Verh. Internat. Verein. Limnol. | 27 | 3626–3631 | Stuttgart, September 2001

Modeling and prediction of phytoplankton growth with equation discovery: case study – Lake Glumsø, Denmark

Boris Kompare, Ljupčo Todorovski and Sašo Džeroski

Introduction

In contrast with traditional modeling methods, which are used to identify parameter values of a (given) model with known structure, equation discovery systems identify the structure of the model as well. The model generated with such systems can give experts a better insight into the measured data and can also be used for predicting future values of the measured variables. This paper presents LAGRAMGE, an equation discovery system that allows the user to define the space of possible model structures and can also make use of domain-specific expert knowledge in the form of function definitions. We use LAGRAMGE to automate the modeling of phytoplankton growth (blooms) in Lake Glumsø, Denmark. The structure of the model constructed with LAGRAMGE agrees with human experts' expectations. The model can be successfully used for long-term prediction of phytoplankton concentration during algal blooms.

Tools that identify structure and parameters of dynamic systems

In a previous paper of this series, KOMPARE (1998) gave an example of a stationary (non-dynamic) system identification using tools that rely on regression (RETIS, FORS). Here we will discuss identification of fully dynamic systems. The task of modeling dynamic systems is to find a model that describes an observed dynamic behavior, i.e. development of the studied process(es) through time. A mathematical model of a dynamic system is usually a set of differential equations that specifies the change of system variables over time. Mainstream system identification methods, surveyed in LJUNG (1993), work under the assumption that the model structure, i.e. the form of the differential equations, is known. The task is then to determine the values of the constant parameters in the equations, so that the model best fits measured data. The structure of the equations is provided by the human expert and is based on the theoretical knowledge about the domain at hand (conceptual model).

Equation discovery systems, such as LAGRANGE (DŽEROSKI & TODOROVSKI 1993) and GOLDHORN (KRIŽMAN et al. 1995), do not assume a prescribed model structure, but rather explore a space of (possibly non-linear) equations. They help human experts to identify the structure of the model as well as the values of the constant parameters. Equation discovery systems can be used for automated modeling of ecological dynamic systems. KOMPARE (1995) used LAGRANGE and GOLDHORN to produce a model for predicting algal growth in the Lagoon of Venice. Several problems arise when using these systems for modeling experimental data. LAGRANGE discovered some equations predicting the optimal temperature for algal growth, but no good equations were discovered, from the viewpoint of what human experts expected. The reason for this was the high level of noise in the data. GOLDHORN incorporates methods for discovery from noisy data, so reasonable equations were discovered, while many equations of unacceptable structure were ranked as better fitted.

These problems led to the idea of restricting the space of possible equations considered in the process of discovery by taking into account the expert's knowledge of the domain at hand. In the area of machine learning, declarative language bias (DEHASPE & DERAEDT 1995) is used to specify the hypothesis space. In the task of equation discovery, this would be the space of all possible equations, or more precisely, the space of all possible equation structures. It has been observed that smaller hypothesis spaces lead to better performance of the learned concept (model) on a test set of unseen cases (NÉDELLEC et al. 1996).

In this paper, we present the LAGRAMGE equation discovery system. This name is a deliberate misspelling of the name of the equation discovery system LAGRANGE, the predecessor of LAGRAMGE. The letter N is replaced with M, so that the second part of the acronym reads **gram** as in **grammar**. Namely, declarative bias based on grammars is used

0368-0770/01/0027-03626 $ 1.50

in LAGRAMGE (Todorovski & Džeroski 1997) that uses context-free grammars as a formalism for specifying the form of discovered equations. The grammar can use the usual mathematical operators defined in C programming language, as well as additional functions defined by the grammar at hand. The grammar is specified according to domain-specific knowledge, and focuses the equation discovery process on equations with structure that is acceptable and comprehensible within the domain of use. The context-free grammar does not necessarily specify the precise structure of the model as in mainstream system identification methods, but can only be used to indicate the form of the expressions on the right side of the equations.

The equation discovery system LAGRAMGE

Problem definition

The problem of equation discovery, as addressed by LAGRAMGE, can be defined as follows.

Given:

- a context-free grammar $G = (N, T, P, S)$ and
- input data $D = (V, v_d, M)$, where $V = \{v_1, v_2,... v_n\}$ is a set of domain variables, $v_d \in V$ is the dependent variable and M is a set of one or more measurements. Each measurement is a table of measured values of the domain variables at successive time points (Table 1).

Find:

an equation for expressing the dependent variable v_d in terms of variables in V. This equation is expected to minimize the discrepancy between the measured and calculated values of the dependent variable. The equation can be either:

- differential, i.e. of the form $dv_d/dt = \dot{v}_d = E$ or
- ordinary, i.e. of the form $v_d = E$,

where E is an expression that can be derived from the context-free grammar G.

Table 1. Table of measured values of the domain variables at successive time points

Time	v_1	v_2	K	v_n
t_0	$v_{1,0}$	$v_{2,0}$	K	$v_{n,0}$
t_1	$v_{1,1}$	$v_{2,1}$	K	$v_{n,1}$
t_2	$v_{1,2}$	$v_{2,2}$	K	$v_{n,2}$
M	M	M	O	M
t_m	$v_{1,m}$	$v_{2,m}$	K	$v_{n,m}$

The declarative bias formalism

The syntax of the expressions on the right hand side of the equation is prescribed with a context-free grammar (Hopcroft & Ullman 1979). A context-free grammar (CGF) contains a finite set of variables (also called non-terminals or syntactic categories) each of which represents expressions or phrases in a language (in equation discovery, non-terminals represent sets of expressions that can appear in the equations). The expressions represented by the non-terminals are described in terms of non-terminals and primitive symbols called terminals. The rules relating the non-terminals among themselves and to terminals are called productions.

We denote a context-free grammar as a tuple $G = (N, T, P, S)$, where N and T are finite disjoint sets of non-terminals and terminals, respectively. P is a finite set of productions; each production is of the form $A \rightarrow \alpha$, where A is a non-terminal and α is a string of symbols from $N \cup T$. We use the notation $A \rightarrow \alpha_1 \mid \alpha_2 \mid ... \mid \alpha_k$ for a set of productions for the non-terminal A: $A \rightarrow \alpha_1$, $A \rightarrow \alpha_2, ...,$ $A \rightarrow \alpha_k$. Finally, S is a special non-terminal called starting symbol. The terminal $const \in T$ is used to denote a constant parameter in an equation that has to be fitted to the input data. The terminals v_i are used to denote variables from the input domain D. Finally, the non-terminal $v \in N$ denotes any variable from the input domain. Productions connecting this non-terminal symbol to the terminals v_i are attached to v automatically, i.e. $\forall v_i \in V$: $v \rightarrow v_i \in P$.

Expressions can be derived by grammar G from the non-terminal symbol S by applying productions from P. We start with the string w consisting of S only. At each step, we replace the leftmost non-terminal symbol A in string w with α, according to the production $A \rightarrow \alpha$ from P. When w consists solely of terminal symbols, the derivation process is over.

LAGRAMGE – the algorithm

Expressions generated by the context-free grammar G contain one or more special terminal symbols *const*. A non-linear fitting method is applied to determine the values of these parameters. A context-free grammar can, in principle, derive an infinite number of expressions (equations). LAGRAMGE thus uses a bound on the complexity (depth) of the derivation used to produce the equation (Todorovski & Džeroski 1997). The LAGRAMGE algorithm exhaustively or heuristically searches for the best equation (according to the selected heuristic function) within the allowed complexity (depth) limits. The whole procedure is described in detail in Todorovski et al. (1998) and Džeroski et al. (1999).

Lake Glumsø

Lake Glumsø (JØRGENSEN et al. 1986) is situated in a sub-glacial valley in Denmark. It is shallow with an average depth of about 2 m and its surface area is 266,000 m^2. For several years, it received mechanically and biologically treated waste water from a community with 3,000 inhabitants and a surrounding area which was mainly agricultural with almost no industry. The high nitrogen and phosphorus concentration in the treated waste water has caused hypereutrophication. The lake contained no submerged vegetation, probably due to the low transparency of the water and oxygen deficit at the bottom of the lake.

Concentrations of phytoplankton (*phyt*), zooplankton (*zoo*), soluble nitrogen (*nitro*) and soluble phosphorus (*phosp*) were considered relevant for modeling the phytoplankton growth. State variables were measured at 14 distinct time points, over a period of 2 months. The amount of measured data itself was far too small for (automatic, i.e. machine) equation discovery, so additional processing was applied to obtain a data set suitable for equation discovery (KOMPARE 1995, DEMŠAR 1996). First, dotted graphs of the measurements were plotted and given to three human experts to draw a curve that, in their own opinion, described the dynamic behavior of the observed system variable between the measured points. A properly plotted expert curve can be regarded as an additional source of reliable data. Curves drawn by the human experts were then smoothed with Besier splines. Finally, three new, more exhaustive data sets were obtained by sampling the splines derived from each of the three human experts' approximations at regular time intervals with time step $h = 0.1$ day. The dynamic behavior of the phytoplankton as represented by each of the three data sets is shown in Fig. 1.

Secondly, a new data set was obtained by applying a more sophisticated smoothing method to the graph plotted by the first human expert (see Fig. 2). The new data set also includes the measurements of the temperature of the water in the lake (*temp*). Due to the fact that the second bloom might not be described by the same model, the last portion of the data was not taken into account.

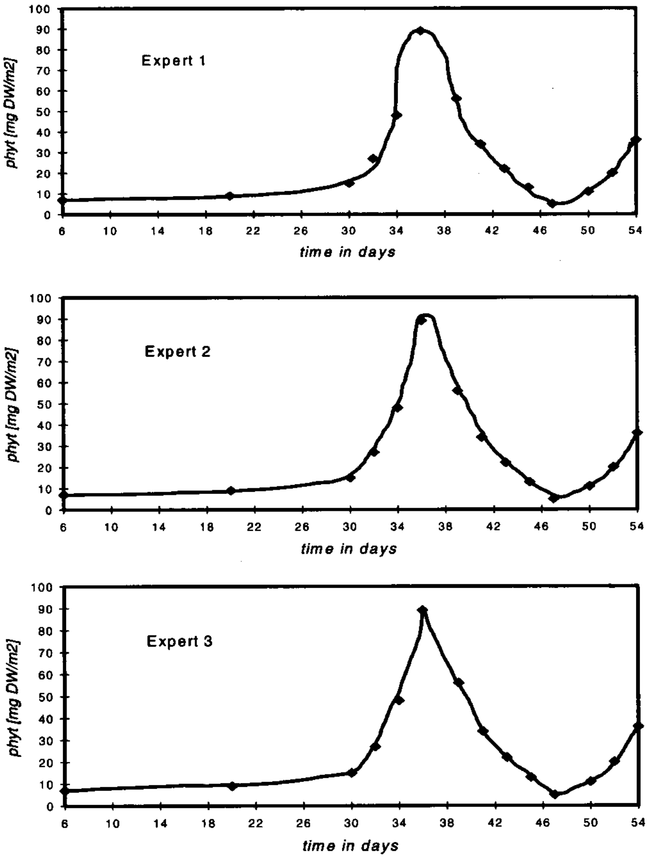

Fig. 1. Phytoplankton growth as seen by three domain experts.

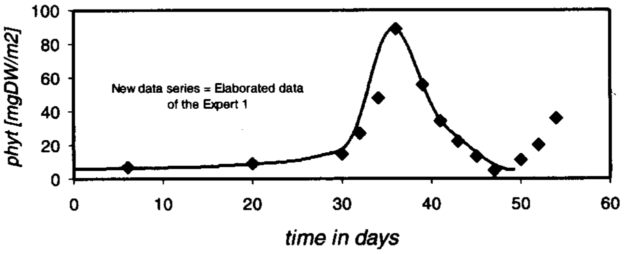

Fig. 2. New data set for phytoplankton growth, derived from the better smoothed curve of the first expert.

Results and discussion

The grammar that describes the algal growth was used in the experiments, taking into account ecological background knowledge on algal growth (MONOD 1942, JØRGENSEN 1986). Phosphorus and nitrogen are nutrients for phytoplankton and can thus appear in monod terms. Other terms describe the decay of phytoplankton (– *const* • *phyt*) and the feeding of zooplankton on phytoplankton (– *const* • *phyt* • *zoo*). At the maximum derivation depth 4 used in our experiments, 72 equations can be derived from the grammar. The values of the constant parameters in the equations specified by the grammar are constrained to be positive.

In a first set of experiments when we worked with three data sets we used the 'leave one out' testing method: LAGRAMGE was given two sets of data for equation discovery, and the best equation discovered was then tested on the remaining data set. The equation was tested on the task of predicting phytoplankton growth. The three equations obtained had the same structure (Eq. 1):

$$\frac{\partial phyt}{\partial t} = const_1 \cdot phyt \cdot \frac{ortP}{const_2 + ortP} - const_3 \cdot phyt \tag{1}$$

The structure of the equations discovered makes sense from an ecological point of view. It correctly identified that phosphorus is a limiting factor for phytoplankton growth in the lake.

We used the obtained equations for predicting the phytoplankton concentration in the lake on the testing set and then calculated the correlation coefficient between the measured and predicted values. The constant parameter values, as well as calculated correlation coefficients are shown in Table 2. It can be seen that all equations give accurate short-term predictions for phytoplankton growth. Note, however, the differences in the values of the equation coefficients, which indicate that experts approximated the dynamics of phytoplankton growth in quite different ways. Furthermore, we tested the robustness of the predictor on increasing the prediction period. The summary of the results (correlation coefficients between measured and predicted values) for prediction periods of 1, 2 and 5 days are given in Table 3.

Finally, we compared the accuracy of the obtained predictor with the accuracies of two simple predictors: no-change and same-change. Although simple, the two predictors have high correlation coefficients for short-term prediction and can thus be used as statistically significant predictors in the absence of better predictors, e.g. statistically derived black-box models, or regression formulae. The no-change predictor predicts that the value of the variable at the next time point will be the same as the present value

$$\hat{phyt}(t+h) = phyt(t) \tag{2}$$

The same-change predictor predicts the same change of the value of the variable, as the change in the previous time step

$$\hat{phyt}(t+h) - phyt(t) = phyt(t) - phyt(t-h) \tag{3}$$

The graphs in Fig. 3 show the dependence of correlation coefficients between the measured values and values predicted by the three different predictors for increasing prediction period for all data sets. The graphs show that the accuracy of the predictions decreases as the prediction time increases, which could be expected. The performance and robustness of all predictors is comparable. The same-change predictor has better performance than the one obtained with LAGRAMGE, especially on the third data set, but the LAGRAMGE predictor is more robust, i.e. has smaller oscillations of performance. The no-change predictor has the lowest accuracy on all data sets.

In the last experiment, we used the new data set, obtained from only the first bloom and also with more elaborated smoothing of the first expert's curve (Fig. 2). The grammar used in the experiments with the new data set was the

Table 2. Constant parameters' values and correlation coefficients for equations (Eq. 1) discovered by LAGRAMGE.

Training data sets	$const_1$	$const_2$	$const_3$	r
1, 2	0.617	0.101	0.442	0.9994
1, 3	0.763	0.080	0.592	0.9989
2, 3	0.383	0.444	0.155	0.9996

Table 3. Correlation coefficients between the actual phytoplankton concentration and the concentration predicted by the discovered equations (Eq. 1) for different prediction time periods.

Training data sets	1 day	2 days	5 days
1, 2	0.9836	0.9413	0.7243
1, 3	0.9849	0.9552	0.8137
2, 3	0.9853	0.9566	0.7267

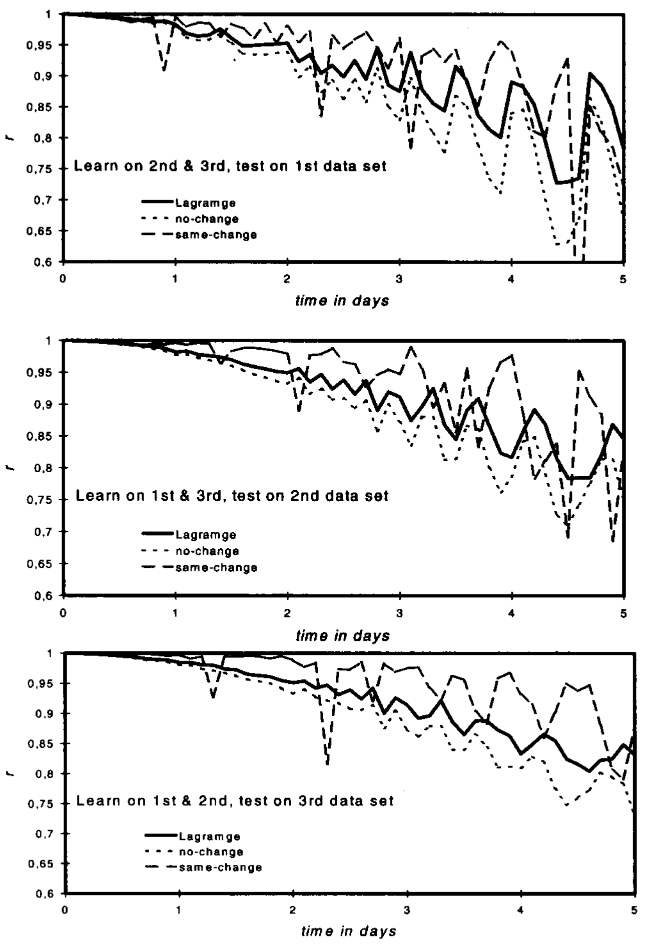

Fig. 3. Dependence of correlation coefficients between the measured values and values predicted with the three different predictors on increasing prediction period for all three original experimental data sets.

same as the one used in previous experiments, except that a new production was added to allow the use of temperature in the monod terms. The best equation discovered by LAGRAMGE that satisfies the constraint for the constant parameters' values is:

$$\frac{\partial phyt}{\partial t} = 0.553 \cdot Temp \cdot phyt \cdot \frac{ortP}{0.0264 + ortP} - 4.35 \cdot phyt - 8.67 \cdot phyt \cdot zoo \qquad (4)$$

Note that the structure of the equation discovered is similar to the structure of equations discovered from the first three data sets. It tells us that phosphorus is the limiting factor for phytoplankton growth in the lake, while phytoplankton concentration and water temperature are controlling the rate of phytoplankton growth. The equation also correctly identifies grazing of zooplankton on phytoplankton.

The graph in Fig. 4 shows the correlation coefficients between the measured values and the values predicted with three different predictors for the new data set as the prediction period increased. We can observe high accuracy and robustness of the predictor obtained with LAGRAMGE. The predictor obtained with LAGRAMGE is also suitable for long-term prediction of phytoplankton growth in Lake Glumsø.

Conclusions

We presented the automatic equation discovery system LAGRAMGE. In contrast with other system identification methods, where the structure of the model has to be provided (in advance) explicitly by the human expert, LAGRAMGE can use a more sophisticated form of representing the expert's theoretical (background) knowledge about the domain at hand. A context-free grammar can be used to specify a whole range of possible equation structures that make sense from the expert's point of view. Therefore, the discovered equations are in a comprehensible form and can give domain experts better or even new insight into the measured data. This distinguishes LAGRAMGE from other methods for automated modeling such as artificial neural networks and polynomial regression, which can be used for obtaining black-box models, i.e. models with incomprehensible structure. Additionally, if plenty of measurement data are available, less restrictive bias (more general equation space) can be used. On the other

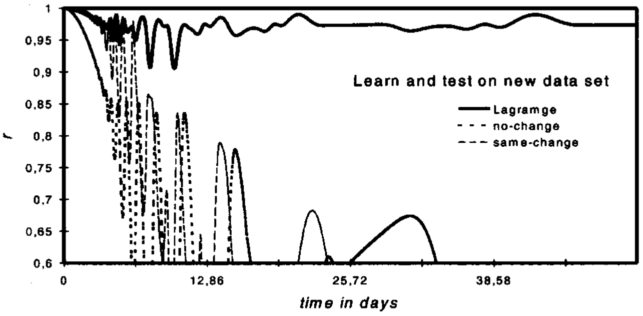

Fig. 4. Dependence of correlation coefficients between the measured values and values predicted with the three different predictors on increasing prediction period for the new experimental data set.

hand, when less data are available, more background knowledge should be used to restrict the space of equations and compensate for the lack of data.

Using background knowledge was essential in the task of modeling phytoplankton growth in Lake Glumsø on the basis of only 14 measurements within the period of 2 months. Even with such sparse measurement time points LAGRAMGE automatically discovered an acceptable and comprehensible model (from the experts' point of view) that can be successfully used for predicting phytoplankton growth in the lake. Experts' knowledge was used in this domain at two different levels. First, experts sketched the dynamic behavior of the observed system variables between the measurement points, which is regarded as an additional source of reliable data. Second, a context-free grammar was built using biological knowledge of population dynamics.

Acknowledgments

This work was supported in part by the Slovenian Ministry of Science and Technology under the project *ILP for Knowledge Discovery in Ecological Databases* and by the ESPRIT IV project *20237 ILP2*. We would also like to thank Professor SVEN-ERIK JØRGENSEN for providing the Lake Glumsø data. BORIS KOMPARE obtained a grant from the Rector of the University of Ljubljana to cover the costs of attending the SIL Conference in Dublin and presenting this paper.

References

DEHASPE, L. & DERAEDT, L., 1995: A declarative language bias for concept-learning algorithms. – *Knowledge Eng. Rev.* 7(3): 251–269.

DEMŠAR, D., 1996: *Experiments in automated modeling of ecological processes in Lake Glumsoe.* – B.Sc. Thesis, Faculty of Computer and Information Science, University of Ljubljana, Slovenia. (in Slovenian).

DŽEROSKI, S. & TODOROVSKI, L., 1993: Discovering dynamics. – In: *Proc. Tenth International Conference on Machine Learning.* – Morgan Kaufmann, San Mateo, CA 97–103.

DŽEROSKI, S., TODOROVSKI, L., BRATKO, I., KOMPARE, B. & KRIŽMAN, V., 1999: Equation discovery with ecological applications. – In: FIELDING, A. H. (ed.): *Machine Learning Methods for Ecological Applications*: 185–207. – Boston; Dordrecht; London: Kluwer Academic Publishers.

HOPCROFT J. E. & ULLMAN, J. D., 1979: *Introduction to Automata Theory, Languages, and Computation.* – Addison-Wesley, Reading, MA.

JØRGENSEN, S. E., 1986: *Fundamentals of Ecological Modelling*: 118–119. – North-Holland, Amsterdam.

JØRGENSEN, S. E., KAMP-NIELSEN, L., CHIRSTENSEN, T., WINDOLF-NIELSEN, J. & WESTERGAARD, B., 1986: Validation of a prognosis based upon a eutrophication model. – *Ecol. Model.* 32: 165–182.

KOMPARE, B., 1995: *The use of artificial intelligence in ecological modeling.* – Ph.D. Thesis, Royal Danish School of Pharmacy, Copenhagen, Denmark.

KOMPARE, B., 1998: Application of artificial intelligence to identify the key processes in a lake: case study – Lake of Bled. – *Verh. Internat. Verein. Limnol.* 26: 2370–2373.

KRIŽMAN, V., DŽEROSKI, S. & KOMPARE, B., 1995: Discovering dynamics from measured data. – *Electrotech. Rev. (Slovenia)* 62: 191–198.

LJUNG, L., 1993: Modelling of industrial systems. – In: *Proc. Seventh International Symposium on Methodologies for Intelligent Systems*: 338–349. – Springer, Berlin.

MONOD, J., 1942: *Recherches sur la croissance des cultures bacteriennes.* – Hermann, Paris. (in French)

NÉDELLEC, C., ROUVEIROL, C., ADÉ, H., BERGADANO, F. & TAUSEND, B., 1996: Declarative bias in ILP. – In: DERAEDT, L. (ed.): *Advances in Inductive Logic Programming*: 82–103. – IOS Press, Amsterdam.

TODOROVSKI, L. & DŽEROSKI, S., 1997: Declarative bias in equation discovery. – In: *Proc. Fourteenth International Conference on Machine Learning*: 376–384. – Morgan Kaufmann, San Mateo, CA.

TODOROVSKI, L., DŽEROSKI, S. & KOMPARE, B., 1998: Modelling and prediction of phytoplankton growth with equation discovery. – *Ecol. Model.* 113: 71–81.

Authors' addresses:

B. KOMPARE*, University of Ljubljana, Faculty of Civil and Geodetic Engineering, Hajdrihova 28, 1001 Ljubljana, P.O. Box 3422, Slovenia.
*Author to whom correspondence should be addressed.
E-mail: bkompare@fgg.uni-lj.si

L. TODOROVSKI, University of Ljubljana, Faculty of Medicine, Vrazov trg 2, 1105 Ljubljana, Slovenia. Present address: Jožef Stefan Institute, Jamova 39, 1111 Ljubljana, Slovenia.

S. DŽEROSKI, Jožef Stefan Institute, Jamova 39, 1111 Ljubljana, Slovenia.