Linking Structural Equation Modelling with Bayesian Network and Coastal Phytoplankton Dynamics in Bohai Bay

Jiangtao Chu and *Yue Yang*

Laboratory of Waterway Environmental Protection Technology, Tianjin Water Transport Engineering Science Research Institute, Tianjin 300456, China

Abstract. Bayesian networks (BN) have many advantages over other methods in ecological modelling and have become an increasingly popular modelling tool. However, BN are flawed in regard to building models based on inadequate existing knowledge. To overcome this limitation, we propose a new method that links BN with structural equation modelling (SEM). In this method, SEM is used to improve the model structure for BN. This method was used to simulate coastal phytoplankton dynamics in Bohai Bay. We demonstrate that this hybrid approach minimizes the need for expert elicitation, generates more reasonable structures for BN models and increases the BN model's accuracy and reliability. These results suggest that the inclusion of SEM for testing and verifying the theoretical structure during the initial construction stage improves the effectiveness of BN models, especially for complex eco-environment systems. The results also demonstrate that in Bohai Bay, while phytoplankton biomass has the greatest influence on phytoplankton dynamics, the impact of nutrients on phytoplankton dynamics is larger than the influence of the physical environment in summer. Furthermore, despite the Redfield ratio indicating that phosphorus should be the primary nutrient limiting factor, our results indicate that silicate plays the most important role in regulating phytoplankton dynamics in Bohai Bay.

1 Introduction

The eco-environment is a complex system encompassing physical, chemical, and biological processes, but our understanding of these natural processes is limited. Modelling these complex systems faces various challenges: simulating all of these processes in one system with different spatial, temporal or functional scales, combining different sources of data and knowledge, properly addressing uncertainty and incomplete data sets, and so on. It is difficult to resolve these problems with traditional mechanistic models. Bayesian networks (BN), which utilize probabilistic expressions to describe relationships among variables, is an increasingly popular method for ecological modelling and environmental management [1-3]. This popularity should be attributed to its features: BN can be used to cope with uncertainty in a natural way, combine data with domain knowledge, provides greater prediction accuracy even with small sample sizes or incomplete datasets, and performs integrated ecological modelling [4-6].

In this study, we devised a method for linking BN with SEM and applied it to a study of phytoplankton dynamics in Bohai Bay, China to examine the effectiveness of using SEM in developing BN models for ecological and environmental research. This was carried out as follows: a) we built three SE models based on expert advice and previous studies to explore the causal relationship between environmental factors and phytoplankton dynamics in Bohai Bay; b) we then explored ways of linking SEM with BN; c) next, we evaluated the performance of the BN model when linked with SEM and d) finally, we examined the role of SEM in developing BN models and its effectiveness in modelling phytoplankton dynamics in Bohai Bay.

2 Methods

2.1. Study Area

Bohai Bay is located in the western region of the Bohai Sea in northern China, covering an area of 15,900 km2 at an average depth of 12 m (Fig. 1). It is a semi-closed, shallow water system, with the exchange of water between Bohai Bay and the outer sea being quite weak [7]; as such, the physical self-purification capacity of Bohai Bay is quite poor. Due to the rapid rates of economic development and population growth in the coastal zone of Bohai Bay, enormous quantities of pollutants from municipal wastewater, urban runoff, animal feeding operations and agricultural fertilizers are being discharged into Bohai Bay from Beijing, Tianjin and Hebei provinces, and most of these nutrients accumulate gradually. As a result, Bohai Bay has been experiencing severe eutrophication events in recent years, and excessive algal blooms and low levels of dissolved

* Corresponding author Yue Yang: lixin1609@tju.edu.cn

© The Authors, published by EDP Sciences. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (http://creativecommons.org/licenses/by/4.0/).

oxygen occur frequently, leading to declines in shellfish populations and an increasing number of fish kills [8]. Because of this, Bohai Bay has been declared as one of the most threatened marine systems in China.

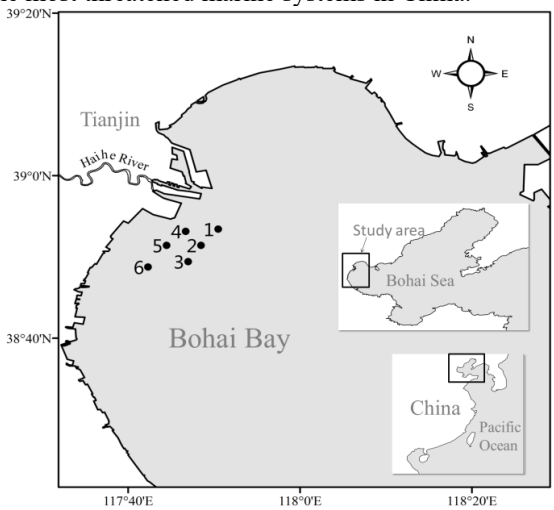


Fig 1 Bohai Bay and the locations of the six sampling stations.

2.2 Data description

The data used in this study were collected from 6 sampling stations in an ecological monitoring area of Bohai Bay, south of the estuary of the Haihe River (Fig. 1), with measurements of chemical, physical and biological processes. For chemical measurements, water samples were taken with Niskin water samplers and filtered immediately through pre-cleaned, 0.45 μ m pore-size acetate cellulose filters into a clean plastic tent, following which saturated HgCl2 solution was added to the filtrates. Nutrient (NO2-, NO3-, NH4+, PO43- and dissolved Si(OH)4 concentrations were quantified using the spectrophotometric methods previously reported in Liu et al. (2009). The analytical precision of NO2-, NO3-, NH4+, PO43- and Si(OH)4 were 0.06 μ mol L-1, 0.06 μ mol L-1, 0.09 μ mol L-1, 0.03 μ mol L-1, and 0.15 μ mol L-1, respectively. For physical measurements, salinity, pH and temperature were measured using YSI-6600 with an accuracy of 0.01 ppt, 0.01 and 0.01℃, respectively. For chlorophyll a measurement, the water samples were poured over Whatman GF/F glass microfiber filters. These filters were then immediately frozen under dark conditions until analysis in the laboratory using spectrophotometry with a Shimadzu UV-2550 [9].

3 Model Development

3.1 Relationship Exploration

In this study, SEM is used to provide a quantitative description of the relationship between phytoplankton dynamics and the coastal environment variables during the summer period in Bohai Bay. The work began by obtaining two competitive structural models, A and B (Fig. 2). Modelling processes described in previous literature asserted that the environmental factors only affect the dynamics of current phytoplankton [10], so the structural model A only considered the influence of the physical environment and nutrients on current phytoplankton dynamics, and did not include the lagged relationships (Fig. 2.a). However, the growth of phytoplankton is a two-step process: the first is to absorb nutrients, and the second to grow with the intercellular nutrient concentration. The response of phytoplankton dynamics may lag behind the effect of environmental factors. Thus, the structural model B evolves from the structural model A and is formulated based on the effects of the physical environment, nutrients and phytoplankton biomass on future phytoplankton dynamics (Fig. 2.b).

After completing the structural models, several sets of indicators should be identified that ‘‘best’’ describe each latent variable in the measurement models. Three observed variables – pH (pHt), salinity (St) and surface temperature (Tt) – were used in our models to measure the latent variable physical environment(the subscripted "t" denotes time and that the time step is 2 weeks).

It should be noted that chlorophyll- α was considered as a perfect measurement (i.e., no measurement error) of phytoplankton in this study. Thus, the latent variable current phytoplankton in the structural model A and structural model B was deputed by the concentration of current chlorophyll-a (Chlat), and the latent variable future phytoplankton in the structural model B was deputed by the concentration of chlorophyll-a two weeks later (Chlat+1). Two actual SE models were thus developed (Fig. 3a.b). In addition, considering that the effects on the total amount of future phytoplankton is different than the change rate of future phytoplankton caused by the environmental factors, the latent variable future phytoplankton was further set to be measured by the ratio of chlorophyll-a concentration after two weeks to current chlorophyll-a concentration (Relative Ratio), then the SE model B was developed to another actual SE model, called SE model C (Fig. 3c).

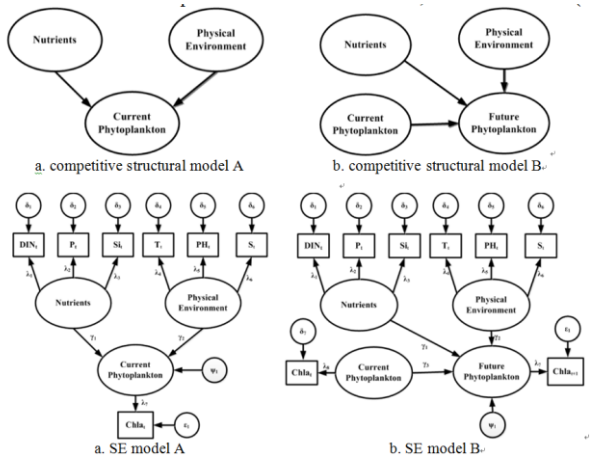


Fig 2 Three actual structural equation models

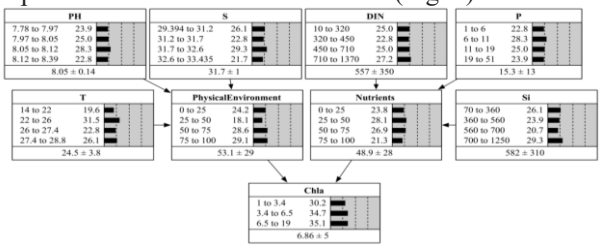
3.2 Models Definition

In general, there are four issues that need to be addressed when constructing a BN model: selecting the key factors (nodes), identifying the causal relationship between factors (nodes), discretizing the nodes and calculating

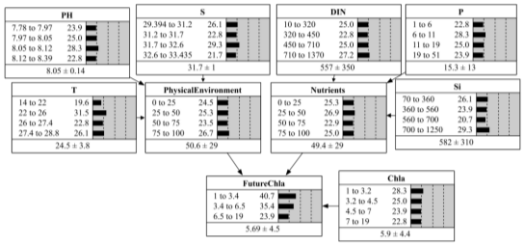
the conditional probability tables (CPTs) associated with the nodes. Of these, the first two problems form the basis for the other issues of BN model, and they determine the accuracy and authenticity of the entire BN model. In this study, we used SEM to cope with these two issues. While the SE model has been shown to fit the observation data well, its measurement and structural models are considered to be realistic, which means that the selected observed variables are effective at measuring latent variables and the causal relationship represented by the structural model is reliable. Thus, in a BN model, the selection of key factors and the identification of the links can follow the measurement model and the structural model, respectively, of the SEM with desired results. In section 2.3, three satisfactory SE models were constructed. To compare and validate them, three BN models – A, B and C – were built based on the SE models A, B and C, respectively (Fig. 4).

For the discretization of nodes, the number of states and the cut-off values of the discrete states need to be defined. In general, nodes with fewer states may provide a more accurate prediction, although more states may provide greater precision. Thus, to balance accuracy and precision, the number of states was kept to four or fewer. In this study, all of the predictor variables were limited to 4 states and all the target variables were restricted to 3 states. Furthermore, to define the cut-off value of each discrete state, every state of the node should include similar amounts of observation data.

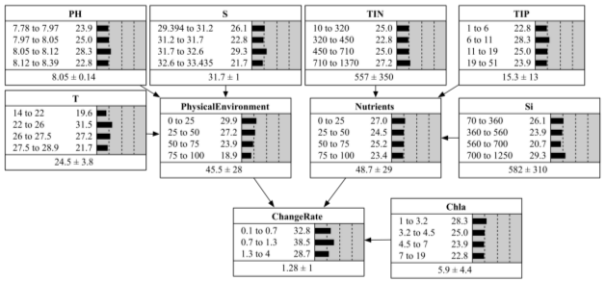
The most crucial problem of a BN model is determining the set of conditional probability tables (CPTs) underlying each node. The function of CPTs is to quantify the cause–effect relationship between variables. At the initial state, due to the analogous number of observation data in each state, all nodes were set to uniform probabilities and all CPTs were assigned to uniform values. Then, the BN models were updated automatically by using learning algorithm with sample data. Because the BN models include latent variables, the expectation maximization (EM) learning algorithm, which could cope with missing data in the case files and automatically calculate CPT values using case data, was used to update the BN models. Solving issues greatly improves the structure of BN models (Fig. 4).



a. BN model A



b. BN model B



c. BN model C

Fig 3 The three non-informed BN models A, B and C.

BN models with different structure perform very differently, which was demonstrated by the error matrix (Table 2). For practical applications, the structure of a BN model is often difficult to specify, especially when a system is poorly understood and hard to characterize. Moreover, there is the possibility of spurious relationships in learning the structure. While these may limit the application of BNs in environmental research, the SEM provides a potential new approach for addressing such problems. Based on expert knowledge and the observation data, ecologists can use SEM to explore and validate causal relationships among ecological factors. After testing several pre-specified SE models, the model that provides meaningful interpretations and perfect goodness-of-fit is considered to represent the true "ecological causal web" of the key environmental correlates. Thus, SEM can provide more credible structure for and improve BN models during the process of structure building.

Compared with the previous method, linking SEM with BN models has several advantages:

a) It is suitable for relatively new and highly contentious issues, such as the initiation of harmful algal blooms in coastal areas, which are very common in ecological research. In general, the structure building mode of BN is often limited to expert-input and/or an initial review of the literature. However, if there is no widely accepted interpretation of the objective process, it would be inadequate for eliciting a prior structure, and may even result in the presence of spurious relationships. As an advanced statistical method, SEM combines expert knowledge and observations. With SEM, ecologists can hypothesize the relevance and influence of ecological variables with expert knowledge and test this causal relationship against observation data. After several specifications and with continual improvement, ecologists can obtain a truly ecological causal web and thereby improve the BN model structure via obtained causal relationships. Thus, SEM can help to minimize the need for expert elicitation and increase the application range of BN models.

b) It provides another reliable way to verify BN models. Verification is an important checkpoint in the construction of a BN model and a very necessary step for the model builder toper form. Generally speaking, verification is carried out to a certain extent according to some subjective criteria, whereas SEM provides an effective means of verifying BN models following objective criteria. Through different solution processes, SEM and BN models can both determine the absolute degree and the rank order of the influence of

environmental factors or latent variables on outcome variables. Thus, by comparing the path coefficient of an SE model with the sensitivity analysis results of a BN model, the modeller can verify the BN model. Such an approach would be an important complement to the traditional verification process of BN models.

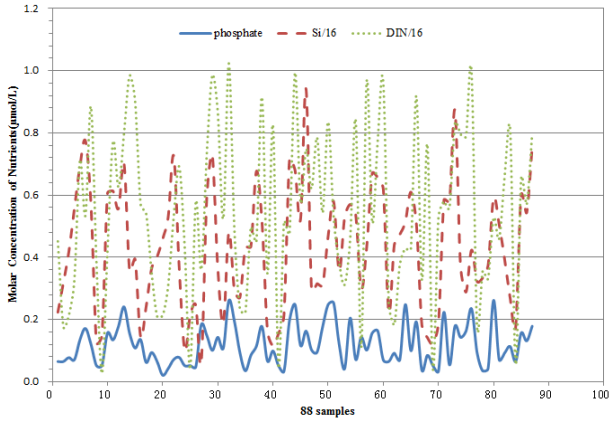


Fig 4 The molar concentration dynamic variation of the three nutrients.

4 Conclusions

In this study, a method of linking an SEM with a BN model is introduced as a means of overcoming the limitations of building BN model structure with inadequate existing knowledge. In the structural building step of a BN model, SEM was used to explore and validate the relationship between environmental factors, based on expert knowledge and observation data. The results indicate that SEM successfully selected the structure that was closer to the true causal relationship and minimizes the need of expert elicitation in the BN modelling process. In addition, SEM also increases the accuracy and reliability of the BN model. Thus, the application of a coupled SEM–BN model shows great promise as a tool for ecological research.

Our model was applied to Bohai Bay and some useful results were obtained: phytoplankton biomass has the largest influence on phytoplankton dynamics, the impact of nutrients on phytoplankton is larger than the physical environment and that controlling nutrient concentrations. To decrease the harmful algal blooms, controlling the amount of the nutrients would be the most effective means of reducing the frequency and magnitude of algal blooms. Finally, although of the three nutrients phosphorus is usually considered to be the limiting nutrient in Bohai Bay, the silicate is significantly correlated with phytoplankton dynamics.

Acknowledgements

This work was supported by “Study of foreign soil water retention, prevent salinization, heavy metal soil remediation, soil backfill and ecological restoration technique in land reclamation. (TKS150218).

References

1. Alameddine, I., Cha, Y., Reckhow, K.H., 2011. An evaluation of automated structure learning with Bayesian networks: An application to estuarine chlorophyll dynamics. Environmental Modelling & Software, 26(2), 163-172.
2. Anderson, R., 2004. Causal modeling alternatives in operations research: Overview and application. European Journal of Operational Research, 156(1), 92-109.
3. Arhonditsis, G., Paerl, H., Valdesweaver, L., Stow, C., Steinberg, L., Reckhow, K., 2007. Application of Bayesian structural equation modeling for examining phytoplankton dynamics in the Neuse River Estuary (North Carolina, USA). Estuarine, Coastal and Shelf Science, 72(1-2), 63-80.
4. Arhonditsis, G., Stow, C., Steinberg, L., Kenney, M., Lathrop, R., McBride, S., Reckhow, K., 2006. Exploring ecological patterns with structural equation modeling and Bayesian analysis. Ecological Modelling, 192(3-4), 385-409.
5. Barton, D., Saloranta, T., Moe, S., Eggestad, H., Kuikka, S., 2008. Bayesian belief networks as a meta-modelling tool in integrated river basin management — Pros and cons in evaluating nutrient abatement decisions under uncertainty in a Norwegian river basin. Ecological Economics, 66(1), 91-104.
6. Borsuk, M., 2004. A Bayesian network of eutrophication models for synthesis, prediction, and uncertainty analysis. Ecological Modelling, 173(2-3), 219-239.
7. Bromley, J., Jackson, N.A., Clymer, O.J., Giacomello, A.M., Jensen, F.V., 2005. The use of Hugin® to develop Bayesian networks as an aid to integrated water resource planning. Environmental Modelling & Software, 20(2), 231-242.
8. Brzezinski, M.A., 1985. THE Si: C: N Ratio of Marine Diatoms: Interspecific Variability and The Effect of Some Environmental Variables1. Journal of Phycology, 21(3), 347-357.
9. Castelletti, A., Soncinisessa, R., 2007. Bayesian Networks and participatory modelling in water resource management. Environmental Modelling & Software, 22(8), 1075-1088.
10. Chen, Y., Lin, L.-S., 2010. Structural equation-based latent growth curve modeling of watershed attribute-regulated stream sensitivity to reduced acidic deposition. Ecological Modelling, 221(17), 2086-2094.