Inducing Process-Based Models of Dynamic Systems from Multiple Data Sets

Darko Čerepnalkoski[1], Katerina Taškova[1], Ljupčo Todorovski[2], and Sašo Džeroski[1]

[1] Jožef Stefan Institute, Jamova 39, SI-1000 Ljubljana, Slovenia
{darko.cerepnalkoski,katerina.taskova,saso.dzeroski}@ijs.si
[2] University of Ljubljana, Gosarjeva 5, SI-1000 Ljubljana, Slovenia
ljupco.todorovski@fu.uni-lj.si

Abstract. In this paper, we explore different modeling scenarios for inducing process-based models from multiple data sets. Namely, when modeling ecosystems, environmentalists expect a single model structure to explain system behavior among different yearly seasons, while the values of the constant model parameters may vary from season to season. We confront this modeling scenario with several others corresponding to multiple (one per season) structure or alternative single-structure models. The empirical evaluation and comparison of these scenarios on two tasks of modeling aquatic ecosystems confirm the expectations: single model structure can explain long-term system behavior, but the values of the model parameters for different seasons are significantly different.

1 Introduction

Models are useful analysis tools that scientists and engineers use to explain the observed and predict the future behavior of a system. A variety of modeling formalisms exist, ranging from purely qualitative ones, that focus on explaining the relations between system entities, to purely quantitative formalisms, based on equations, usually used to simulate/predict the future system behavior. Process-based models [4] integrate aspects of both qualitative and quantitative models. On one hand, a process model consists of processes that causally link system variables and entities, possibly through unobserved theoretical terms. On the other hand, each process has a quantitative model; when put together these models serve as an equation-based model that allows for a quantitative analysis (simulation and prediction) of system behavior.

Establishing a process-based model of an observed system involves two tasks. The first, often referred to as structure identification, is to identify the processes that govern the behavior of the observed system. Processes specify generic functional forms of the relations between system variables; when put together they define the structure of the model equations. The second task is to estimate the values of the model constant parameters that lead to an optimal match between the measured values of the system variables and the values obtained by simulating the model. Existing algorithms for process-based modeling combine heuristic search through the space of candidate model structures (constrained by domain knowledge) to identify model structure with non-linear optimization methods for estimating the values of the model constant parameters [8, 4].

In this paper, we deal with the problem of inducing process-based models from multiple data sets. When experts build a model from multiple data sets of the observed system collected over longer periods of time, they expect a single model structure to explain the overall behavior of the system. However, different values of the constant model parameters may be optimal for different data sets. A specific modeling task we address in this paper is modeling the change of population densities in an ecosystem. Ecologists collect data about an ecosystem over consecutive yearly seasons of and would expect a single model structure to explain the food web between the different populations (e.g., species in a lake), while the constant model parameters specifying the particular details of food web interactions can change from year to year. Existing algorithms for inducing process models do not support such modeling scenario, leaving ecologists with two choices. The first is to induce a single model with the same structure and parameter values for all data sets, which makes an unrealistic assumption that model parameters do not change. The second is to induce a separate model (both structure and parameters) for each data set. However, realistic modeling assumptions lead to an in between scenario: inducing a single model structure with multiple parameter settings, one for each of the corresponding data sets. We developed a modification of the Lagramge algorithm [8] for inducing process-based models that supports this scenario.

In the following section, we introduce the formalism of process-based models and briefly present the algorithms for inducing them with the focus on the changes necessary to support the specific modeling scenario for dealing with multiple data sets outlined above. Section 3 reports the results of the empirical evaluation of the method. In Section 4, we conclude the paper, put the presented method in the context of related research and outline further work.

2 Process-Based Modeling

Process-based models [4] integrate explanatory aspect of the qualitative models with quantitative equations that allow effective simulation and/or prediction of the (future) system behavior. When dealing with dynamic systems, scientists and engineers often refer to processes that govern system dynamics and entities that are influenced by those processes. Processes causally link system variables and entities, possibly through unobserved theoretical terms. To allow quantitative analysis of system behavior process models specify quantitative model for each process; when put together these models serve as a model that takes a form of ordinary differential equations.

Table 1 presents a process model of a phytoplankton growth in an aquatic environment (lake). The four processes explain the change in concentration of phytoplankton through time. Two other entities, environment *e* and inorganic nutrient *nitro*, have important impact on the phytoplankton dynamics. The

Table 1. A process model of a phytoplankton growth in an aquatic environment. The notation $\frac{d}{dt}$ X indicates the derivative of X with respect to time t.

```
model phytoplankton_dynamics
    entities phyto{primary_producer}, nitro{nutrient}, e{environment}

    process phyto_growth
        entities phyto
        equations $\frac{d}{dt}$ phyto.conc = 0.1 · phyto.growth_rate · phyto.conc

    process nitro_growth_limitation
        entities phyto, nitro
        equations phyto.growth_rate = nitro.conc / (nitro.conc + 5)

    process temperature_growth_limitation
        entities phyto, e
        equations phyto.growth_rate = (e.water_temp - 4) / (21 - 4)

    process phyto_loss
        entities phyto
        equations $\frac{d}{dt}$ phyto.conc = -0.5 · phyto.conc
```

Table 2. A generic entity and a generic process for modeling phytoplankton growth in any ecosystem.

```
 generic entity primary_producer
     variables conc{sum}, growth_rate{prod}
     constants max_growth_rate{0,Inf}

generic process primary_producer_growth
     entities P
     equations $\frac{d}{dt}$ P.conc = P.max_growth_rate · P.growth_rate · P.conc
```

phyto_growth process together with the following two, *nitro_growth_limitation* and *temperature_growth_limiation*, state the rate of phytoplankton growth and specifies how environment (in particular, water temperature) and the concentration of inorganic nutrient limit the growth. The process of *phyto_loss* refers to an unlimited exponential mortality of the phytoplankton population. These processes identify the main driving forces that influence the phytoplankton dynamics in the observed aquatic environment and the limiting factors for the phytoplankton growth.

In addition to this qualitative explanatory information, the process model from Table 1 specifies the quantitative (equation) models for each process. Combined together, these equations give the model of the phytoplankton growth. The issue that is immediate obvious is combining the influences of several processes on the same variable. By default, the influences are summed up, but other aggregation function can be specified by the expert. This specification is part of the background knowledge for building process models. Note that all the specific entities and processes in the model are instances of more general forms, generic processes and entities, that can apply to any ecosystem. These generic processes and entities serve as background knowledge for induction of process models. Table 2 includes examples of a generic entity and a generic process for modeling phytoplankton growth.

The generic process *primary_producer_growth* is the general form of the phytoplankton growth process; note the replacement of the constant parameter value 0.1 with a generic constant *P.max_growth_rate* and the entity *phyto* with a typed identifier *P*. The generic entity *primary_producer* corresponds to the specific entity *phyto* from the model of phytoplankton growth. The generic entity includes three properties that correspond to the (current) concentration, growth rate, and maximal growth rate of the primary producer. The first two properties vary through time, while the third correspond to a model constant parameter (the value of which should be between 0 and infinity, i.e., positive). Note that multiple influences on the variable *conc* are summed up, while those on the variable *growth_rate* are multiplied (declarations $\{sum\}$ and $\{prod\}$). Following these aggregation functions, we can combine the models of the individual processes from the model in Table 1 into the following differential equation:

$$\frac{d}{dt}\text{phyto} = 0.1 \cdot \frac{\text{nitro.conc}}{\text{nitro.conc} + 5} \cdot \frac{\text{e.water\_temp} - 4}{21 - 4} \cdot \text{phyto} - 0.5 \cdot \text{phyto}$$

where, for simplicity, we replaced *phyto.conc* with *phyto*. Given the initial concentration of phytoplankton, one can simulate this equation to produce trajectory that reflects phytoplankton dynamics.

After introducing the notion of process-based models, we can now present the task of inductive process modeling as:

Given

- observations of a set of continuous variables in consecutive time points (in the example above, this set would include concentrations of phytoplankton and nitrogen as well as water temperature);
- a set of entities expected to be included in the model;
- generic processes and entities specifying the modeling knowledge in the domain at hand;

Find a specific process-based model that explains the observed data (and predicts unseen data accurately).

There are several algorithms for inducing process-based models from time course data. Lagramge 2.0 [8] learns process-based models by transforming the modeling knowledge (generic processes and entities along with the specific model entities) into a grammar that specifies the set of candidate models for the particular task at hand. Lagramge than search this space and find an optimal model for the observed system behavior (time course data). IPM [4], on the contrary, performs heuristic search directly through the space of process-based models. Given the set of specific model entities, IPM generate all the possible instances of the generic processes and uses them as model components. In the next step, IPM searches through the space of model components combinations to find the optimal one. To avoid combinatorial explosion due to exploring all possible combinations, HIPM [7] introduces structural constraints specifying basic modeling rules in the domain at hand, such as "these two processes are mutually exclusive" or "these two processes should always be together in the model". All three algorithms employ standard non-linear least squares method [5] to fit the values of the constant model parameters against observed trajectories.

3 Handling Multiple Data Sets

In this paper, we explore different modeling scenarios that involve induction of process-based models from multiple data sets. In the domain of modeling aquatic ecosystems multiple data sets correspond to multiple ecosystem seasons. Ecologists expect the model for different data sets to have the same structure (same set of processes); only values of the constant parameters may change from year to year. This scenario is not directly supported by the inductive process modeling algorithms surveyed in the previous section. We altered the procedure for model evaluation in Lagramge, so it fits a separate set of constant parameter values for each data set, and sums up the model error estimates.

Having altered Lagramge, we have four modeling scenarios to explore. The first base-line scenario is to handle each data set separately and induce a separate process-model for each. In this case, we get process-models with different structures and parameter values; we refer to this scenario as multiple-structures, multiple-parameter-settings (MS-MP). The second scenario is the single-structure, multiple-parameter-settings one, we introduced in the previous paragraph. Following this scenario we get a single model structure that explains system behavior. The third scenario explores another way to get a single model structure for all the data sets: single-structure, single-parameter-settings (SS-SP) scenario. This approach handles multiple data sets as a single one and thus makes an (often unrealistic) assumption that model parameters do not change across data sets. Finally, we can also follow the SS-SP scenario to get the overall model structure and then fit the model parameters on each data set separately. This is an alternative SS-MP approach, that we will refer to as SS-MP*.

We empirically evaluate and compare the alternative modeling scenarios to two ecological tasks of modeling phytoplankton dynamics in lakes Bled [2] and Greifensee [1]. We have six data sets for Bled corresponding to one-year seasons from 1997 to 2002 and four data sets for Greifensee from 1988 to 1991. In both cases, we have regular measurements of concentrations of nutrients, plankton species, and environmental variables, such as water temperature and light intensity. Given these data, the task is to induce a process-based model for the change of the phytoplankton concentration. We compare the performance of the models obtained following the different modeling scenarios in terms of model accuracy and model complexity. We measure the model accuracy using the correlation coefficient between the measured values of phytoplankton concentration and the

Table 3. Coefficient of correlation between the measured values of phytoplankton concentration and the values obtained by simulating the four models (induced following the four modeling scenarios) on the one-year data sets for the lakes Bled and Greifensee.

	Lake Bled						Lake Griefensee			
Modeling scenario	1997	1998	1999	2000	2001	2002	1988	1989	1990	1991
MS-MP	0.98	0.93	0.91	0.97	0.98	0.99	0.70	0.92	0.94	0.93
SS-MP	0.88	0.89	0.95	0.96	0.95	0.98	0.62	0.84	0.94	0.95
SS-MP*	0.21	0.87	0.88	0.33	0.96	0.91	0.47	0.50	0.57	0.52
SS-SP	-0.19	-0.15	0.90	0.77	-0.80	0.46	0.37	0.55	0.72	0.51

Table 4. Complexities of the models (in terms of number of different processes) induced from multiple data sets for the lakes Bled and Greifensee.

Modeling scenario	Bled (1997-2002)	Greifensee (1988-1991)
MS-MP	22	15
SS-MP, SS-MP*, and SS-SP	9	8

values obtained with simulating the induced model. We measure the model complexity as number of different processes included in the induced models.

Table 3 compares the performance of the process-based models in terms of the correlation coefficient between the measured values of phytoplankton concentration and the values obtained by simulating the models induced with Lagramge. Note that high values indicate high model performance (the value of 1 indicates a perfect match between measurement and simulation). The results confirm the expectations: The SS-SP scenario does not work, since forcing a single parameter setting for all years leads to poor model performance. As one could expect, models that fit multiple structures (MS-MP) are better than those that fit a single one (SS-SP and SS-MP), with rare exceptions due to the difference between the heuristic used to guide search in Lagramge and the metric used to evaluate models. However, the performance difference between MS-MP and SS-MP is insignificant, while we observe a significant performance difference between SS-MP and SS-SP.[3] Note furthermore that the idea to fit the structure on all the data sets at once and then re-fit the parameter values for each data set separately (SS-MP*) does not really work and compares poorly to the more complex SS-MP scenario. In sum, a single structure can well explain the dynamics of phytoplankton growth in different seasons; however, seasons have strong influence on the values of the model parameters.

Table 4 compares the complexities of the models induced following the four scenarios. For the MS-MP scenario, we counted the number of different processes appearing in the model structures induced for different years, while for all the

[3] We performed single-tailed t-test with 90% significance threshold to test the statistical significance of the differences.

other (SS) scenarios, we counted the number of processes in the single structure induced. Again, the results are as expected: the complexity of the single-structure model is half the one of the multi-structure models.

4 Conclusion and Further Work

In this paper, we empirically evaluate and compare four modeling scenarios for inducing process-based models from multiple data sets on two tasks of modeling phytoplankton dynamics in lakes. The comparison confirms the initial hypothesis of environmentalists that there should be a single model structure capable of explaining the ecosystem dynamics in all the seasons; however model parameter can significantly change between the seasons. Furthermore, comparison shows that single-structure model is much simpler in terms of number of processes needed to explain system dynamics. Finally, another contribution is the development of a modified Lagramge algorithm for inducing process-based models to support such single-structure multiple-parameter-settings scenario.

Note that the issue of inducing models from multiple data sets have been previously addressed in [3]. However, there the focus is on integrating the structure of multiple process-based models induced from different samples of the same data set. In contrast, here we handle multiple data sets (not random samples of the same one) and explore the ways how to get single process-based model structure explaining all of them at the same time.

Immediate direction for further work is more elaborate empirical evaluation of the modeling scenarios. Note that in the preliminary experiments presented here, we have not tested the predictive performance of the models on unseen test data; these experiments would provide a proper and complete evaluation of modeling scenarios presented here. Further direction for further work is to analyze the relation between the complexity of the models obtained following the different scenarios and their performance. The minimum description length principle [6] can be used as a basis for developing a framework for principled comparison of process-based models induced from multiple data sets in terms of both performance and complexity.

Acknowledgments

We would like to thank Špela Remec Rekar, Johanna Mieleitner and Nataša Atanasova for providing the data on Lake Bled and Lake Greifensee.

References

1. Atanasova, N., Mieleitner, J., Džeroski, S., Todorovski, L., Kompare, B.: Construction of Lake Greifensee conceptual model combining machine learning and expert knowledge. In: Proceedings of International Conference on Ecological Modelling, pp. 262–263. Yamaguchi University, Japan (2006)

2. Atanasova, N., Todorovski, L., Džeroski, S., Remec-Rekar, Š., Recknagel, F., Kompare, B.: Automated modelling of a food web in lake Bled using measured data and a library of domain knowledge. Ecological Modelling Journal 194, 37-48 (2006)
3. Bridewell, W., Bani Asadi, N., Langley, P., Todorovski, L.: Reducing overfitting in process model induction. In: Proceedings of the Twenty-Second International Conference on Machine Learning, pp. 81–88. Omnipress, Madison, WI (2005)
4. Bridewell, W., Langley, P., Todorovski, L., Džeroski, S.: Inductive Process Modeling. Machine Learning Journal 71, 1-32 (2008)
5. Bunch, D. S., Gay, D. M., Welsch, R. E.: Algorithm717:subroutines for maximum likelihood and quasi-likelihood estimation of parameters in nonlinear regression models. ACM Transactions on Mathematical Software 19, 109–130 (1993)
6. Grnwald, P. D.: The Minimum Description Length Principle. MIT Press, Cambridge, MA (2007)
7. Todorovski, L., Bridewell, W., Shiran, O., Langley P.: Inducing hierarchical process models in dynamic domains. In: Proceedings of the Twentieth National Conference on Artificial Intelligence, pp. 892–897. AAAI Press, Pittsburgh, PA (2005)
8. Todorovski, L. Džeroski, S.: Integrating Domain Knowledge in Equation Discovery. In: Computational Discovery of Scientific Knowledge. LNAI, vol. 4660, pp. 69-97. Springer, Heidelberg (2007)