A Spatio-Temporal Bayesian Network Approach for Revealing Functional Ecological Networks in Fisheries

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

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

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. Machine learning techniques can allow such complex, spatially varying interactions to be recovered from collected field data. In this study, we apply structure learning techniques to identify functional relationships between trophic groups of species that vary across space and time. Specifically, Bayesian networks are created on a *window of data* for each of the 20 *geographically* different and *temporally* varied sub-regions within an oceanic area. In addition, we explored the *spatial* and *temporal* variation of *pre-defined functions* (like predation, competition) that are generalisable by experts' knowledge. We were able to discover meaningful ecological networks that were more precisely *spatially-specific* rather than temporally, as previously suggested for this region. To validate the discovered networks, we predict the biomass of the trophic groups by using dynamic Bayesian networks, and correcting for spatial autocorrelation by including a *spatial node* in our models.

1 Introduction

In recent decades it has become clear that ecosystem structure and function can change over relatively short time [13]. Functional changes can significantly affect the abundance and distribution of fish populations, either directly or by affecting prey or predator populations [11]. The effect of predators has been shown to influence prey populations and vice versa and has been described to be of the same or greater magnitude than fishing alone [11]. 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.
One way to understand community structure and stability is examination of the

functional relationships (such as prey-predator) between species 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 [8]. In this study, we aggregate individual species into trophic species (functional groups of taxa that share the same set of predators and prey within a food web): invertebrates, pelagics (pelagic fish that live in the pelagic zone of ocean waters - being neither close to the bottom nor near the shore), small piscivorous (fish-eating species) and large piscivorous and top predators from the northern Gulf of St. Lawrence groundfish and shrimp summer survey since 1990 and examined how the learned functional relationships between the trophic groups varied in time and space.

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 [4]. 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 to predict community stability. Recently, an approach has arisen in biology that is capable of inferring network structures, capturing nonlinear, stochastic and arbitrary combinatorial relationships: Bayesian Networks (BNs) [10]. Formally, a BN exploits the conditional independence relationships over a set of variables, represented by directed acyclic graphs (DAG) [6]. 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 into the network, apply inference and inspect the posterior distribution. Structure learning of these models from data is an NP-hard problem and many studies have been conducted on this subject, leading to three different approaches: constraint-based methods, score-based and hybrid methods [2]. We focus in this paper on BN structure learning using score-based method, specifically learning a distinct network for each sub-region of the Gulf of St. Lawrence oceanic area.

In this paper, we examine how aggregated species interact at different *spatial* scales and over *time* 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 and resilience. At larger spatial scales, although fishing can still be the dominant driver of functional changes, the consequences of fishing are not predictable without understanding the food web dynamics [11].

2 Methods

2.1 Species Collection

We analysed data from the northern Gulf of St. Lawrence (48.00°N, 61.50°W, Fig.1a) groundfish and shrimp summer survey (1990-2013). The survey utilises a

stratified random sampling design [3] with a standard tow using a benthic otter trawl. For each tow, all the fish were weighed and a subsample (200 individuals per species) was taken for computing length-frequency distributions. These length-frequency distributions were the basis of the data used here.

2.2 Data Preparation

K-means [9] was applied to limit the number of variables and cluster the number of sampling stations (originally over 200 sampling stations per year, Fig.1b) on the mean latitude and longitude, resulting in 20 spatial clusters (or subregions, Fig.1c). Note that differences in density of the clustered stations could explain the slight spatial contrast between Fig.1b and Fig.1c. The number of stations varied within each cluster so the biomass (the total quantity or weight of organisms in a given area or volume) was averaged 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. The nature of individual species summed into the trophic guilds varied between the spatial clusters but this was not of importance since they were always aggregated into the correct trophic group. This was performed for each of the 20 clusters and for each year in the time window: 1990-2013, ending up with four variables for each spatial cluster across the time window.
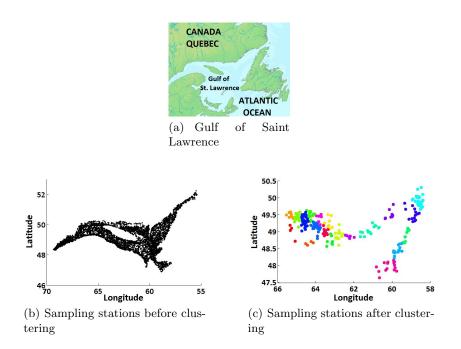

(a) Gulf of Saint Lawrence

(b) Sampling stations before clustering

(c) Sampling stations after clustering

Fig. 1: Locations of the oceanic region of St. Lawrence (a) and the sampling stations before clustering (b) and after clustering (c).

2.3 Structure Learning of BNs

Our model is a BN in which nodes represent trophic groups and edges (connections between nodes) represent potential species interactions. Note we infer static BNs from temporal data for each of the 20 spatial clusters. Hill-climbing procedure was applied for learning the static BN structure. The search 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* or *delete arc* and choose the one change that improves the score the most. We used the Bayesian Information Criterion (BIC) for scoring candidate networks [14]. The BIC function is a combination of the model log-likelihood and a penalty term that favours less complex models- as such it is similar to the minimum description length: $BIC = \log P(\Theta) + \log P(\Theta|D)$ - 0.5 k $\log(n)$ 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.

The hill-climb structure learning approach was conducted with 10 random restarts. In this approach, we apply the search until we hit a local maximum. Then, we randomly perturb the network structure and repeat the process for some number of iterations, in the case of the network analysis for individual clusters alone (20 clusters, each matrix with the size of *4*x*24*), we apply the learning procedure for 500 iterations. In addition, to learn the model 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.

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 [1]. To incorporate potential spatial autocorrelation in our model, we connect each node in the network to an enforced parent node that represents the average biomass from the spatial neighbourhood (the four nearest neighbours) of the current geographic location (or cluster) [1]. By applying the windowing approach, we produced two variants of our BN model: one that excludes a spatial node and one including the spatial node.

2.4 Detection of Pre-defined Functions

A library of simple BNs, representing species interactions or functional relationships, based on expertise knowledge (Table 1, **I**-invertebrates, **P**-pelagics, **SP**-small piscivorous and **LP**-large piscivorous and top predators) was created. Then, the experiment was conducted, in which each 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. Our model 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 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

2.5 Dynamic Bayesian Networks and Prediction

As well as learning functional relationships over space and time, we also explore network predictions over time. We choose to validate the networks through prediction by inferring dynamic Bayesian networks (DBNs) for each cluster and comparing the predicted biomass by either including or removing the spatial node from the model. Modelling time series is achieved by the DBN where nodes represent variables at particular time slices [6]. More precisely, a DBN defines the probability distribution over $\mathbf{X}[t]$ where $\mathbf{X}=X_1...X_n$ are the n variables observed along time t. To predict the biomass of each trophic group, we first infer the biomass at time t by using the observed evidence from time t-1. Two sets of experiments were then conducted: one that excludes the spatial node (DBN) and in the other, spatial node was included in the model (DBN+ spatial) to see if the node improves prediction. Non-parametric bootstrap analysis [6] was applied 250 times for each variant of the model (resulting in two model variants for each of the clusters) to obtain statistical validation in the predictions.

3 Results and Discussion

We were able to discover meaningful networks of functional relationships from ecological data, giving us confidence in the novel methods and results presented

here. 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 [12].

3.1 Functional Relationships Revealed by Hill-climbing

We now examine how learned by the model relationships amongst trophic groups of species vary across time and space. The relationship between invertebrates and pelagics (I-P) was found to be strongly significant (range: 0.3-1) and consistent in time and space (Fig.2a,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. 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.2c,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, 15) whilst clusters 11 and 19 increased around the same time and in most recent years.

The most highly confident small piscivorous- large predators (SP-LP) relationship (Fig.2e,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 to 2008 (clusters: 15, 5 and 7), whilst in other clusters increase in confidence was found for more recent years (for example clusters 1, 9).
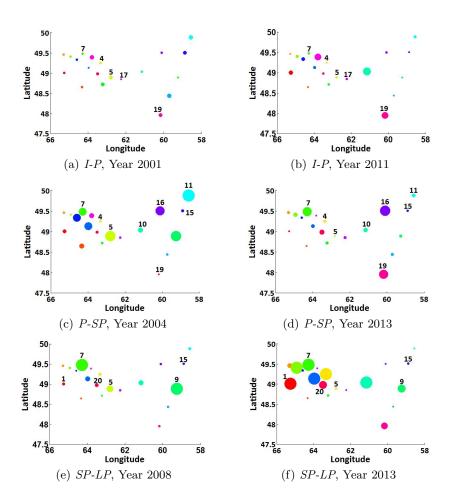
Fig. 2: The learned *I-P*, *P-SP* and *SP-LP* relationships for all 20 spatial clusters (size of scattered bubbles is equivalent to the estimated confidence by the hill-climb). The clusters mentioned in 3.1 are numbered.

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 [7]. Individual year effects are very strong for this area as time increases, as already suggested by [5] which makes it difficult to determine temporal trends. However, some of the clusters' temporal increase in early to mid-2000 (specifically for *I-SP* (cluster 5), *P-LP* (cluster 7) and *SP-LP* (cluster 5), Fig.3a,b,c) 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.3b,c) we suggest to be due to predation release of small abundant species by the selective fishing of larger predators [7]. 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.

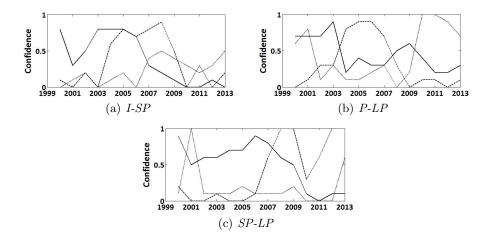
(a) *I-SP*

(b) *P-LP*

(c) *SP-LP*

Fig. 3: 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.

3.2 Summary of Discovered Functional Relationships

Next, we consider the variation of the pre-defined known functional relationships (Table 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.

3.3 DBNs and Prediction

We now turn to the generated predictions by the DBNs for each spatial cluster. To recall, two variants of each model were produced: DBN excluding the spatial node and DBN+ spatial 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 2). 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 DBN. In some clusters (for example 5 and 15, Fig.4), the predictive accuracy was significantly improved by the DBN+ spatial. 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. For example resource availability, habitat selection, processes like dispersal and metapopulation effects [7] but also commercial fishing could have influence on the local community stability and structure, resulting in our modelling approach identifying spatially-specific differences.

Table 2: SSE of DBN and DBN+ spatial. 95% confidence intervals reported in brackets

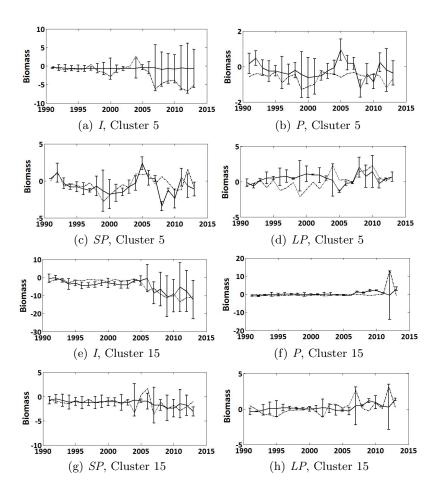| DBN	DBN+ spatial	DBN	DBN+ spatial
1. 5.58 ($\pm$9.29)	**1.** 4.38 ($\pm$7.08)	**11.** 12.44 ($\pm$20.56)	**11.** 16.54 ($\pm$34.34)
2. 0.24 ($\pm$0.36)	**2.** 0.14 ($\pm$0.12)	**12.** 69.90 ($\pm$308.02)	**12.** 30.55 ($\pm$64.30)
3. 16.20 ($\pm$29.92)	**3.** 10.76 ($\pm$17.16)	**13.** 12.68 ($\pm$16.63)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

Fig. 4: Biomass predictions generated by DBN+ spatial 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.

4 Conclusion

In this paper 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. We have also used knowledge of functional interactions between species to identify changes over time. 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. Future work will involve detailed analysis of each individual cluster with expansion on the functional networks.

References

1. Aderhold, A., Husmeier, D., Lennon, J.J., Beale, C.M., Smith, V.A.: Hierarchical Bayesian models in ecology: Reconstructing species interaction networks from non-homogeneous species abundance data. Ecological Informatics 11, 55–64 (2012)
2. Chickering, D.M., Geiger, D., Heckerman, D., et al.: Learning Bayesian networks is NP-hard. Tech. rep., Citeseer (1994)
3. Doubleday, W.: Manual on groundfish surveys in the Northwest Atlantic. Tech. rep., NAFO (1981)
4. Dunne, J.A., Williams, R.J., Martinez, N.D.: Network structure and biodiversity loss in food webs: robustness increases with connectance. Ecology letters 5(4), 558–567 (2002)
5. Duplisea, D.E., Castonguay, M.: 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 (2006)
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. Frisk, M.G., Duplisea, D.E., Trenkel, V.M.: Exploring the abundance-occupancy relationships for the Georges Bank finfish and shellfish community from 1963 to 2006. Ecological Applications 21(1), 227–240 (2011)
8. Gaston, K.J., Blackburn, T.M., Greenwood, J.J., Gregory, R.D., Quinn, R.M., Lawton, J.H.: Abundance–occupancy relationships. Journal of Applied Ecology 37(s1), 39–59 (2000)
9. Hartigan, J.A., Wong, M.A.: Algorithm as 136: A k-means clustering algorithm. Applied statistics pp. 100–108 (1979)
10. Heckerman, D., Geiger, D., Chickering, D.M.: Learning Bayesian networks: The combination of knowledge and statistical data. Machine learning 20(3), 197–243 (1995)
11. 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)
12. Milns, I., Beale, C.M., Smith, V.A.: Revealing ecological networks using Bayesian network inference algorithms. Ecology 91(7), 1892–1899 (2010)
13. Scheffer, M., Carpenter, S., Foley, J.A., Folke, C., Walker, B.: Catastrophic shifts in ecosystems. Nature 413(6856), 591–596 (2001)
14. Schwarz, G., et al.: Estimating the dimension of a model. The annals of statistics 6(2), 461–464 (1978)