Reduction of complex models using data-mining and nonlinear projection techniques

Bernhardt, K. [a], Wirtz, K.W.

[a]Institute for Chemistry and Biology of the Marine Environment (ICBM)
Carl-von-Ossietzky University Oldenburg
P.O. Box 2503
26111 Oldenburg, Germany

Abstract: Complex models of environmental systems typically depend on a large amount of uncertain parameters. Therefore, they are often difficult to handle and do not provide an insight into effective modes of the underlying system's dynamics. Unlike earlier analytical attempts to find more effective model representations, we present a new combination of methods that only relies on data generated by complex, process-based models. These methods are taken from the field of data-mining and enable the recognition of patterns in measured or modelled data by unsupervised learning strategies. As these methods do not directly lead to a better understanding of the systems' driving processes, we suggest the linkage between pattern recognition and process identification by a multi-stage approach. In a first step, a large data-base was produced by a mechanistic model for species competition in a virtual ecosystem for a range of parameter settings. Using Vector Quantisation and nonlinear projection techniques such as Self-Organising Maps and nonlinear Principal Component Analysis, typical states of the complex model's dynamics as well as major pathways connecting these states were then identified. The visualisation of the results points to the existence of nonlinear transformations of former model state variables and parameters to few effective variables. Effective variables built this way preserve most of the model's dynamic behaviour, while they are nonetheless easier to use and require much less parameterisation effort.

Keywords: model reduction; data-mining; Self-Organising Map; nonlinear projection; multispecies competition

1 INTRODUCTION

Process-based models are widely used for modelling key mechanisms ruling ecosystem dynamics. The vast number of potentially relevant interactions and adaptations in ecosystems thereby increases the corresponding model complexity. Secondly, process identification is rarely unique, i.e. data can be reproduced with a variety of models or parameterisations (see e.g. Beven [2001]). Given this high complexity and the sparseness of data, parameter uncertainty is difficult to handle in these models.

An alternative way to reproduce and extrapolate data is the use of methods taken from the field of data-mining, such as Neural Networks (NN), clustering methods or (non-)linear projection techniques which are able to 'learn' distinct features of a dataset (Fodor [2002]). No knowledge of the underlying system is required using data-driven methods. New understanding of underlying key mechanisms however, can not be gained and generic models with a large domain of applicability are not provided.

The aim of this work is the construction of a new type of deterministic but reduced and efficient model from a classic complex mechanistic model by only using information contained in the data generated by the complex model. A nonlinear statistical analysis and reduction of this data should reflect the overall dynamics even for uncertain model parameterisations and should yield interpretable information on dominant internal modes. We propose a multi-step analysis using data-mining and nonlinear projection techniques to extract these modes or 'effective variables' (Wirtz and Eckhardt [1996]). Those variables have been shown to successfully replace complex descriptions of adaptive processes in biosystems (e.g.Wirtz [2002]). Up to now they had to be built using intuitive modelling knowledge which is a major impediment for a broader use.

The recombination of the effective variables resulting in a reduced-form deterministic model consequently combines the benefits of the process-oriented as well as data-mining approaches. The existence of such a reduced representation is supported by the finding, that even huge ecosystem models have a limited number of internal dynamical modes (Ebenhöh [1996]).

In principal, the proposed reduction scheme can be applied to any deterministic process-based model. In this study, we present the extraction of effective variables using a combination of Vector Quantisation algorithms such as the Self-Organising Map (Kohonen [1997]) and nonlinear Principal Component Analysis (Kramer [1991]).

We have chosen the reduction of a prominent model of multispecies competition with rich dynamics including chaotic behaviour (Huisman and Weissing [1999]) as a test case.

2 A MODEL OF SPECIES COMPETITION

The model analysed in this study was proposed by Huisman and Weissing [1999]. It describes competition for resources like phytoplankton species concurring for nitrogen and phosphorus.

Consider n_P phytoplankton species and n_N nutrients. Let state variables $\hat{P}_i$ and $\hat{N}_j$ be the population abundance of species i and the availability of resource j respectively. The dynamics of species i follows

$$\frac{d\hat{P}_i}{dt} = \hat{P}_i \cdot (\mu_i - \omega_i) \qquad i = 1, \ldots, n_P \tag{1}$$

where ω_i are model parameters describing the mortality. The growth rate μ_i is controlled by the most limiting resource via a minimum of Monod functions with K_{ji} denoting the half-saturation constant for resource j of species i and g_i the maximal growth rate:

$$\mu_i = \min_v \left(\frac{g_i \hat{N}_v}{K_{vi} + \hat{N}_v} \right) \qquad v = 1, \ldots, n_N. \tag{2}$$

The time evolution of the abiotic resource j is described as

$$\frac{d\hat{N}_j}{dt} = D \cdot \left(S_j - \hat{N}_j \right) - \sum_i c_{ij} \cdot \mu_i \cdot \hat{P}_i \qquad j = 1, \ldots, n_N \tag{3}$$

where D is a constant factor describing the nutrient turnover rate, S_j is the supply concentration and parameters c_{ij} quantify the content of nutrient j in species i.

For different choices of model parameters the system can be driven into attractors with different topologies containing fixed-point dynamics (no changes in species abundances for one or more species), limit cycles (fluctuating coexistence of species) or chaotic behaviour. For further details on the parameter settings see Huisman and Weissing [1999, 2001].

To keep the analysis simple, we numerically integrated (1) and (3) to produce 16 time series of 2000 points each for a model configuration with five species ($n_P = 5$) and three abiotic resources ($n_N = 3$) by varying only two of the half-saturation constants (k_{21} and k_{25}). The other model parameters (D, S_j, ω_i, g_i and c_{ij}) were kept at the fixed values used in Huisman and Weissing [2001]. Time series modelled this way show all sorts of dynamics described above (see Figure 1).

3 THE SELF-ORGANISING MAP

The Self-Organising Map (SOM) algorithm was introduced by Kohonen [1997]. It resembles a neural network variant consisting of topologically ordered nodes on a grid of predefined dimensionality.

A SOM is able to 'learn' structures of high-dimensional input data-vectors and to project them onto a lower-dimensional output space. It is therefore often used for Vector Quantisation (VQ) where a reduced representation of complex datasets is built by replacing the data-vectors with a smaller subset of so-called prototype vectors. Additionally, the existence of the typically two-dimensional output grid simplifies the visual inspection of the dataset and helps to identify patterns inherent to the data.

The algorithm transforms a dataset consisting of vectors $\mathbf{x}(t) = (x_1(t), x_2(t), \ldots, x_n(t))^T \in \Re^n$ with discrete-time coordinate $t = 0, 1, 2, \ldots$, e.g. measurements of n variables over time. In this case, each $\mathbf{x}(t)$ is a ten-dimensional vector with entries $\hat{P}_i$, $\hat{N}_j$, k_{21} and k_{25} ($n = n_P + n_N + 2$).

The SOM-network consists of a z-dimensional array of k nodes associated with prototype-vectors $\mathbf{m}_k \in \Re^n$ with orthogonal or hexagonal neighbourhood relationships between adjacent nodes.

The data-vectors are iteratively compared with all $\mathbf{m}_k$ by using euclidean distances to find the best-matching node denoted by c. The updating procedure for prototype s then follows

$$\mathbf{m}_s(t+1) = \mathbf{m}_s(t) + h_{cs} \cdot [\mathbf{x}(t) - \mathbf{m}_s(t)]\,, \tag{4}$$

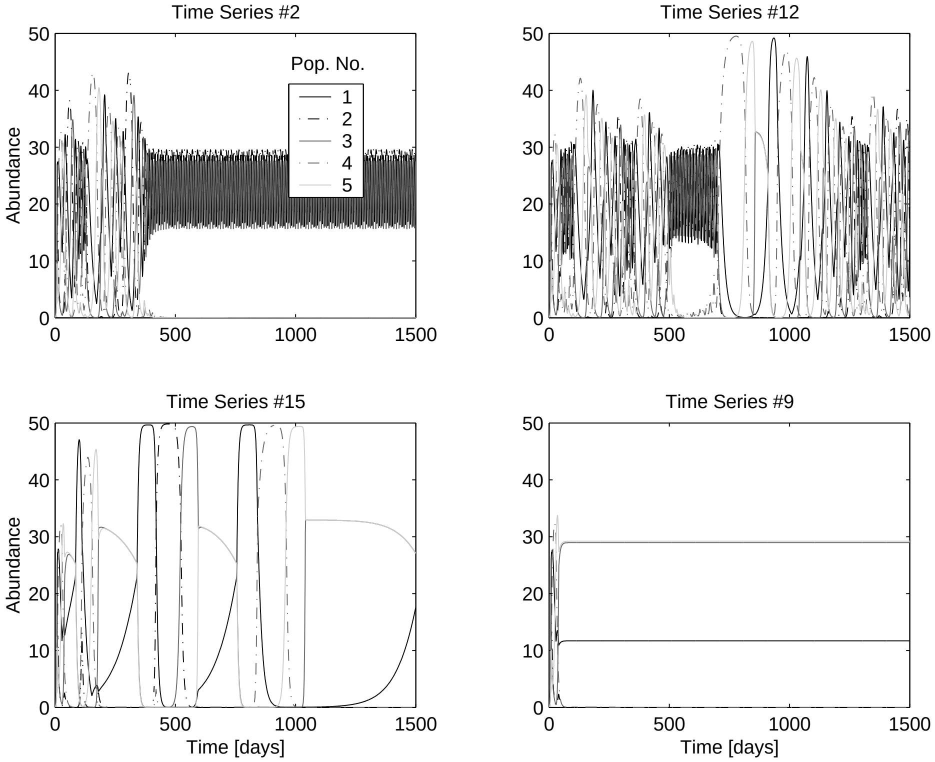

Figure 1: Time series of $\hat{P}_i (i = 1, \ldots, 5)$ created by parameter variation of the test model. Parameter settings $\{k_{21}, k_{25}\}$ are $\{0.2, 0.325\}$ for time series #2, $\{0.275, 0.4\}$ for time series #12, $\{0.325, 0.35\}$ for time series #15 and $\{0.275, 0.275\}$ for time series #9.

where h_{cs} is a neighbourhood function that asserts the convergence of the algorithm for $h_{cs} \to 0$ when $t \to \infty$. Mostly, the Gaussian function in dependence of $\|r_c - r_s\|$ is used, where $r_c \in \Re^z$ and $r_s \in \Re^z$ are the location vectors of nodes c and s. Additionally, h_{cs} is multiplied by the learning-rate factor $\alpha(t) \in [0, 1]$ that decreases monotonously over time to prevent the distortion of already ordered parts of the map at later time steps.

In measuring the quality of the SOM-mapping (see e.g. Bauer and Pawelzik [1992]; Villmann et al. [1997]) a compromise has to be made between an optimised reproduction of the data vectors and the minimisation of the topological distortion by neighbourhood violations. In this work the SOM-Toolbox 2.0 package (Vesanto et al. [1999]) was used that calculates the average quantisation error and the topographic error (Kiviluoto [1996]). The best network of different map configurations was assumed to minimise the sum of these two measures. This procedure tends to find solutions overfitting the dataset but this drawback was accepted as the details of the VQ step were found to be of minor importance for the following analysis.

3.1 Vector quantisation of the dataset

In advance of the analysis the data-matrix was standardised by mean and standard deviation of the individual variables ($\hat{P}_i \to P_i$). To incorporate information about control parameters of the competition model into the learning procedure, the constant time series of k_{21} and k_{25} were added as additional variables to the training dataset.

SOM networks of different map configurations were trained and the best network with 50 x 50 prototype vectors was found to explain 96.2% of the data variance.

4 NONLINEAR PROJECTION

Even though the SOM itself represents a kind of nonlinear projection technique it is not very well suited for the extraction of distinct modes of the underlying dynamics as the vectors spanning the SOM network can not be interpreted in terms of variable model entities. This limitation also exists for other unsupervised learning strategies that construct relevant topological constraints directly from the data (e.g. Martinetz and Schulten [1991]; Baraldi and

Alpaydin [2002]). Hence, the need for finding 'directions' along which features of the system vary continuously remains. A promising technique to extract these effective variables is nonlinear principal component analysis (NLPCA) put forward by Kramer [1991].

The NLPCA relies on an autoassociative feedforward neural network as depicted in Figure 2. It projects data-vectors $\mathbf{x}(t)$ onto a so-called bottleneck layer u and compares the decoded vectors $\mathbf{x}'(t) = (x'_1(t), x'_2(t), \ldots, x'_n(t))^T$ with the input data to minimise the cost function $J = \langle \|\mathbf{x}(t) - \mathbf{x}'(t)\|^2 \rangle$.

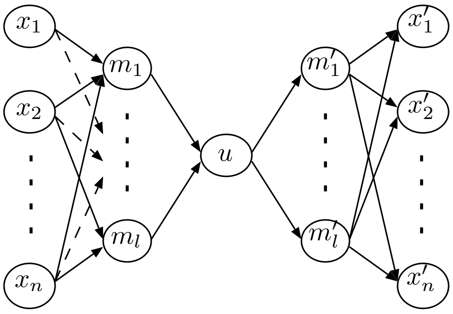

Figure 2: Example for an autoassociative neural network with l nodes $m_1, \ldots, m_l$ in the first and $m'_1, \ldots, m'_l$ in the second hidden layer.

The mappings $\mathbf{x} \rightarrow \mathbf{m}$ and $\mathbf{m}' \rightarrow \mathbf{x}'$ are typically performed by nonlinear transfer functions (e.g. the sigmoidal function), whereas mappings from and to the nonlinear principal component $\mathbf{u}$ use the identity function. The number of nodes in the hidden layers of the network determines the approximation quality of the data.

Typical problems arising during neural networks training are overfitting and local minima in the cost function J. In our analysis we employ the NeuMATSA (Neuralnets for Multivariate and Time Series analysis) package (Hsieh [2001]) where multiple runs and penalty terms for the network weights smooth the nonlinear responses of the transfer functions to obtain results less sensitive to local minima. Only by the data reduction of the preceding SOM analysis the NLPCA step is made applicable. Thus, an immense speed-up of the minimisation of the neural networks' weights is gained and the already smoothed SOM representation additionally accounts for the avoidance of local minima in the cost function.

4.1 NLPCA of the SOM-filtered data

To prevent the NLPCA from overfitting, 20% of the SOM-filtered dataset were chosen randomly as test-dataset and ensembles of 25 runs were selected for configurations of nodes in the hidden layers ranging from one to five. The analysis was terminated when the quality of the mapping as quantified by the mean squared error (MSE) for the test set decreased subsequently to an initial rise.

After extraction of the first nonlinear PCA, further components were iteratively found by subtracting earlier solutions from the SOM dataset and by repeating the analysis using the residuals.

The first nonlinear mode found this way explained 61%, whereas the second and third mode accounted for 16.5 and 2.7% of the SOM networks' variance, respectively. Thus, the dataset can be assumed to be essentially two-dimensional and the first two nonlinear modes extracted by NLPCA can be interpreted as effective variables (EV) of the underlying model. Figure 3 shows the first two modes in state space $\{P_1, P_2\}$. Clearly, variation of the original model variables is constrained indicating implicit model trade-offs and the existence of a reduced EV representation.

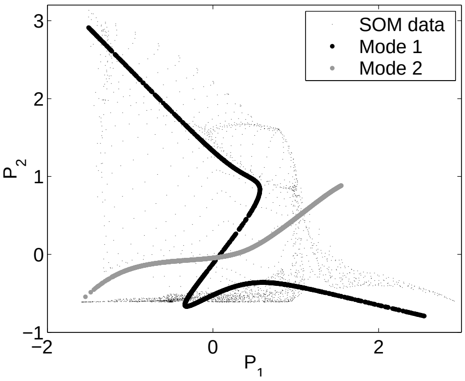

Figure 3: Nonlinear PCA modes 1-2 in a VQ state subspace with dots representing SOM-filtered model data.

4.2 Interpretation of modes

Figure 4 shows successive segments of an example time series projected into the EV space. The typical cycles with varying periods found in the dataset (see Figure 1) are clearly separated from each other and resemble the form of limit-cycles in the complex model's phase-space. Successive changes between these cycles illustrate the ability of the method to separate dynamical states.

To further investigate and interpret the effective nonlinear modes in terms of former model variables, the projected data was aggregated into bins of equal size with a minimum of five datapoints per class. Figure 5 shows the class distributions of the first mode.

The smooth course of the distributions together with the relatively small inner class variability even for densely covered bins (e.g. small positive values of the first nonlinear PCA) may enable a meaningful interpretation of the nonlinear modes. For example, a comparison of the chaotic transition (from small negative to small positive values of PCA 1 in Figures 4 and 5) for P_i provides an insight into the particular case when species coexist. This type of coexistence can thus be imagined as an occupation of 'dynamically separated' niches.

5 DISCUSSION

Improvements of the methodical parts of this work, as discussed in section 4 for the SOM algorithm, can be thought of. As outlined in Malthouse [1998], NLPCA solutions for the projection problem are only suboptimal and alternatives like the Principal Curves approach of Hastie and Stuetzle [1989], for example, can be tested as well. In this work however, the projection discrepancy does not constrain the usefulness of NLPCA as smooth solutions following mean features of the dataset are explicitly requested.

First outcomes of this work show that a combination of Vector Quantisation and nonlinear projection can already provide valuable insights into the dynamics underlying process-oriented models. The extraction of relevant nonlinear modes describing a model on a higher or aggregated level is a first step towards effective variable models that are easier to use and better to interpret than their complex model equivalents.

The results shown here point to the existence of non-linear but nonetheless simple transformations of former model state variables and parameters to effective variables. The projections of quantised model data along the first two nonlinear modes, from which only a subset is presented in Figure 5, already support a piecewise linear transformation from the original space of model entities to new aggregated variables. In future studies reduced-form models will be formulated using effective variables as provided by the approach put forward in this study. We will thereby rely on results and techniques presented here comprising (i) the simultaneous incorporation of model outcomes and varied model coefficients into the analysis, (ii) internal trade-offs between model variables for different attractors of the model dynamics and (iii) the smoothness of the projections of a small set of effective variables to the original model space. The extraction of these nonlinear transformations constitutes an analytical means to interpret the nonlinear principal components in terms of simulated processes as an essential step towards a reduced-form representation of complex, mechanistic models.

REFERENCES

Baraldi, A. and E. Alpaydin. Constructive feedforward art clustering networks - part ii. *IEEE Transactions on Neural Networks*, 13(3):662–677, 2002.

Bauer, H.-U. and K. Pawelzik. Quantifying the neighborhood preservation of self-organizing feature maps. *IEEE Transactions on Neural Networks*, 3(4):570–579, 1992.

Beven, K. On modelling as collective intelligence. *Hydrological Processes*, 15:2205–2207, 2001.

Ebenhöh, W. Stability in models versus stability in real ecosystems. *Senckenbergiana Maritima*, 27: 251–254, 1996.

Fodor, I. A survey of dimension reduction techniques. Technical Report UCRL-ID-148494, Lawrence Livermore National Laboratory, 2002.

Hastie, T. and W. Stuetzle. Principal curves. *Journal of the American Statistical Association*, 84(406): 502–516, 1989.

Hsieh, W. Nonlinear principal component analysis by neural networks. *Tellus*, 53A:599–615, 2001.

Huisman, J. and F. J. Weissing. Biodiversity of plankton by species oscillations and chaos. *Nature*, 402:407–410, 1999.

Huisman, J. and F. Weissing. Fundamental unpredictability in multispecies competition. *The American Naturalist*, 157(5):488–493, 2001.

Kiviluoto, K. Topology preservation in self-organizing maps. In *Proceedings of IEEE International Conference on Neural Networks*, volume 1, pages 294–299, 1996.

Kohonen, T. *Self-organizing maps*. Springer, Berlin, 2nd edition, 1997.

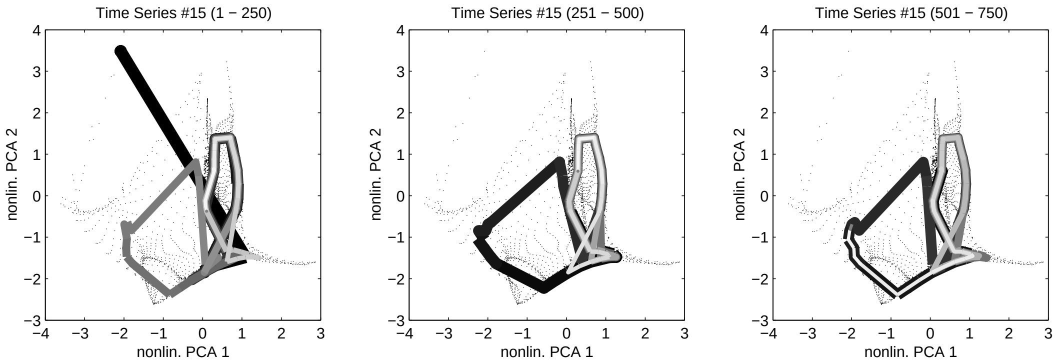

Figure 4: Plot of time series examples in EV space (spanned by the first two NLPCA modes). Shown are the first 750 time steps of series #12. Earlier time steps are drawn in thick dark and later ones in thin light lines.

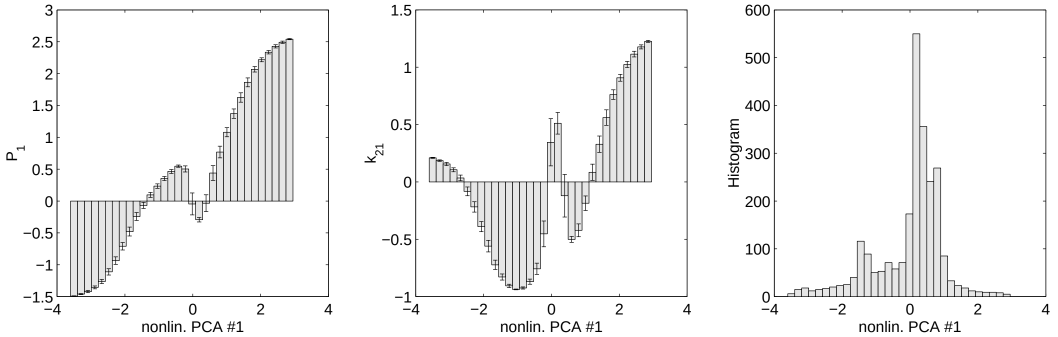

Figure 5: Aggregation of the projected data along the first nonlinear mode together with a histogram of bin occupancy.

Kramer, M. Nonlinear principal component analysis using autoassociative neural networks. *AIChE Journal*, 37(2):233–243, 1991.

Malthouse, E. Limitations of nonlinear pca as performed with generic neural networks. *IEEE Transactions on Neural Networks*, 9(1):165–173, 1998.

Martinetz, T. and K. Schulten. A ”neural-gas” network learns topologies. In *Artifical Networks*, pages 397–402. North-Holland, Amsterdam, 1991.

Vesanto, J., J. Himberg, E. Alhoniemi, and J. Parhankangas. Self-organizing map in matlab: the SOM toolbox. In *Proceedings of the Matlab DSP Conference*, pages 35–40, Espoo, Finland, 1999.

Villmann, T., R. Der, M. Herrmann, and T. Martinetz. Topology preservation in self-organizing feature maps: exact definition and measurement. *IEEE Transactions on Neural Networks*, 8(2): 1291–1303, 1997.

Wirtz, K. A generic model for changes in microbial kinetic coefficients. *Journal of Biotechnology*, 97:147–162, 2002.

Wirtz, K. and B. Eckhardt. Effective variables in ecosystem models with an application to phytoplankton succession. *Ecological Modelling*, 92: 33–53, 1996.