]
[3] Centre for Environment, Fisheries and Aquaculture Science, Lowestoft, UK[3]

Abstract. In this study, dynamic Bayesian networks have been applied to predict future biomass of geographically different but functionally equivalent fish species. A latent variable is incorporated to model functional collapse, where the underlying food web structure dramatically changes irrevocably (known as 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. The hidden variable appears to reflect some of the metrics' characteristics in terms of identifying regime shifts that are known to have occurred. It also appears to capture changes in the variance of different species biomass. Including metrics in the models had an impact on predictive accuracy but only in some cases. 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. A non-parametric bootstrap in combination with a greedy search algorithm was applied to estimate the confidence of features of networks learned from the data, allowing us to identify pairwise relations of high confidence between species.

1 Introduction

Some spectacular collapses in fish stocks have occurred in the past 20 years but the most notable is the once largest cod (*Gadus morhua*) stock in the world, the Northern cod stock off eastern Newfoundland, which experienced a 99% decline in biomass (the total quantity or weight of organisms in a given area or volume). Such regions have experienced a "regime shift" or moved to an "alternative stable state" and are unlikely to return to a cod dominated community without some influence beyond human control [9]. 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 regime shift) [3].

Regime (functional) changes can affect the abundance and distribution of fish populations, either directly or by affecting prey or predator populations [9]. 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 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.

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 latent variables. Formally, a Bayesian Network (BN) exploits the conditional independence relationships over a set of variables, represented by directed acyclic graphs (DAG) [6]. Modelling time series is achieved by the DBN, where nodes represent variables at particular time slices [6]. Closely related to the DBN is the Hidden Markov Model (HMM) which models the dynamics of a dataset through the use of a latent or hidden variable. This latent variable is used to infer some underlying state of the series and can be applied through an autoregressive link which can capture relationships of a higher order. Hidden variables can also be incorporated to model unobserved variables and missing data by using the EM algorithm [2]. 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 such complex interdependencies between the acting variables. See Fig. 1 for an illustration of the architecture of the DBN used in this paper including a hidden variable.

In this paper, we investigate the reliability of our modelling approach in detecting *early-warning* signals of functional change across *different geographic regions*. We explore how the latent 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 latent 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 [12] explores functionally equivalent species but here we further adopt the approach to predict functional collapse by investigating *early-warning signals* and comparing learned 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 (relating to the feeding habits of different organisms in a food

chain) dynamics [9]. A clear example is the Scotian Shelf, where fishing has led to a restructuring of the ecosystem [9]. 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.

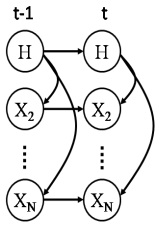


Fig. 1. The Dynamic Bayesian Model with N variables used in this study where H denotes the unmeasured hidden variable.

2 Methods

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. 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). 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. In addition to survey data on fish abundance, grey seal abundance and plankton time series were also included in the analysis.

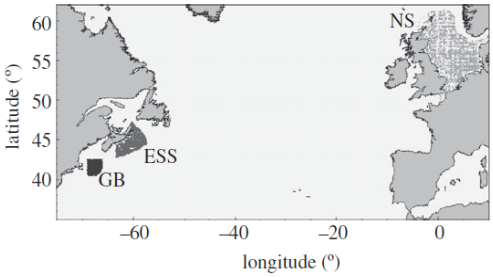


Fig. 2. Regions of the three surveys (shaded area) corresponding to the three datasets: Georges Bank (GB), East Scotian Shelf (ESS) and the North Sea (NS).

The experiments involve the prediction of a pre-selected variable (here functional collapse, represented by the latent variable) based on the values of other variables (here species biomass). 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 [4] is used to build the BN classifiers. A bootstrap approach is employed to repeat the following 1000 times: 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 [8] and scores each variable by taking into account their interaction with other variables through the use of a classifier model. Next, we identify the equivalent species in the other two datasets using the features discovered using Algorithm 1. The *functional equivalence* search algorithm [12] 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 [10] 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 [4]. In Algorithm 1 *UnifRand* represents a random value generated from uniform distribution with limits between (0,1).

After choosing the species, we want to predict their biomass and the functional collapse in the relevant geographic region. For example to predict functional collapse we compute $P(H^t|X^t, X^{t-1})$, where H^t represents the hidden variable (functional collapse) and X^t represents all observed variables at times t. First, we infer the biomass at time t (Fig.1) by using the observed evidence and then use the predicted variable states to infer the hidden state at time t. The hidden variable was parameterised using the EM algorithm.

The metrics: variance and autocorrelation 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 (HDBN) and in the other, metrics were included in the model (HDBN + metrics) to see if they improve prediction of species biomass. Non-parametric bootstrap analysis [6] was applied 250 times for each variant of the model to obtain statistical validation in the predictions. 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 expected collapse [7]. 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 [7]. All statistical tests were reported at 5% significance level.

For the next part of the study- learning the model structure, the species biomass data was discretised and a greedy search algorithm: REVEAL [11] was applied

to learn the structure of the DBN model for each region. The 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 [6].

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 $dscore \geq 0$ 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: $vars_2$

3 Results

The wrapper feature selection approach managed to identify the species likely to be associated with cod collapse on GB, Table 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 [1].
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 2, confidence reported in brackets). 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. 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. 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.

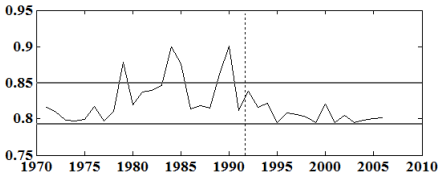
Table 1. Wrapper feature selection results for GB region.

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)

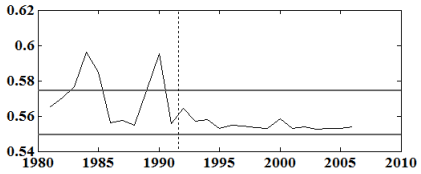
Table 2. The functionally equivalent species to GB dataset for ESS and NS. These are each ordered based upon their relevance to species in Table 1.

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)

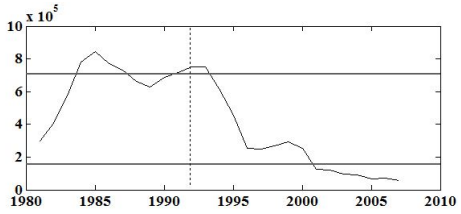
We now explore the latent variable models for ESS and NS learnt from the selected functionally equivalent species. We also focus on the relationship of the latent variable to the two regime metrics. The expected value of the hidden variable for ESS managed to capture some of the key predictive qualities of the metrics in terms of identifying a regime shift that is known to have occurred. The ESS latent variable model (HDBN) in Fig. 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.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). 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 [5]. According to theoretical expectations of critical slowing down, both latent 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 latent variable was more stable and still reflective of capturing the correct dynamics and characterised by rising trends in time prior to the expected transition (HDBN + metrics in Fig. 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 HDBN, suggesting that the latent variable in combination with the metrics might be performing better in earlier detection of change in the time series, though having a negative impact on the predictive performance of biomass (SSE HDBN: 4.83 and SSE HDBN+ metrics: 13.65) (Fig.4).



(a) Expected hidden value (HDBN)



(b) Expected hidden value (HDBN+ metrics)



(c) Mean variance

Fig. 3. The expected values of the discovered hidden variable from HDBN (a), HDBN+ 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.

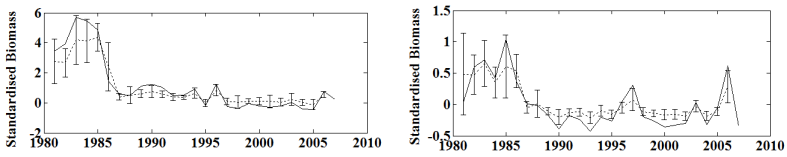
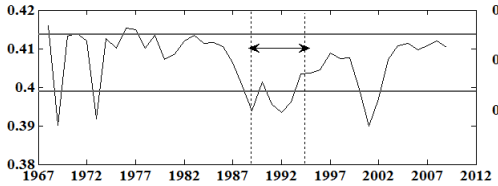


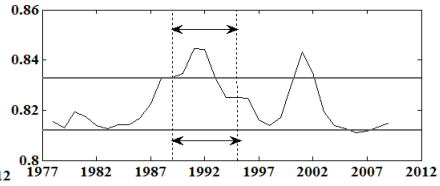
Fig. 4. Biomass predictions generated by HDBN+ metrics of cod (left), and silver hake (right) for ESS region. 95% confidence intervals report bootstrap predictions' mean and standard deviation. Dashed line indicates predictions by the model, whilst solid indicates standardised observed biomass for the time period 1980-2006.

The expected value of the hidden variable for NS (Fig.5a) was characterised by some fluctuation up to early 1980s followed by a small decrease below the lower confidence 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 latent variables in GB and ESS. Perhaps this is not surprising as it was found in [12] that the latent 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 latent variable and variance (Fig.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 [1]. 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 latent variable from the HDBN+ metrics (Fig.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 [3]. However, the effect of the metrics on the latent variable assisted in the correct identification of the time

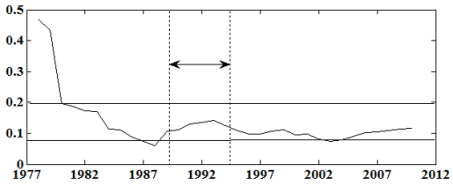
period where we would expect some functional change or disturbance in NS. Results for NS showed reliable prediction of species biomass, with improved ability of the dynamic models when used in combination with the published metrics (SSE model: 5.50, SSE model+ metrics: 2.82).



(a) Expected hidden value (HDBN)



(b) Expected hidden value (HDBN+metrics)



(c) Mean variance

Fig. 5. The expected value of the discovered hidden variable from HDBN (a), HDBN+ 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.

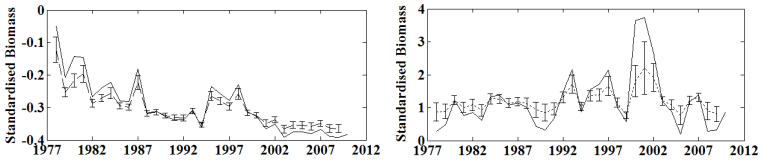
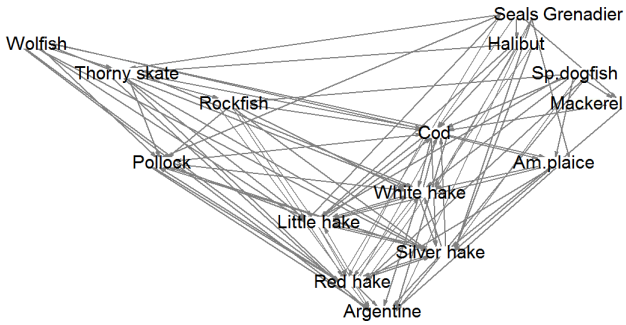


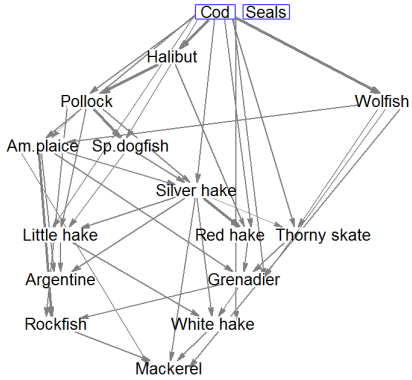
Fig. 6. Biomass predictions generated by HDBN+ metrics of cod (left), and haddock (right) for NS region. 95% confidence intervals report bootstrap predictions' mean and standard deviation. Dashed line indicates predictions by the model, whilst solid indicates standardised observed biomass for the time 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 latent variable alone managed to reflect the ecosystem dynamics but that was more evident in the ESS region with a larger regime shift.
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 latent variable models (timing identified from F-test significant results in the first part of the study). Some high

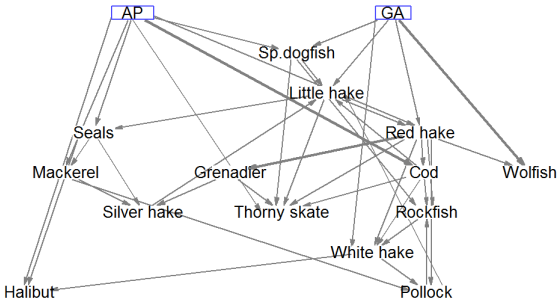
confidence relationships were identified which represent likely models of the functional interactions between species. The direction of the discovered significant links did not mean causation and it was not considered in the comparison with the generating diet matrix as we were interested in finding 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.7b) (as found by HDBN+metrics) identified 8 significant links (four true positives) and after (Fig.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.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.
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 HDBN+metrics) identified four significant links and after: one significant link, no true positives. The relative simplicity of the NS networks and much lower number of correctly identified associations 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. The NS diet matrix was also relatively "poor" compared to ESS in terms of quantity of species recorded.
To summarise, the bootstrap methodology of learning the model structure in combination with the data split from our latent 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



(b) Network before regime shift



(c) Network after regime shift

Fig. 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.

4 Conclusion

In this paper we have explored the use of regime metrics in conjunction with latent variables which proved useful (compared to these without) in terms of detecting *early-warning* signals (significantly rising variance and latent variable fluctuations) of functional changes but it seemed to have an impact on biomass prediction. The latent variables fitted to models that exclude these metrics appear to reflect some of their characteristics in terms of capturing the correct trophic dynamics. The learned network links managed to find some overlap with the diet matrices, though not many, maybe due to implicit correlations (and so more latent 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. Further work will involve informative priors based upon available expertise to create scenarios for environmental management purposes.

References

1. Beaugrand, G., Brander, K.M., Lindley, J.A., Souissi, S., Reid, P.C.: Plankton effect on cod recruitment in the north sea. Nature 426(6967), 661–664 (2003)
2. Bilmes, J.A., et al.: A gentle tutorial of the em algorithm and its application to parameter estimation for gaussian mixture and hidden markov models. International Computer Science Institute 4(510), 126 (1998)
3. Carpenter, S.R., Brock, W.A., Cole, J.J., Pace, M.L.: A new approach for rapid detection of nearby thresholds in ecosystem time series. Oikos (2013)
4. Cooper, G.F., Herskovits, E.: A bayesian method for the induction of probabilistic networks from data. Machine learning 9(4), 309–347 (1992)
5. Dakos, V., Van Nes, E.H., D'Odorico, P., Scheffer, M.: Robustness of variance and autocorrelation as indicators of critical slowing down. Ecology 93(2), 264–271 (2012)
6. Friedman, N., Goldszmidt, M., Wyner, A.: Data analysis with bayesian networks: A bootstrap approach. In: Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence. pp. 196–205. Morgan Kaufmann Publishers Inc. (1999)
7. Gröger, J.P., Missong, M., Rountree, R.A.: Analyses of interventions and structural breaks in marine and fisheries time series: Detection of shifts using iterative methods. Ecological Indicators 11(5), 1084–1092 (2011)
8. Inza, I., Larrañaga, P., Blanco, R., Cerrolaza, A.J.: Filter versus wrapper gene selection approaches in dna microarray domains. Artificial intelligence in medicine 31(2), 91–103 (2004)
9. Jiao, Y.: Regime shift in marine ecosystems and implications for fisheries management, a review. Reviews in Fish Biology and Fisheries 19(2), 177–191 (2009)
10. Kirkpatrick, S., Jr., D.G., Vecchi, M.P.: Optimization by simmulated annealing. science 220(4598), 671–680 (1983)
11. Liang, S., Fuhrman, S., Somogyi, R., et al.: Reveal, a general reverse engineering algorithm for inference of genetic network architectures. In: Pacific symposium on biocomputing. vol. 3, p. 2 (1998)
12. Tucker, A., Duplisea, D.: Bioinformatics tools in predictive ecology: applications to fisheries. Philosophical Transactions of the Royal Society B: Biological Sciences 367(1586), 279–290 (2012)