available at www.sciencedirect.com

ScienceDirect

journal homepage: www.elsevier.com/locate/ecolmodel

Application of automated model discovery from data and expert knowledge to a real-world domain: Lake Glumsø

Nataša Atanasova[a,*], *Ljupčo Todorovski*[b], *Sašo Džeroski*[b], *Boris Kompare*[a]

[a] *Faculty of Civil and Geodetic Engineering, University of Ljubljana, Slovenia*
[b] *Jožef Stefan Institute, Slovenia*

ARTICLE INFO

Article history:
Published on line 26 November 2007

Keywords:
Automated modeling
Lagramge
Modeling knowledge base
Lake Glumsø
Phytoplankton model

ABSTRACT

In this paper, we apply automated modelling method Lagramge to the task of modelling phytoplankton dynamics in Lake Glumsø, Denmark. The approach is based on integrating expert knowledge in the process of automated model induction from measured data. It supports modelling of ecosystem dynamics with ordinary differential equations by following the mass conservation law. The data set used in this paper comprises 2 years daily measurements of data needed for phytoplankton modelling in lake. In order to have sufficient data set for training and testing the models, the entire data set was divided in two parts, each containing 1 year of daily measurements. The expert knowledge supplied to Lagramge consists of elementary models of the basic ecological processes related to the food web dynamics and rules for combining elementary into complex models of the whole system. By applying Lagramge on Lake Glumsø we discovered a set of phytoplankton models that showed good fit on the training data set. The models were evaluated by simulating them on testing data set, which revealed good performance of the models.

1. Introduction

Lake ecosystems are complex dynamic systems. Modelling of such ecosystems is a great challenge to the scientists, who are progressively improving and making more and more complex models. In general, we distinguish between two basic approaches to mathematical modelling. Following the deductive approach (knowledge driven), the model is derived from the basic background knowledge (e.g. basic physical, chemical and biological principles) from the domain of use. The second, inductive approach (data driven) is based on exploring some space of candidate models and asses their accuracy against measured data. The model that fits measured data best is the result of the induction.

In this paper we apply an approach to modelling, which combines advantages of both the domain expert knowledge and induction from measured data. The domain knowledge is gathered in a knowledge library, which is used to guide the process of induction. The library consists of a set of elementary models, mainly descriptions of basic generic processes as well as rules for building complex models as interactions of these processes. Knowledge encoded in the library typically follows the basic principles in the domain of interest (Todorovski and Dzeroski, 2001; Langley et al., 2002; Todorovski, 2003). In the early days of the development of these tools (Todorovski and Džeroski, 1997), the knowledge had to be provided as an explicit definition of the space of candidate models. Now, these tools allow the user to provide higher-level domain

* *Corresponding author*. Tel.: +386 1 425 4052x111; fax: +386 1 251 9897.
E-mail address: natanaso@fgg.uni-lj.si (N. Atanasova).

doi:10.1016/j.ecolmodel.2007.10.032

knowledge about building mathematical models of complex real-world systems.

In this paper we focus on the application of the newly developed knowledge library about modelling of lake ecosystems (Atanasova, 2005; Atanasova et al., 2006a) on a real-world domain, i.e., Lake Glumsø, Denmark. The library comprises many known concepts in the domain of lake modelling, which can be found in literature (e.g. Jørgensen and Bendoricchio, 2001; DeAngelis, 1992; Chapra, 1997; and so on). Together with the automated modelling method Lagramge 2.0 the library was already successfully applied to other lakes, i.e., Lake Bled (Atanasova et al., 2006b) and Lake Kasumigaura (Atanasova et al., 2006c).

Lake Glumsø has been tackled with machine learning methods previously (Todorovski et al., 1998; Todorovski, 2003) with the earlier version of Lagramge, i.e., the version V 1.0 that required a hand crafted grammar has been used to discover a phytoplankton model (Todorovski et al., 1998). The same model was (re)discovered with the latest version V 2.0 of Lagramge (Todorovski, 2003), by using a simple knowledge library. Slightly different model was discovered by implementing a complex knowledge library (Atanasova, 2005). All of these experiments were performed on a small data set that did not allow for successful model evaluation. Just recently we obtained additional data for Lake Glumsø (Jørgensen, 2004), i.e., 2 years data set of daily measurements. This data set allows for correct model induction and successful model evaluation as well, i.e., inducing models on 1 year data and testing on the other year (unseen) data.

The paper is organized as follows: in the next section we briefly explain the method and the procedure of introduction of the expert knowledge about a specific ecosystem to the model discovery tool, i.e., Lagramge (if not specified, version V 2.0 is meant). In Section 3 we present the data set and the experiments performed. Section 4 gives the results and discussion. Finally, the conclusions are summarized in Section 5.

2. The method: automated modelling framework

2.1. *Conceptual modelling of dynamic ecosystems*

In order to compose a mathematical model of a dynamic ecosystem we typically start with setting a conceptual model of the system model. In the conceptual model, ecological modeller determines (1) the relevant variables in the system and (2) the bio-geo-chemical processes that connect these variables. In the next step the modeller selects mathematical formulations for the processes included, and finally estimates the constant parameters' values in the mathematical model (by hand or numerically, by fitting the parameters to the measurements). In the later phase the modeller evaluates the obtained model by performing its simulations on test data sets. However there is an important issue to be raised here, i.e., has the modeller selected the correct mathematical structure (and parameters) for the selected concept? This issue can be addressed systematically using computational methods for modelling.

2.2. *Lagramge method*

Lagramge is an automated modelling tool for a given conceptual model, perform heuristic searches for optimal mathematical model structure and optimal model parameters, that is structure and parameters that fit measurement data best. The search is performed based on a knowledge

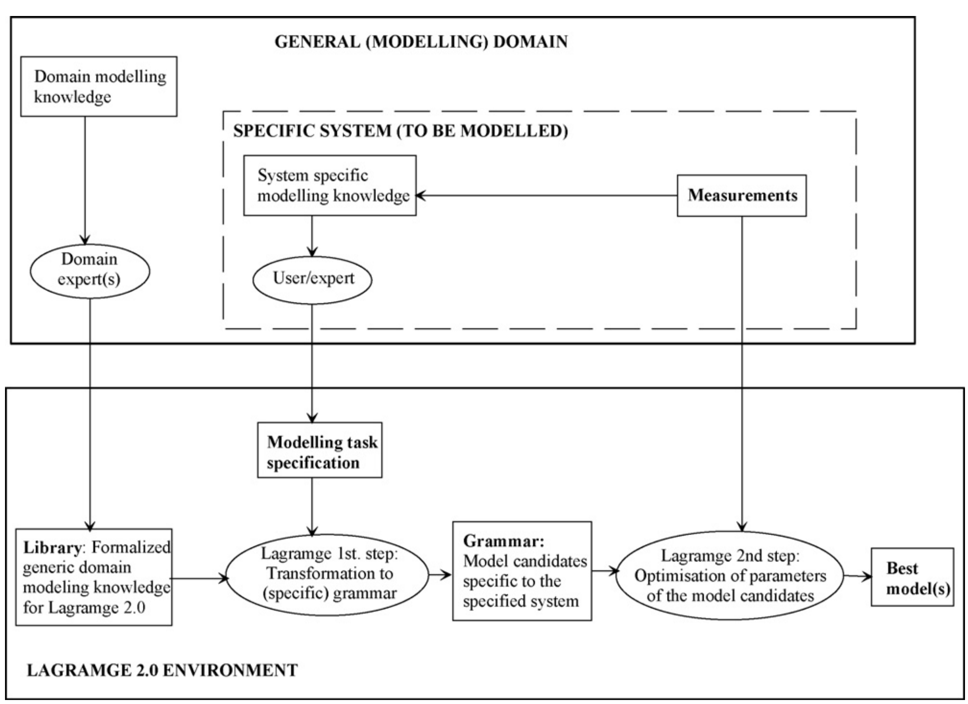

Fig. 1 – An automated modelling framework based on the integration of domain-specific modelling knowledge in the process of equation discovery.