

Modelling the growth of algae in the Lagoon of Venice with the artificial intelligence tool GoldHorn

Boris Kompare¹, Sašo Džeroski² & Viljem Križman²

¹ *University of Ljubljana, Institute of Sanitary Engineering,
Hajdrihova 28, SI 1001 Ljubljana, P.O.Box 3422, Slovenia,
E-mail: bkompare@fgg.uni-lj.si*

² *Jožef Stefan Institute, Jamova 39, SI 1111 Ljubljana, Slovenia,*

Abstract

Classical construction of ecological models follows one of the following two approaches: (1) either the measured data are analyzed with a statistical approach and a black-box statistical model is constructed, or (2) the model is deduced from basic (physical, chemical, etc.) principles and the measured data are used to calibrate and validate the model. A combination of the both ways is used very rarely, although it could combine the benefits of both.

Machine learning tools (ML), developed within the area of artificial intelligence (AI), are able to analyze a data base (perform data mining), using also some domain (expert) knowledge. As a result, ML tools can automatically, or with limited expert help, induce compact and easy to understand models from measured data. In this paper a description of the ML tool GoldHorn that induces differential or difference (algebraic) equations from measured time series data is given. Its successful application to the prediction of algae growth in the Lagoon of Venice, Italy, is also described in detail.

The model was induced using some background knowledge, as the tackled domain was too complex for data mining alone. Our experiments show that for noisy and complex domains inducing difference equations models could be more effective than inducing differential equations models. The induced (discovered) difference equation model for predicting algae growth in the Lagoon contains only one difference equation in contrast to classical models which contain several differential equations. Nevertheless, the simulation and predictive accuracy of the discovered model is comparable to the best deductionistic models of the Lagoon.

1 Introduction

Our companion paper (Kompare *et al.* [1]) describes two machine learning (ML) tools that construct regression trees from data only. These tools can analyze data, identify system structure and the discriminating values of the

important system parameters. The learned knowledge (constructed model) is represented in the form of a tree. We have described a successful application of these tools for predicting algae growth, i.e., system identification of the Lake of Bled. In this paper we will focus on ML tools which can discover differential, difference, or algebraic equations.

The first attempts to use AI in ecological modelling were applications of expert systems (ES) of the first generation (Kompore *et al.* [2]). In many cases ES did not give the expected results. One of the reasons for this was the knowledge acquisition bottle-neck. Namely, a knowledge engineer (computer expert) was supposed to extract the needed data and knowledge from a domain expert (ecologist) through a dialogue. This was a manual and consequently time consuming process. Due to the different backgrounds of the two experts, the efficiency of communication and knowledge acquisition was greatly reduced. The next generation of ES is attempting to automate knowledge acquisition as much as possible.

Today, a broad palette of AI techniques is available which can help in the knowledge acquisition process. The main issues addressed are system identification, knowledge representation, and model construction from given (measured) data or examples alone, or in combination with some basic domain knowledge. Some of the methods just represent data in a form more suitable for humans, but almost all of them seek for a more intelligent representation, i.e. perform learning. In fact, this is generation of domain knowledge from data. Automatic, or machine learning and deep knowledge representation are the main features built in the second generation of ES. One of the research fields in AI, within ES and ML, is machine discovery (MD) of mathematical description of dynamic systems. MD programs can find algebraic, differential, or difference equations.

In this paper we tried to test the suitability of MD tools for finding proper description of complex ecological systems. In the following, two MD tools will be briefly presented and illustrated with an example of their use for prediction of algae growth in the Lagoon of Venice, Italy.

2 The tools

2.1 Discovering Equations with LAGRANGE

An algorithm for discovering the dynamics of a system from a time series of measured relevant quantities was designed by Džeroski and coded in C by Todorovski (Džeroski & Todorovski [3, 4]; Todorovski [5]). The algorithm produces a set of algebraic and ordinary differential equations describing the system. The algorithm is implemented in the program LAGRANGE and is running on SUN workstations. Križman ([6]; Križman *et al.* [7]), see below, improved the algorithm, especially regarding robustness to noise in the data, and developed a PC version, called GoldHorn. We will refer to both programs as L&GH.

The LAGRANGE algorithm consists of three main stages (Džeroski & Todorovski [3, 4]):

1. Introduce time derivatives of the measured parameters.

The time derivatives are calculated from the given time series using numerical derivation. The quality of the derivatives (robustness to noise) depends on the derivation formula used, as shown by Križman [6] and Križman *et al.* [7], see below.

2. Recursively introduce new variables.

New variables are introduced by multiplying existing variables, i.e., the measured parameters and their time derivatives, calculated at the first stage. In each step, the depth (number of multiplicands) is increased by 1, and all possible combinations of multiplicands are constructed. The values of the new variables are calculated for all time points.

3. Generate equations and test on goodness of fit to measured data.

From the terms constructed in the previous step, linear equations with constant coefficients are constructed. The term with the greatest depth (complexity) is chosen as the dependent variable and is expressed as a linear combination of the remaining terms. The coefficients in the equations are obtained by linear regression. The suitability of an equation can be judged by the multiple correlation coefficient R . The equation is considered significant if $R > 1 - t_R$, where t_R is a user prescribed threshold. If the equation appears to be significant it is added to the model.

Džeroski & Todorovski [3, 4] made first experiments with LAGRANGE on synthetic, i.e. simulated data. LAGRANGE was then given the tabulated values of the simulated data. LAGRANGE discovered adequate models for several given problems from elementary physics (mass on a spring), through fluid dynamics (U-tube, cascaded tanks), through chemistry (chemical reactions, oscillating chemical reactor), to biological domains (microbial kinetics, population dynamics) etc. The most complex system identified was balancing an inverted pendulum (in its upper, labile equilibrium).

Attempts to solve real-life problems were unsuccessful due to noise in the data. The numerical derivation scheme used to introduce time derivatives of the measured quantities was amplifying the errors in the given data. A new approach was thus designed to tackle the noise in the data, implemented in the GoldHorn system (Križman [6], Križman *et al.* [7]).

2.2 Discovering Equations with GoldHorn

GoldHorn (Križman [6]; Križman *et al.* [7]) upgrades LAGRANGE in several directions.

1. First, it allows for introducing new variables through the use of the functions e^x , $\log(x)$, $|x|$, and $\text{sign}(x)$, in addition to $\sin(x)$ and $\cos(x)$ that are used in LAGRANGE.
2. Second, it expresses the highest order derivatives explicitly as rational functions of the system variables and their lower order derivatives. By doing

this, GoldHorn avoids the need to use the highest order numerical derivatives, which are most sensitive to noise. To estimate equation coefficients, as well as the quality of the explicit equations, GoldHorn uses numerical integration, rather than derivation. This procedure again reduces the error propagation due to noise.

3. In addition, GoldHorn allows the measured data to be pre-processed with digital filters that alleviate the effect of noise to a certain degree. Due care should be exercised when applying filters to the measured signals - applying an inappropriate filter may not only remove noise but also distort the original signal.
4. Finally, GoldHorn can be used to discover difference equations, thus eliminating the problems with numerical derivation. Instead of introducing $\dot{X}(t) = dX(t)/d(t)$, we introduce $X'(t) = X(t+h)$. In this case, numerical integration and derivation are not used.

GoldHorn is implemented in Pascal and runs on IBM PCs or compatibles (386 and above). It includes an interactive interface through which all tasks are done, i.e. (1) acquisition and plotting of measurements, (2) digital filtering of the data, (3) definition of new variables in terms of existing ones, using basic arithmetic operations and built in functions, (4) running the search for suitable equations, (5) sorting the constructed equations according to the coefficient of correlation R with data, or normalized deviation S , or the sum of squared error E , (6) post-processing the constructed equations with a built-in equation editor, etc.

3 Modelling algae growth in the Lagoon of Venice

The Lagoon of Venice measures 550 km², but is very shallow, with an average depth of less than 1 m. It is heavily influenced by anthropogenic inflow of nutrients - 7000 t/a of nitrogen and 1400 t/a of phosphorus (Bendoricchio *et al.* [8]). These loads are above the Lagoon's admissible trophic limit and generate its dystrophic behavior, which is characterized by excessive growth of algae, mainly *Ulva rigida*.

Four sets of measured data were available (Coffaro *et al.* [9]). The data were sampled weekly for slightly more than one year at four different locations in the Lagoon. Location 0 was sampled in 1985/86, locations 1, 2, and 3 in 1990/91. The sampled quantities are nitrogen in ammonia NH_3-N , nitrogen in nitrate NO_3-N , phosphorus in orthophosphate PO_4-P (all in µg/l), dissolved oxygen DO (in % of saturation), temperature $Temp$ (°C), and algae biomass Bio (dry weight in g/m²). In some experiments, we used the total nitrogen concentration N_{tot} instead of ammonia and nitrate nitrogen separately, as *Ulva* can use them both without greater difference as long as ammonia is not present in toxic concentrations (Coffaro *et al.* [9]; Bendoricchio *et al.* [8]).

We applied LAGRANGE and GoldHorn (L&GH) to the problems of predicting the total algae biomass and predicting the change of biomass from week t to the next week $t+1$. The independent variables were the measured values for

week t and week $t-1$, and in some cases week $t-2$. So besides the data for the current week t , we used one $(t, t-1)$ or two $(t, t-1, t-2)$ weeks of history. In the first experiments with LAGRANGE we used data of three stations as a learning set, and data of the remaining station as the testing set. Latter on we found out that the sampling stations are not similar, so only one station was used for learning. No test with independent set of data for the same station could have been done, as we had data for only one year, which is the minimum that covers the yearly dynamics of algae growth. Still, tests with cross-validation will be done in the future.

We started with the simplest case, trying to predict biomass for week $t+1$, knowing all the data for week t . Our first results with LAGRANGE were discouraging. We found out that LAGRANGE is not able to tackle real-life problems where noise is present in the data. The numeric derivation from noisy data produced too large errors. So the improved version of the LAGRANGE algorithm, GoldHorn, was used for further experimentation. Tests of GoldHorn on some real-life domains with moderate noise have given good results (Križman [6]). But the application on this really noisy domain, with estimated errors of the order of 20-50% for biomass, was still too much to discover plausible differential equations. As a by-product, some quite acceptable algebraic equations were found, e.g., law of conservation of mass in a rudimentary form of linear combination of given attributes, temperature dependence of growth (Arrhenius law), etc., but no acceptable dynamical equations.

Once we found out that the sampling sites are very diverse in their dynamics, we tried to tackle only one site (Station 0). The improvement was slight, almost negligible as compared to the improvement observed when we used regression trees (Kompore & Džeroski [10, 11]). We believe that the problem is for now too complex for GoldHorn. Namely, regression tools first split the problem in subproblems and then tackle them separately. L&GH tackle all the data as being one homogeneous set, i.e., describing the same phenomenon. Thus we tried to help the GoldHorn program by preparing the data in a more suitable way. We have taken the equations for nutrient limitation and growth dynamic from the conceptual model of Coffaro *et al.* [9] and Bendoricchio *et al.* [8] and calculated the growth rate μ and the mortality ω . The original equations and values of parameters in the conceptual model of Bendoricchio *et al.* [8] are given below:

$$\frac{dBio}{dt} = (\mu - \omega) Bio \quad (1)$$

$$\mu = \mu_{\max} \cdot f(I) \cdot f(N, P) \cdot f(T) \quad (2)$$

where $\mu_{\max}$ is the maximal growth rate [0.45 day^{-1}], and the f 's are specific functions (factors) limiting growth rate, i.e. $f(I)$ is the function of light intensity, $f(N, P)$ is the function of nutrients availability, and $f(T)$ is the temperature limiting factor, defined as follows:

$$f(I) = \frac{Fp}{Zm} \cdot \int_0^{Zm} \frac{I(z)}{I(z) + KmI} dz \quad (3)$$

$$I(z) = I_0 \cdot \exp\left\{-\left(K_0 + Kext \cdot Bio\right) \cdot z\right\} \quad (4)$$

$$f(N, P) = \min\left\{\left[1 - \left(\frac{QNmin}{N}\right)^{KN}\right], \left[1 - \left(\frac{QPmin}{P}\right)^{KP}\right]\right\} \quad (5)$$

$$f(T) = \exp\left\{-KT \cdot (T - Topt)^2\right\} \quad (6)$$

$$Topt(t) = Tmean + Tdev \cdot \left[1 - \cos\left(2\pi \frac{t - Hsft}{365}\right)\right] + Vsft \quad (7)$$

In the above equations Fp means daylength function [1], Zm is maximal depth [= 1 m], KmI is half-saturation constant for photosynthetic activity [= 0.0028 kcal m⁻² s⁻¹], I_0 is the function of daily insolation at the surface [kJ m⁻² day⁻¹], K_0 is extinction coefficient for the light in water [= 0.3 m⁻¹], $Kext$ is extinction coefficient in the presence of the algae [(g DW)⁻¹ m⁻¹], z is current depth [m], $QNmin$ is minimal intracellular quota of nitrogen [= 10 mg (g DW)⁻¹], KN is empirical exponent [= 3.5], $QPmin$ is minimal intracellular quota of phosphorus [= 1.1 mg (g DW)⁻¹], KP is empirical exponent [= 2.8], T is temperature of water [°C], KT is empirical exponent [= 0.007 °C⁻¹], $Topt$ is optimal temperature for algae growth [°C], $Tmean + Vsft$ is yearly mean temperature [= 5.1 + 15 °C] and $Tdev$ is its deviation [= 11.3 °C], $Hsft$ is horizontal shift of the time scale [= 105 day].

When calculating μ we did not take into account the function of light limiting $f(I)$ neither during the time of the year t nor along the vertical z . We considered this factor of minor importance as being almost constant at the surface during the year. Self shading of the algae, which is represented by extinction of light along vertical $I(z)$, could not be reproduced from the given measurements - it is purely expert background knowledge. So we decided to leave it out for the first experiments with the AI tools to see whether the AI tools can generate acceptable models without this notions. We namely wanted not to include the very specific background knowledge into the models as long as possible. Introduction of concepts for growth and mortality rate μ and ω is of course an application of background knowledge, but is far more general than light extinction and can be regarded as a way of data preparation (especially if we have in mind that L&GH have already discovered the Monod's relation on synthetic domains (Džeroski & Todorovski [3, 4], Križman *et al.* [7])). Because of the omission of the light limiting factor, we could also omit the constant factor μ_{max} , as the growth rate μ is not as accurately calculated as in the conceptual model.

Further, when calculating ω we did not take into account the term of toxic effect during anoxia, but only regular decay, i.e. we omitted the term $Ftox$ in the

following equation of Bendoricchio *et al.* [8]. The rationale to do so was that the oxygen was measured during daytime and close to the surface.

$$\omega = Dmax \cdot \Theta^{(T-30)} \cdot \frac{\frac{Bio}{f(I) \cdot f(N, P) \cdot f(T)}}{Kd + \frac{Bio}{f(I) \cdot f(N, P) \cdot f(T)}} + Ftox \quad (8)$$

where

$$Ftox = ToxMax \cdot \exp\{Ktox \cdot (T - 30)\}, \quad \text{if } DO < 5 \text{ mg/l}$$

$$Ftox = 0, \quad \text{if } DO > 5 \text{ mg/l}$$

and $Dmax$ is maximal mortality rate [= 0.5 day⁻¹], Θ is temperature adjustment coefficient [= 1.13], and Kd is half-saturation constant for mortality [= 600 g DW].

These two simplified background concepts, i.e. μ and ω , were then supplied to GoldHorn together with other (measured) data. This time GoldHorn found some well behaved differential equations describing the change of biomass, or total biomass next week. One of the generated equations is given as eq. 9 below (for the sake of simplicity we dropped the indexes (t) for the current week):

$$\frac{dBio}{dt} = \frac{1}{1 - 1.21\mu} \left[0.0122\mu \cdot Bio - 0.109\omega \cdot Bio - 0.4775 + 0.8604\omega - 89.39\omega^2 \right] \quad (9)$$

We were in fact expecting an equation of the form of eq. 1, i.e. $dBio/dt = \mu \cdot Bio - \omega \cdot Bio$, so we have to find plausible explanation for the other terms. The differential equations could not have been satisfactorily interpreted according to expert knowledge, although they gave acceptable predictions. We still have to investigate them, to possibly develop some plausible conceptual models from them. Perhaps the "surplus" terms are trying to describe the omitted light limitation and toxic terms. The discovered (algebraic) difference equations can be explained more easily. Eq. 10 is perhaps the best equation found by GoldHorn.

$$Bio(t+1) = -\frac{0.6111}{\omega} - 2077\omega + 0.6535DO + 0.6626Bio + 7.490Temp \quad (10)$$

In eq. 10 the term DO gives an indication that DO indeed influences the growth and consequently the mortality, as was conceptualized in the toxicity term of Coffaro *et al.* [9] and Bendoricchio *et al.* [8].

We simulated the equations generated by GoldHorn to visually check their behavior. Among the best for one week ahead prediction is eq. 10, which is the best for one year ahead prediction. One year ahead prediction is carried out by

using at the time t the calculated values of biomass at time $t-1$; the other attributes were given as measured. This kind of simulation reveals possible inner instabilities of the constructed formula. The simulation runs are given in Fig. 1.

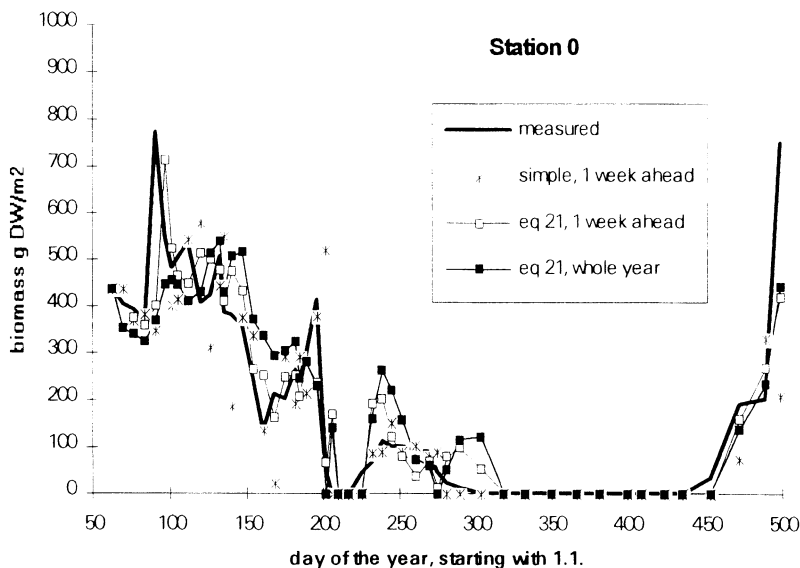


Figure 1: A diagram showing the predicted and measured total algae biomass in the Lagoon of Venice at Station 0. Prediction is made using eq. 10 (on Figure denoted by 21), discovered by GoldHorn on the training set of Station 0 only.

4 Discussion & conclusions

We have described two machine learning programs, LAGRANGE and GoldHorn, their theoretical background and examples of their use. The learned knowledge is represented by a model, consisting of algebraic or differential equations.

On real life domains LAGRANGE had serious problems (Džeroski & Todorovski [3, 4]). They are mainly due to the noise in the data which increases through the procedure of numerically calculating the derivatives of the given measured functions. Indeed, without properly filtering the data, no model could have been constructed. As a consequence of this bitter experience GoldHorn was constructed which to a great extent alleviates the influence of the noise in the data with several mechanisms, e.g. built in filters, expressing the highest order derivative in an explicit way, estimating the coefficients with integration, rather than derivation, etc.

As the programs construct new variables by multiplying existing ones, very complex formulae can be found given only the relevant measured behavior. Still, some concepts as fractions, e.g. the Monod's term, have to be given

explicitly into the background knowledge of the programs, if we want to obtain exactly the same-looking equations as were used for integration. Otherwise, mathematically analogous equations are found - but these equations are mostly difficult to recognize as appropriate by the domain experts. Some work will still be needed to improve the representation of the constructed equations in a more suitable form.

The application to modelling algae growth in the Lagoon of Venice shows the efficiency of the GoldHorn program and its suitability for use in scientific modelling, especially when tackling complex problems such as environmental modelling. The domain experts confirmed that the generated equations capture important properties of the underlying systems and processes which are comparable and/or complementary to expert knowledge built through years of work and study. Although the equations are far less complex than the conceptual model of the comparable behavior (Bendoricchio *et al.* [8]; Coffaro *et al.* [9]), the experts accepted the constructed equations - at least they are not contradictory to their knowledge or experience.

As compared to earlier experiments in predicting algae growth in the Lagoon of Venice with machine learning techniques (Kompere & Džeroski [10, 11]), we have recognized that the four stations have to be modelled separately in order to achieve good models. This resulted in the construction of satisfactory equations by GoldHorn, which was given intermediate concepts (growth and mortality rates) from a conceptual model (Bendoricchio *et al.* [8]; Coffaro *et al.* [9]).

Of course, having in mind that the models discovered by GoldHorn have been constructed from extensive analysis of the data plus from very limited basic knowledge of the modelled domain, we can not expect that the models will be very general. Nevertheless, the behavior of the discovered model, depicted in Fig. 1 clearly shows the usefulness of the discovered model for that particular site and conditions. This is from our point of view enough strong argument to be satisfied with the model and conclude that the used AI tools have great future in ecological modelling.

5 Acknowledgments

A part of this research was done within the TEMPUS JEP No. 4724 in Padua (Italy), Copenhagen (Denmark), and Ljubljana (Slovenia). The data for the Lagoon of Venice were provided by experts Prof. Giuseppe Bendoricchio, and Giovanni Coffaro (both Dept. of Industrial Chemistry, University of Padua, Italy). Special thanks are due to these experts for interpreting the constructed trees and equations, and giving us insights in their conceptual models. B. Kompere acknowledges the financial help of the Rector's Fund which facilitated participation to this Conference. S. Džeroski was supported in part by the Slovenian Ministry of Science and Technology and in part by the ESPRIT IV project 20237 ILP2.

Thanks also to the anonymous reviewers who helped to improve the final version of the paper.

6 References

1. Kompare, B., Džeroski, S., & Karalič, A., 1997. Identification of the Lake of Bled ecosystem with the artificial intelligence tools M5 and Fors. *Proceedings of the 4th International Conference Water Pollution 97: Modelling, Measuring and Prediction*, Bled, Slovenia, 18-20 June 1997.
2. Kompare, B., Bratko, I., Steinman, F., and Džeroski, S., 1994. Using machine learning techniques in the construction of models. Part I: Introduction, *Ecological Modelling*, 75/76, 1994, pp. 617-628.
3. Džeroski, S., and Todorovski, L., 1993. Discovering dynamics, *Proc. Tenth International Conference on Machine Learning*, pp 97-103, Univ. of Massachusetts, Amherst, June 27-29, 1993, Morgan Kaufmann, San Mateo, CA, 1993.
4. Džeroski, S., and Todorovski, L., 1995. Discovering dynamics: From inductive logic programming to machine discovery, *Journal of Intelligent Information Systems*, 4, 1995, pp. 89-108.
5. Todorovski, L., 1993. Modelling Dynamic Systems with Machine Discovery. B.Sc. Thesis, Univ. of Ljubljana, Faculty of Computer and Information Science, Ljubljana, Slovenia. In Slovenian.
6. Križman, V., 1994. *Handling noisy data in automated modelling of dynamical systems*, M.Sc. Thesis, University of Ljubljana, Faculty of Computer and Information Science, Ljubljana, 1994. In Slovenian.
7. Križman, V., Džeroski, S., and Kompare, B., 1995. Discovering dynamics from measured data. *Electrotechnical Review* (Special Issue on Application of Artificial Intelligence in Control Systems), Ljubljana, Slovenia, Vol. 62(3-4), pp. 191-198.
8. Bendoricchio, G., Coffaro, G., and De Marchi, C., 1994. A trophic model for *Ulva Rigida* in the Lagoon of Venice, *Ecological Modelling*, 75/76, 1994, pp. 485-496.
9. Coffaro, G., Carrer, G., Bendoricchio, G., 1993. *Model for Ulva rigida Growth in the Lagoon of Venice*, UNESCO MURST Research Project Venice Lagoon Ecosystem, Padua, 1993.
10. Kompare, B., and Džeroski, S., 1994. Two artificial intelligence methods for knowledge synthesis from environmental data. In P. Zannetti (Ed.), *Proceedings of Envirosoft '94, Computer Techniques in Environmental Studies V, Vol. II: Environmental Systems*, San Francisco, 16.-18. November 1994, Computational Mechanics Publications, pp. 265-272.
11. Kompare, B., and Džeroski, S., 1995. Getting more out of data: Automated modelling of algae growth with machine learning. *Proceedings of COSU '95, Coastal Ocean Space Utilization*, May 30 - June 2, 1995, Yokohama, Japan, pp. 209-220.