VATTEN – Journal of Water Management and Research 69: 15–19. Lund 2013

PAIRING MULTIVARIATE DATA ANALYSIS AND ECOLOGICAL MODELING IN THE BIOMANIPULATED LAKE ENGELSHOLM, DENMARK

Parvis multivariat dataanalys och ekologisk modellering i den biomanipulerade sjön Engelsholm, Danmark

by FÁBIO PEREIRA[1], CARLOS RUBERTO FRAGOSO JR[2], CINTIA UVO[1], DAVID DA MOTTA MARQUES[3]
1 Department of Water Resources Engineering, Lund University, Box 118, 221 00 Lund, Sweden
e-mail address: fabio.pereira@tvrl.lth.se
2 Center for Technology, Federal University of Alagoas, Maceió, Brazil
3 Institute for Hydraulics Research, Federal University of Rio Grande do Sul, Porto Alegre, Brazile




Abstract

A multivariate analysis was applied to monthly climatic, hydrologic and water quality series of Lake Engelsholm, a shallow lake that was subjected to biomanipulation. In this study, we focused on identifying the most important patterns of monthly time series after biomanipulation. Moreover, we tried to quantify how those variables are interconnected with each other. A multivariate data analysis was made by using Principal Component Analysis (PCA) and Cluster Analysis. A statistical model using Canonical Correlation Analysis (CCA) was developed to predict the biological variables. Its efficiency was compared to a conceptual ecological model called IPH-ECO. The predictions made by IPH-ECO presented a better performance than those predictions using multivariate analysis in the period after biomanipulation.

Key words – Ecological modeling, principal component analysis, cluster analysis, biomanipulation, Lake Engelsholm

Sammanfattning

En multivariat statistisk analys genomfördes med månadsvisa klimat-, hydrologi- och vattenkvalitésdata för sjön Engelsholm, Danmark, en grund sjö där biomanipulation genomförts. Vi fokuserade i den här studien på att indentifiera de viktigaste mönstren i tidsserierna och månadsmedel efter biomanipulationen av det akvatiska ekosystemet. Vidare försökte vi kvantificera hur dessa variabler är förbundna med varandra. Multivariat statistisk analys genomfördes med Principal Component Analysis (PCA) och Cluster Analysis för att identifiera mönster. Vidare utvecklades en statistisk modell med hjälp av Canonical Correlation Analysis (CCA) för att förutsäga de biologiska variablerna utifrån observerade variabler. Beräkningens duglighet jämfördes med en konceptuell, ekologisk modell, IPH-ECO. De biologiska variablerna som beräknades m.h.a. den konceptuella, ekologiska modellen gav bättre korrelation mellan observerade och simulerade värden i perioden efter biomanipulationen jämfört med den multivariata statistiska analysen.

Introduction

Stochastic and deterministic models have been widely used in short and long-term predictions of variables which describe a range of physical processes in the atmosphere, oceans and lakes. However, stochastic models provide their outputs based on a probability or an element of randomness of their input data whereas deterministic models represent physical, biological or chemical processes by conceptual formulas and equations (Clarke, 1973).

Several stochastic models present multivariate analysis to reproduce environmental processes (Dillon and Goldstein, 1984). One of the advantages of these models is

that they are not limited by numerical stability requirements such as Courant-Friedrich-Lewy (CFL) and Courant number condition (Casulli and Cattani, 1994). However, uncertainties associated with their predictions are related to how stationary and large are their database (Scheffer et al., 1993).

Although deterministic models also use two or more variables for environmental modeling (Tucci, 1998), their application requires a database large enough for their calibration and validation. Once calibrated and validated, deterministic models can accurately predict a range of physical, chemical or biological phenomena (Rosman, 2000; Fragoso Jr, 2009).

Both types of models are often employed for solving realistic problems in limnology and water resources (Bernardi et al, 2001; Fragoso Jr, 2009; Pereira, 2010). Recently, a way to group or associate biological functions or variables with regional climate and hydrology in lakes has been discussed by many authors (Scheffer et al., 1993; Alcântara et al., 2004; Léon et al., 2006).

Thus, this study presents a multivariate statistical analysis applied to monthly time series of climate, hydrologic and water quality variables for Lake Engelsholm, a small and shallow Danish lake which was biomanipulated. Predictions performed using multivariate analysis were compared to predictions made using a deterministic ecological model called IPH-ECO.

Methods

In this work, two different methodologies were applied to Lake Engelsholm in order to predict and identify trends in biological variables after its biomanipulation. A biomanipulation is defined as any alteration of an ecosystem by adding or removing species. The first methodology consists of using a multivariate analysis whereas the second one is based on using a conceptual ecological model (IPH-ECO).

A multivariate data analysis was performed to capture important patterns in the database after biomanipulation as well as how variables are interrelated with each other. Firstly, a Principal Component Analysis (PCA) was employed on Lake Engelsholm data set that contains external (i.e. inflow, outflow, phosphorus, nitrogen and silica inputs), water quality (i.e. Chlorophyll a, NO3, NH4, PO4 etc.) and climatic (i.e. air temperature, solar radiation etc.) variables. The purpose was to identify variables that present similar trends as well as those which drive Lake Engelsholm dynamics.

This PCA has been followed by Cluster Analysis. A Cluster Analysis was used for grouping and determining external, water quality or climatic variables of high similarity. The four first eigenvectors from the PCA were provided as input data to the Cluster Analysis. As output, a dendrogram chart was created in order to show the relations among all variables in Lake Engelsholm database.

As a last step, a stochastic model was developed using Canonical Correlation Analysis to predict the variability of biological variables in the Lake Engelsholm and its predictions were compared to predictions made by a conceptual ecological model (IPH-ECO).

Study area

Lake Engelsholm is a shallow lake with a maximum depth of 6.1 m, a mean depth of 2.4 m, surface area of 44 ha and volume of approx. $1.4 \cdot 10^6$ m^3. It has a hydraulic retention time of 61–104 days increasing to 177–292 days in the summer. Nowadays, the lake catchment (15.2 km^2) consists of cultivated areas (78 %), forested hills (16 %) and scattered dwellings (6 %). Considering an expansion of agriculture over its catchment during the 70's, Lake Engelsholm was heavily loaded by nutrients leading to an increasing of algal biomass with blooms of cyanobacteria and low water transparency. The chronology of events in the Lake Engelsholm is presented in Figure 1 where a fish manipulation has been made from April 1992 to September 1994 (Sondergaard et al., 1999). It consisted of removing cyprinids (planktivorous fish) in order to reduce the fish pressure on zooplanktons.

Data analysis

As the biomanipulation changes several physical and biological features of the Lake Engelsholm (e.g. species composition, exchanges between water column and

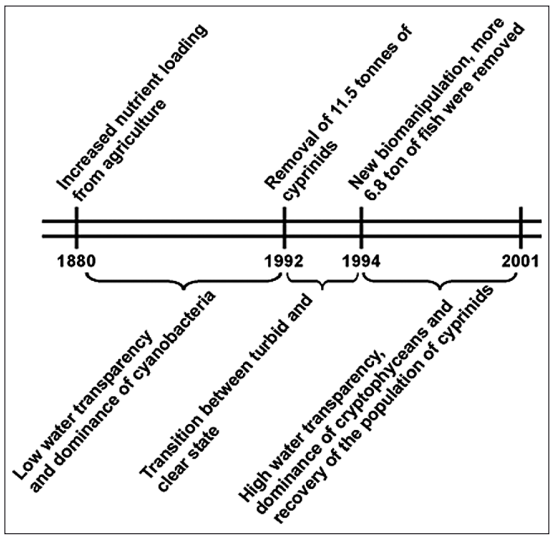

Fig. 1. *Chronology of events and conditions in the Lake Engelsholm (Font: Fragoso Jr, 2009).*

Table 1. *Mean, deviation standard and variance of all variables of the Lake Engelsholm database.*

Variables	Name	Mean	Standard deviation	Variance
External	Inflow ($m^3\ s^{-1}$)	0.0137	0.00443	0.00002
	Outflow ($m^3\ s^{-1}$)	0.0143	0.00514	0.00003
	InputP (kgP d^{-1})	1.192	0.546	0.298
	InputN (kgP d^{-1})	87.024	31.33	981.542
	InputSi (kgSi d^{-1})	46.174	14.902	222.084
Water quality	Chla (µg L^{-1})	25.193	25.038	626.914
	NH4 (mg L^{-1})	0.0875	0.090	0.00812
	NO3 (mg L^{-1})	0.9415	0.868	0.753
	PO4 (mg L^{-1})	0.0138	0.012	0.00015
	Schecci (m)	2.236	1.222	1.494
	Zoo (mg L^{-1})	0.8115	0.645	0.416
	SiO2 (mg L^{-1})	4.915	2.496	6.230
	Total P (mg L^{-1})	0.0565	0.026	0.00065
	Total N (mg L^{-1})	1.533	0.866	0.751
Climatic	Water Temp. (°C)	10.149	6.543	42.806
	Air Temp. (°C)	8.153	5.934	35.210
	Wind (m s^{-1})	4.183	0.792	0.627
	Solar Radiation ($Wm^{-2}\ d^{-1}$)	113.376	78.090	6098.027
	Precipitation (mm d^{-1})	2.276	1.215	1.476
	Evaporation (mm d^{-1})	1.338	1.149	1.321

sediment, mortality and predation rates), data analysis were conducted considering only observations after the biomanipulation. Thus, a total period of 80 months (Jan 1994 to Sept 2000) yielded 80 values of monthly average per each variable. Table 1 shows a full list of the variables used with their mean, standard deviation and variance.

Ecological model

IPH-ECO model is a complex ecosystem model developed at Institute for Hydraulics Research (IPH-UFRGS) for investigations involving physical, chemical and biological processes in lakes and reservoirs. It consists of hydrodynamic module coupled to chemical and biological modules. A detailed description of its application in Lake Engelsholm can be found in Fragoso Jr (2009).

Results and discussions

Principal Component Analysis

Principal component analysis applied to Lake Engelsholm data shows that almost 60 % of the variance is accounted by the first two principal components where the first principal component explains 45.43 % while the second principal component represents 13.29 % of the total variance in the data (Tab. 2).

Table 2. *Variance explained by each principal component (eigenvalues) and the weight of each variable into its components (eigenvectors).*

Variables	Mod 1	Mod 2
Explained Variance (%)	45.43	13.29
Explained Variance Accum (%)	45.43	58.72
Inflow ($m^3\ s^{-1}$)	–0.2765	–0.2653
Outflow ($m^3\ s^{-1}$)	–0.2894	–0.2571
InputP (kgP d^{-1})	–0.2215	–0.2766
InputN (kgP d^{-1})	–0.2711	–0.1920
InputSi (kgSi d^{-1})	–0.2765	–0.2653
Chla (µg L^{-1})	0.1890	–0.2349
NH4 (mg L^{-1})	–0.1709	0.0754
NO3 (mg L^{-1})	–0.2880	0.1428
PO4 (mg L^{-1})	–0.1456	–0.0627
Schecci (m)	–0.0560	0.3319
Zoo (mg L^{-1})	0.0147	0.2517
SiO2 (mg L^{-1})	–0.0054	–0.1261
Total P (mg L^{-1})	0.1397	–0.3380
Total N (mg L^{-1})	–0.2743	0.1002
Water Temp. (°C)	0.2944	–0.2201
Air Temp. (°C)	0.2741	–0.2557
Wind (m s^{-1})	–0.2376	–0.0168
Solar Radiation ($Wm^{-2}\ d^{-1}$)	0.2591	–0.1302
Precipitation (mm d^{-1})	–0.1262	–0.3615
Evaporation (mm d^{-1})	0.2687	–0.1384

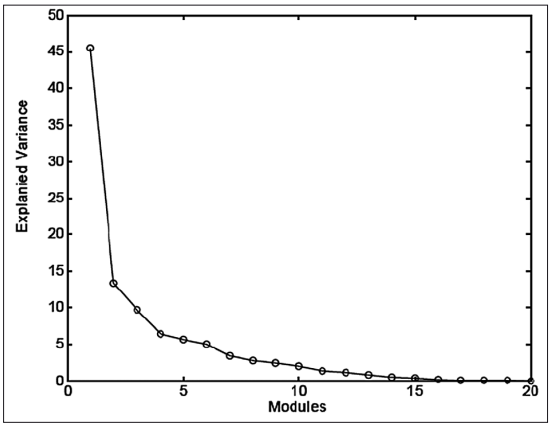

Fig. 2. *Contribution of principal components to the total explained.*

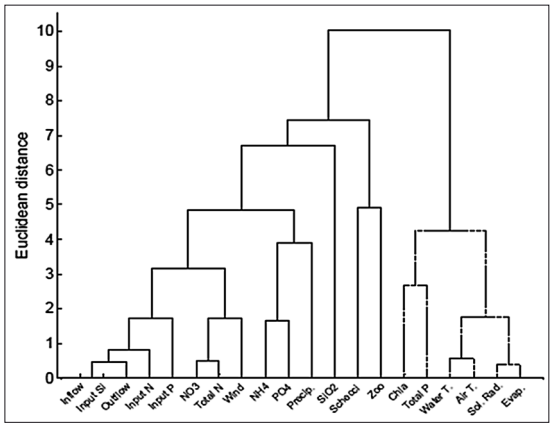

Fig. 3. *Cluster Analysis in the variables which present similar patterns in Lake Engelsholm.*

A scatter plot of the scores of the all principal components is shown in Fig. 2. Since the explained variance presents peaks during the summer and winter, it implies that water quality variables are induced by climatic variables in Lake Engelsholm.

The eigenvector of the first principal component indicates that the external variables are well correlated with some of the water quality variables such as nitrate (NO3), ammonia (NH4), orthophosphate (PO4), silicate (SiO2) and total nitrogen (Total N). It means that the variability of these variables is driven by external nutrient loads originated from agricultural land use within the Lake Engelsholm watershed. Water column transparency (measured with Secchi disk) follow the same trend as wind intensity since it affects sediment resuspension which increases the exchange of organic matter and suspended solids across the sediment-water interface. The variability of total phosphorus loads, however, is explained by phytoplankton population densities. Seasonal patterns in algal biomass (measured by chlorophyll -a) are closely related to climatic variables whose values are directly linked to primary production (i.e. water temperature, air temperature, evaporation, and solar radiation). On the other hand, the variability of nutrient loads responds inversely to variations in chlorophyll-a (chl-a) once algal production increases nutrient assimilation in the water column.

The second principal component indicates that chlorophyll-a is related to both external variables as climatic variables. It implies the external variables may also affect primary production in Lake Engelsholm where nutrient assimilation plays an important role on changes in algal biomass.

The cluster analysis corroborated the results of the principal component analysis. Two main groups were identified (Fig. 3). The first one is composed by external and water quality variables which showed a good correlation among all variables analysed. The second group includes chlorophyll-a and climatic variables which means variations in chlorophyll-a may be explained by the regional climate in Lake Engelsholm.

Stochastic and deterministic predictions

A stochastic model was generated by applying Canonical Correlation Analysis (CCA) to the observations at Lake Engelsholm where external, climatic and water quality variables were entered as predictors and biological variables as predicted variables. The canonical vectors are shown in Fig. 4.

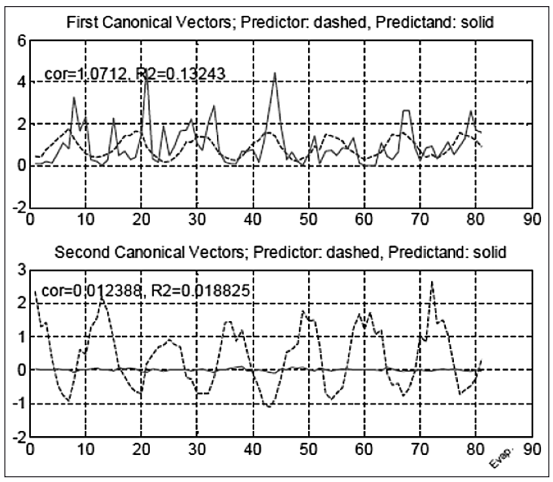

Fig. 4. *The new time series (canonical vectors) transformed according to CCA.*

By cross-validation, the performances of both stochastic and deterministic models were evaluated. The full cross-validation indicated weaker correlation between predictions made with the stochastic model (R^2 = 0.238 for chl-a and R^2 = 0.184 for zooplanktons) than predictions using the deterministic model (R^2 = 0.643 for chl-a and R^2 = 0.554 for zooplanktons) in the period after biomanipulation.

Conclusions

The use of multivariate statistical analysis in the Lake Engelsholm database has shown to be an important tool for identifying short and long-term trends in the climatic and water quality variables as well as to analyse how these trends may affect nutrients dynamics in the Lake Engelsholm. Therefore, multivariate statistical analysis may support deterministic models by creating relationships between variables, providing new parameters and adjusting processes.

The predictions of biological variables made with a stochastic model based on Canonical Correlation Analysis presented lower accuracy than those predictions using a deterministic model (IPH-ECO). However, deterministic models as IPH-ECO requires a longer time for pre and post-processing data and results than stochastic models, a good calibration, validation beyond the user's familiarity with the processes that a deterministic model covers. Stochastic models, on the other hand, provide faster solutions and their accuracy depend on the stationary and size of the database.

Acknowledgments

The development of this work has been supported by the Swedish Research Council (Vetenskaprådet) and the Crafoord Foundation.

References

Alcântara, E., Mochel, F., Amorim, A., Thevand, A. (2004) Modelagem da Profundidade Secchi e da Concentração de Clorofila a no Estuário do Rio Anil, São Luís-MA, Caminhos da Geografia, vol. 5, pp. 19–40.

Bernardi, J., Fowler, H., Landim, P. (2001) Um estudo de impacto ambiental utilizando análises estatísticas espacial e multivariada, HOLOS Environment, vol. 1, pp. 162–172.

Casulli, V., Cattani, E. (1994) Stability, accuracy and efficiency of a semi-implicit method for three-dimensional shallow water flow, Computers and Mathematics with Applications, vol. 27, pp. 99–112.

Clarke, R. (1973) Mathematical Models in Hydrology, New York: Environmental Science Series.

Dillon, W. and Goldstein, M. (1984) Multivariate Analysis – Methods and applications, New York: John Wiley & Sons.

Fragoso Jr, C. (2009) Modelagem Tridimensional da Estrutura Trófica em Ecossistemas Aquáticos Continentais Rasos, PhD. Thesis, Instituto de Pesquisas Hidráulicas, Universidade Federal do Rio Grande do Sul, Porto Alegre (in Portuguese).

Fragoso Jr, C., Van NES, E., Motta Marques, D., Jeppensen, E. (Submitted). Modelling the biomanipulated Lake Engelsholm: are ecosystem changes after a large disturbance predictable?

Hotelling, H. (1936) Relations between two sets of variants, Biometrika, vol. 28, pp. 321–377.

Jackson, J. (1991) A user's guide to principal components, New York: John Wiley & Sons.

Jeppensen, E., Jensen, J., Sondergaard, M., Lauridsen, T., Moller, F., Sandby, K. (1998) Changes in nitrogen retention in shallow eutrophic lakes following a decline in density of cyprinids, Archiv Fur Hydrobiologie, vol. 142, pp. 129–151.

Leon, L., Lam, D., Schertzer, W., Swayne, D., Imberger, J. (2006) Towards Coupling a 3D Hydrodynamic Lake Model with the Canadian Regional Climate Model: Simulation on Great Slave Lake, Enviromental Modelling & Software, vol. 22, pp. 787–796.

Pereira, F. (2009) Modelo Hidrodinâmico e de Transporte Bidimensional de Grade Não Estruturada para Lagos Rasos, Master Thesis (Mestrado em Recursos Hídricos e Saneamento Ambiental), Instituto de Pesquisas Hidráulicas, Universidade Federal do Rio Grande do Sul, Porto Alegre (in Portuguese).

Rosman, P. (2000) Referência Técnica do SisBaHiA – Sistema Base de Hidrodinâmica Ambiental, Programa COPPE: Engenharia Oceânica, Área de Engenharia Costeira e Oceanográfica, Rio de Janeiro, Brasil.

Scheffer, M., Hosper, S., Meijer, M.-L., Moss, B., Jeppensen, E. (1993) Alternative Equilibria in Shallow Lakes, Trends in Ecology and Evolution, vol. 8, pp. 275–279.

Sondergaard, M., Jensen, J., Jeppesen, E. (1999) Internal phosphorus loading in shallow Danish lakes, Hydrobiologia, vol. 408, pp. 145–152.

Tucci, C. (1998) Modelos Hidrológicos, Porto Alegre: Editora da UFRGS.