PREDICTIVE PROCESS-BASED MODELING OF AQUATIC ECOSYSTEMS

Nina Vidmar[1], Nikola Simidjievski[2,3], Sašo Džeroski[2,3]
Faculty of Mathematics and Physics, University of Ljubljana, Ljubljana, Slovenia[1]
Jožef Sefan Institute, Ljubljana, Slovenia[2]
Jožef Stefan International Postgraduate School, Ljubljana, Slovenia[3]
e-mail: nina.vidmar@student.fmf.uni-lj.si, {nikola.simidjievski, saso.dzeroski}@ijs.si

ABSTRACT

In this paper, we consider the task of learning interpretable process-based models of dynamic systems. While most case studies have focused on the descriptive aspect of such models, we focus on the predictive aspect. We use multi-year data, considering it as a single consecutive dataset or as several one-year datasets. Additionally, we also investigate the effect of interpolation of sparse data on the learning process. We evaluate and then compare the considered approaches on the task of predictive modeling of phytoplankton dynamics in Lake Zürich.

1 INTRODUCTION

Mathematical models play an important role in the task of describing the structure and predicting the behavior of an arbitrary dynamic system. In essence, a model of a dynamic system consists of two components: a structure and a set of parameters. There are two basic approaches to constructing models of dynamic systems, i.e., theoretical (knowledge-driven) modeling and empirical (data-driven) modeling. In the first, the model structure is derived by domain experts of the system at hand, the parameters of which are calibrated using measured data. In contrast, the later uses measured data to find the most adequate structure-parameter combination that best fits the given task of modeling. In both approaches, models often take the form of ordinary differential equations (ODEs), a widely accepted formalism for modeling dynamic systems, allowing the behavior of the system to be simulated over time.

Equation discovery [1, 2] is the area of machine learning dealing with developing methods for automated discovery of quantitative laws, expressed in the form of equations, from collections of measured data. The state-of-the-art equation discovery paradigm, referred to as process-based modeling [3], integrates both theoretical and empirical approaches to modeling dynamics. The result is a process-based model (PBM) – an accurate and understandable representation of a dynamic systems.

The process-based modeling paradigm has already been proven successful for modeling population dynamics in a number of aquatic ecosystems, such as: lake ecosystems [4, 5, 6, and 7] and marine ecosystems [3]. However, these studies focus on obtaining explanatory models of the aquatic ecosystem, i.e., modeling the measured behavior of the system at hand, while modeling future behavior is not considered. In contrast, Whigham and Recknagel [8] discuss the predictive performance of process-based models in a lake ecosystem. However, either they assume a single model structure and focus on the task of parameter identification, or explore different model structures where the explanatory aspect of the model is completely ignored. The method proposed by Bridewell et.al [9] focuses of establishing robust interpretable process-based models, by tackling the over-fitting problem. Even though this method provides estimates of model error on unseen data, these estimates are not related to the predictive performance of the model, i.e., its ability to predict future system behavior beyond the time-period captured in training data. Most recently, the study of Simidjievski et.al [10] focuses on the predictive performance of process-based models by using ensemble methods. However, while their proposed ensemble methods improve the predictive performance of the process-based models, the resulting ensemble model is not interpretable.

In this paper we tackle the task of establishing an interpretable predictive model of a dynamic system. We focus on predicting the concentration of phytoplankton biomass in aquatic ecosystems. Due to the high dynamicity and various seasonal exogenous influences [6, 7], most often process-based models of such systems are learned using short time-periods of observed data (1 year at most). Note however, this short time-periods of data are very sparse, i.e., consist of very few measured values, thus, most often the measurements are interpolated and daily samples are obtained from the interpolation.

The initial experiments to this end, indicate that the predictive performance of such models is poor: While providing dense and accurate description of the observed behavior, they fail at predicting future system behavior. To address this limitation we propose learning more robust process-based models. We conjecture that by increasing the size of the learning data, more general process-base models will be obtained, thus yielding better predictive performance while maintaining their interpretability.

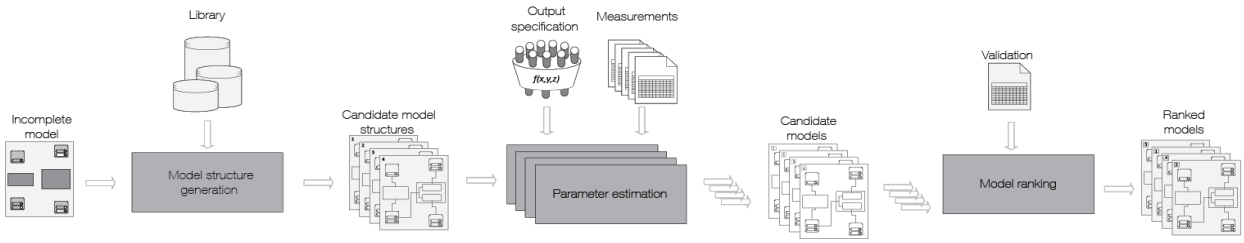

Figure 1: Automated modeling with ProBMoT.

The main contribution of this paper are the approaches to handling the learning data. The intuitive way of increasing the size of the learning data is by sequentially adding predeceasing contiguous datasets, thus creating one long time-period dataset, i.e., learning from sequential data (LSD). In contrast, when learning from parallel data (LPD), the model is learned from all the datasets simultaneously. Figure 2 depicts the both approaches. The two approaches, in terms of learning process-based models, are described in more detail in Section 3.

We test the utility of the two approaches on a series of tasks of modeling phytoplankton concentration in Lake Zürich. We use eight yearly datasets, using six for training, one for validation and one for testing the predictive performance of the obtained models. The aim of this paper is two-fold: besides validating the performance of the two approaches to handling data when learning predictive process-based models, we also test the quality of the training data. For that purpose, we perform additional set of experiments, similar to the previous. However, instead of using the interpolated data for learning the models – we use the original (sparse) measured values, thus examining the influence of the interpolation on the predictive performance of the process-based models.

The next section provides more details of the task of process-based modeling, and introduces a recent contribution to the area of automated process-based modeling, i.e., the ProBMoT [4, 10] platform. Section 3 depicts the task of predictive process-based modeling of aquatic ecosystems. Section 4 describes the data used in the experiments, the design of the experiments, and the task specification. Section 5 presents the results of the experiments. Finally, Section 6 discusses the findings of this paper and suggests directions for future work.

2 PROCESS-BASED MODELING AND PROBMOT

The process-based modeling paradigm, addresses the task of learning process-based models of dynamic systems from two points of view: qualitative and quantitative. The first, provides a conceptual representation of the structure of the modeled system. Still, this depiction does not provide enough details that would allow for simulation of the system's behavior. In contrast, the later, treats the process-based model as a set of differential and/or algebraic equations which allows for simulation.

A process-based model consists of two basic types of elements: entities and processes. Entities correspond to the state of the system. They incorporate the variables and the constants related to the components of the modeled system. Each variable in the entity has its role. The role specifies whether the variable is exogenous or endogenous. Exogenous variables are explanatory/input variables, used as forcing terms of the dynamics of the observed system (and are not modeled within the system). Endogenous variables, are the response/output (system) variables. They represent the internal state of the system and are the ones being modeled. The entities are involved in complex interactions represented by the processes. The processes include specifications of the entities that interact, how those entities interact (equations), and additional sub-processes.

From the qualitative perspective, the unity of entity and processes allows for conceptual interpretation of the modeled system. On the other hand, the entities and the processes provide further modeling details that allow for transformation from conceptual model to equations and therefore simulation of the system, i.e., providing the quantitative abilities of the process-based model. The equations define the interactions represented by the processes including the variables and constants from the entities involved.

The process-based modeling paradigm allows for high-level representation of domain-specific modeling knowledge. Such knowledge is embodied in a library of entity and process templates, which represent generalized modeling blueprints. The entity and process templates are further instantiated in specific entities and processes that correspond to the components and the interactions of the modeled system. These specific model components and interactions define the set of candidate model structures.

The algorithm for inducing models employs knowledge-based methods to enumerate all candidate structures. For each obtained structure, a parameter estimation is performed using the available training data. For this reason each structure is compiled into a system of differential and algebraic equations, which allows for the model to be simulated. In essence, this includes minimizing the discrepancy between the values of the simulated behavior obtained using the model and the observed behavior of the system.

Recent implementations of the PBM approach include Lagrame2.0 [11], HIPM [12] and ProBMoT (Process-Based Modeling Tool) [4, 10], which is next described.

The Process-Based Modeling Tool (ProBMoT), is a software platform for simulating, parameter fitting and

inducing process-based models. Figure 1 illustrates the process of automated modeling with ProBMoT. The first input to ProBMoT is a conceptual model of the modeled system. The conceptual model specifies the expected logical structure of the modeled system in terms of entities and processes that we observe in the system at hand. The second input is the library of domain-specific modeling knowledge. By combining the conceptual model with the library of plausible modeling choices, candidate model structures are obtained.

The model parameters for each structure are estimated using the available training data (third input to ProBMoT). The parameter optimization method is based on meta-heuristic optimization framework jMetal 4.5 [13], in particular, ProBMoT implements the Differential Evolution (DE) [14] optimization algorithm. For the purpose of simulation, each model is transformed to a system of ODEs, which are solved using CVODE ODE solver from the SUNDIALS suite [15].

Finally, the last input, is a separate validation dataset. In both cases (LSD and LPD), the model which has best performance on the validation dataset is the output of automated modeling process.

3 PREDICTIVE PROCESS-BASED MODELING OF AQUATIC ECOSYSTEMS

ProBMoT has been used extensively to model aquatic ecosystems [4, 5, 6]. Most of the case-studies, however, have focused on descriptive modeling – focusing on the content/interpretation of the learned models and not on their accuracy and predictive performance (with the exception of [10]). Predominately, models have been learned from short time-period (one-year) datasets, as considered long time-periods worth of data resulted in models of poor fit. These models, however, had poor predictive power when applied to new (unseen) data.

We use ProBMoT to learn predictive models of aquatic ecosystems from long time-period (multi-year) datasets. ProBMoT supports predictive modeling, as the obtained models can be applied/evaluated on a testing dataset. Taking the input/exogenous variable values from the test dataset, ProBMoT simulates the model at hand, and makes predictions for the values of the output/endogenous (system) variables. Using the output specifications, the values of the output variables of the model are calculated and compared to the output variables from the test set, thus allowing for the predictive performance of the model to be assessed.

Concerning the use of long time-period datasets, ProBMoT supports two different approaches, i.e., learning from sequential data (LSD) and learning from parallel data (LPD). The parameter optimization algorithm uses the available training data from the observed system to estimate the numerical values of the parameters. When learning from sequential data, illustrated in Figure 2a, ProBMoT takes as an input one training dataset. The training dataset is comprised of several contiguous short time-period datasets, thus the parameters are estimated over the whole time-span.

One the other hand, when learning from parallel data, depicted in Figure 2b, ProBMoT takes as an input several short time-period training datasets. The parameter optimization algorithm handles the short time-periods in parallel, i.e., it estimates the optimal model parameters by minimizing the discrepancy between the simulated behavior and each individual training set.

ProBMoT offers wide range of objective functions for measuring model performance such as sum of squared errors (SSE) between the simulated values and observed data, mean squared error (MSE), root mean squared error (RMSE), relative root mean squared error (ReRMSE), which is used in all experiments presented here for when learning the models and evaluating their performance. Relative root mean squared error (*ReRMSE*) [16] is defined as:

$$ReRMSE(m) = \sqrt{\frac{\sum_{t=0}^{n}(y_t - \hat{y}_t)^2}{\sum_{t=0}^{n}(\bar{y} - \hat{y}_t)^2}}, \qquad (1)$$

where n denotes the number of measurements in the test data set, $\hat{y}_t$ and y_t correspond to the measured and predicted value (obtained by simulating the model m) of the system variable y at time point t, and $\bar{y}$ denotes the mean value of the system variable y in the test data set.

The data on the aquatic systems are very sparse (e.g. measure on a monthly basis). In the above mentioned studies, often they have been typically interpolated and sampled at a daily interval. Here, to assess the effect of the interpolation to the performance of the models, we also consider using only the original measured values when establishing the predictive process-based model.

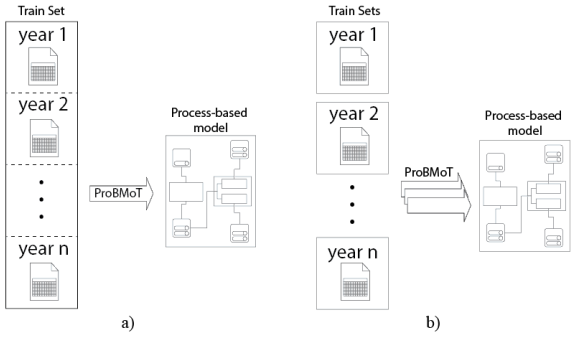

Figure 2: Two approaches to predictive modeling. a) Learning from sequential data (LSD), and b) Learning from parallel data (LPD).

4 EXPERIMENTAL SETUP

In this study, we apply the automated modeling tool ProBMoT to the task of predictive modeling of phytoplankton dynamics in Lake Zürich, Switzerland. We empirically evaluate the two different approaches for learning predictive models, LSD and LPD, on this task. We apply those two on interpolated data (sampled daily) and on the original (sparse) data.

4.1 Data & domain knowledge

The datasets used for our experiments were obtained from the Water Supply Authority of Zürich. Lake Zürich is a lake in Switzerland, extending southeast of the city of Zürich. It has an average depth of 49 m, a volume of 3.9 km^3 and a surface area of 88.66 km^2. The measurements consist of physical, chemical and biological data for the period from 1992 to 1999, taken once a month at 19 different sites, and averaged to the respective epilimnion (upper ten meters) and hypilimnion (bottom ten meters) depths.

The data were interpolated with a cubic spline algorithm and daily samples were taken from the interpolation [17]. Both the original and interpolated data from the first six years were used for training the models (1992-1997), data from year 1998 for validation and data from 1999 to estimate the predictive performance of the learned models.

The population dynamics model considered, consists of one endogenous/output (system) variable and multiple exogenous/input variables structured within a single ODE. The phytoplankton biomass is represented as a system variable, while the exogenous variables include: the concentration of zooplankton, dissolved inorganic nutrients (nitrogen, phosphorus, and silica) and two environmental influences of water temperature and global solar radiation (light).

The library for process-based modeling of aquatic ecosystems used in our experiments, is the one presented by Atanasova [18]. Particularly, to reduce the computational complexity of our experiments, we use a simplified version of the library which results in total of 128 candidate models.

4.2 ProBMoT parameter settings

For the parameter calibration procedure we use Differential Evolution with rand/1/bin strategy, 1000 evaluations over a population space of 50 individuals. For simulating the ODEs we use the CVODE simulator with absolute and relative tolerances set to 10^{-3}. For measuring the model performance we use objective function *ReRMSE*, described in Section 3. To further assess the significance of the differences in performance between the single dataset approach and multiple datasets approach we use Wilcoxon test for statistical significance [19] as presented by Demšar [20].

4.3 Experimental design

In this paper we compare the performance of the two different approaches (LSD and LSP) to learning predictive process-based models. For each approach we learn six process-based models using the available training data of six successive years (1992-1997). For both cases, we start with one short time-period training dataset (year 1997), and continue for five steps adding one preceding year to the training data set. At each step we learn the process-based models accordingly to the two approaches described in the previous section.

First, we apply this two approaches on the interpolated data, or more precisely, daily samples of interpolated data. Second, we apply the two learning approaches to the original (sparse) training data. In all of the experiments the validation dataset (year 1998) and the test dataset (year 1999) remain the same.

5 RESULTS

Table 1 summarizes the performance comparison between models learned from sequential data (LSD) and models learned from parallel data (LPD), using both interpolated (left-hand side) and original (right-hand side) training data. Note that, in both cases, learning from sequential data, yields better predictive performance than learning from parallel data. The results of the Wilcoxon test (in Table 1 below) shows that using LSD is better than using LPD, however, the difference in performance is not substantial nor significant (p-value=0.11).

Table 1: Comparison of the predictive performances (ReRMSE on test data) of models learned from sequential data (LSD) and models learned from parallel data (LPD), from both interpolated and original samples. The numbers in bold indicate the best result for the given years.

Train data (years)	Interpolated		Original	
	LSD	LPD	LSD	LPD
‘97	**1.398**	**1.398**	1.074	1.074
’96-’97	**1.099**	1.391	1.381	1.469
’95-’97	1.006	1.044	**0.984**	1.084
’94-’97	**0.986**	1.094	1.004	1.112
’93-’97	**1.075**	1.109	1.105	1.085
’92-’97	**0.934**	0.998	1.074	0.974
Wilcoxon test	LSD > LPD; p-value = 0.11		LSD > LPD; p-value = 0.11	

Next, as shown in Table 1, using the original measured values when learning the models, did not improve their predictive performance.

Finally, most importantly, from both experiments performed, we can conclude that using large amounts of training data (even interpolated) improves the overall predictive performance of the learned process-based models. Note however, that for one case (’93-97) the performance of the models does not improve. Further investigations are required to determine whether this phenomena is due to the quality of the data of that particular dataset (‘93), or to the dynamics of the system at that particular period significantly differing from the rest.

6 CONCLUSION

In this paper, we tackle the task of learning predictive interpretable process-based models of dynamic systems. In the process of establishing general and robust predictive models, we investigate learning from parallel data (LPD), in contrast to the state-of-the-art approach of learning from sequential data (LSD). We apply the both approaches to the task of modeling phytoplankton dynamics in Lake Zürich, using ProBMoT, a platform for simulating, parameter fitting and inducing process-based models. Additionally, besides validating the performance of the approaches to learning

predictive process-based models, we also test the quality of the training data by learning models from the original measured values, in contrast to learning models from daily samples of interpolated data.

The general conclusion of this paper is that using larger amounts of training data for learning process-based models yields improved predictive performance for tasks of modeling aquatic ecosystems. Both, Atanasova et al [5] and Taškova et al. [6] clearly state that one-year datasets produce models with poor predictive performance. We show that using data from a longer period, considered either consequently (LSD) or parallel (LPD) helps in deriving more general models, and therefore, better predictive models.

Even though the statistical significance comparison shows that the LSD approach has better performance than the LPD approach, the difference in performance is neither substantial nor significant. Nevertheless, when learning from sequential data, due to the mater of simulation and parameter optimization, the available training data considered for learning process-based models should be contiguous. On the other hand, one useful feature of the LPD approach is that can handle missing data (e.g. intermediate period with no measurements) for establishing robust process-based models.

Our empirical evaluation of learning from the original uninterpolated and sampled interpolated data, showed that the interpolation does not affect the performance of the learned process-based models. On the contrary, the models learned using the interpolated values yielded better performance than the ones learned using the original values. We conjecture that this is due to the sparsity of the original measured values (~ 12 time-points per year), which is insufficient to capture the dynamics of such a system. Moreover, considering the relative performance between the two approaches, the LSD approach performed insignificantly better than the LPD approach

Taken all together, some new questions arise for further investigation. How strongly the quality of measurements affects the results? Would the results change significantly in the case of ideal measurements? Considering this, possible directions for further work are as follows. First, performing more experiments using multiple parallel sets of data from different periods and, data from various different lake ecosystems should be used. In order to achieve more controlled experiments, we consider testing the presented approaches on synthetic data, that is, data obtained by simulating a well-established model of an arbitrary aquatic ecosystem. Finally, we would like to extend our approach to different ecosystems and other domains.

Acknowledgements

We would also like to acknowledge the support of the European Commission through the project MAESTRA - Learning from Massive, Incompletely annotated, and Structured Data (grant number ICT-2013-612944).

References:

[1] P. W. Langley, H. A. Simon, G. Bradshaw, J. M. Zytkow. Scientific Discovery: Computational Explorations of the Creative Processes. MA: The MIT Press, Cambridge, MA, USA. 1987.

[2] S. Džeroski, L. Todorovski. Learning population dynamics models from data and domain knowledge. *Ecological Modelling 170*, pp. 129–140. 2003.

[3] W. Bridewell, P. W. Langley, L. Todorovski, S. Džeroski. Inductive process modeling. *Machine Learning 71*, pp. 1–32. 2008.

[4] D. Čerepnalkoski, K. Taškova, L. Todorovski, N. Atanasova, S. Džeroski. The influence of parameter fitting methods on model structure selection in automated modeling of aquatic ecosystems. *Ecological Modelling 245* (0), pp. 136–165. 2012.

[5] K. Taškova, J. Šilc, N. Atanasova, S. Džeroski. Parameter estimation in a nonlinear dynamic model of an aquatic ecosystem with meta-heuristic optimization. *Ecological Modelling 226*, pp. 36–61. 2012.

[6] N. Atanasova, F. Recknagel, L. Todorovski, S. Džeroski, B. Kompare. Computational assemblage of Ordinary Differential Equations for Chlorophyll-a using a lake process equation library and measured data of Lake Kasumigaura. In: Recknagel, F.(Ed.), Ecological Informatics. Springer, pp. 409–427. 2006a.

[7] N. Atanasova, L. Todorovski, S. Džeroski, R. Remec, F. Recknagel, B. Kompare. Automated modelling of a food web in Lake Bled using measured data and a library of domain knowledge. *Ecological Modelling 194* (1-3), pp. 37–48. 2006c.

[8] P. Whigham, F. Recknagel, F. Predicting Chlorophyll-a in freshwater lakes by hybridising process-based models and genetic algorithms. *Ecological Modelling 146* (13), pp. 243–251. 2001.

[9] W. Bridewell, N. B. Asadi, P. Langley, L. Todorovski. Reducing overfitting in process model induction. In: Proceedings of the 22nd International Conference on Machine learning. (ICML '05). ACM, pp. 81–88. 2005.

[10] N. Simidjievski, L. Todorovski, S. Džeroski. Learning ensembles of population dynamics models and their application to modelling aquatic ecosystems. *Ecological Modelling (In Press)*. 2014.

[11] L. Todorovski, S. Džeroski. Integrating domain knowledge in equation discovery. In: Džeroski, S., Todorovski, L. (Eds.), Computational Discovery of Scientific Knowledge. Vol. 4660 of Lecture Notes in Computer Science. Springer Berlin, pp. 69–97. 2007.

[12] L. Todorovski, W. Bridewell, O. Shiran, P. W. Langley. Inducing hierarchical process models in dynamic domains. In: Proceedings of the 20th National Conference on Artificial Intelligence. AAAI Press, Pittsburgh, USA, pp. 892–897. 2005.

[13] J. J. Durillo, A. J. Nebro. jMetal: A Java framework for multi-objective optimization. *Advances in Engineering Software 42*, pp. 760–771. 2011.

[14] R. Storn, K. Price. Differential Evolution – A simple and efficient heuristic for global optimization over continuous spaces. *Journal of Global Optimization 11* (4), pp. 341–359. 1997.

[15] S. D. Cohen, A. C. Hindmarsh. CVODE, a stiff/nonstiff ODE solver in C. Computers in Physics 10 (2), pp. 138–143. Mar. 1996.

[16] L. Breiman. Classification and Regression Trees. Chapman & Hall, London, UK. 1984.

[17] A. Dietzel, J. Mieleitner, S. Kardaetz, P. Reichert. Effects of changes in the driving forces on water quality and plankton dynamics in three swiss lakes long-term simulations with BELAMO. *Freshwater Biology 58* (1), pp. 10–35. 2013.

[18] N. Atanasova., L. Todorovski, S. Džeroski, B. Kompare. Constructing a library of domain knowledge for automated modelling of aquatic ecosystems. *Ecological Modelling 194* (13), pp. 14–36. 2006b.

[19] F. Wilcoxon. Individual comparisons by ranking methods. Biometrics, 1:80–83. 1945.

[20] J. Demšar. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research 7*, pp. 1–30. Dec. 2006.