Mining co-variation patterns from ecological data: a process to aid the construction and validation of computer models

Florent Arthaud[1,2] and Serge Fenet[3] and Adriana Prado[3]

[1] Université de Lyon, CNRS, Université Lyon 1, LEHF, UMR5023, F-69622, France
florent.arthaud@univ-lyon1.fr
[2] ISARA-Lyon, 23, rue Jean Baldassini 69364 LYON CEDEX 07, France
[3] Université de Lyon, CNRS, Université Lyon 1, LIRIS, UMR5205, F-69622, France
serge.fenet@liris.cnrs.fr / adriana.bechara-prado@insa-lyon.fr

1 Ecological context

Shallow lakes and ponds represent a high proportion of aquatic ecosystems and, despite being often very well managed, they have been shown to be hot spots of biodiversity ([1]). Besides their direct contribution to biodiversity, the aquatic plants living in such lakes provide shelters, support, and food for many organisms. Therefore, they significantly contribute to an increase in the complexity of trophic chains and, consequently, in biodiversity. However, aquatic plants are potentially vulnerable to human activities and global changes, for example:

- eutrophication (increase of algae population, known as turbidity), which leads to an increase in competition pressure by phytoplankton and ultimately to the disappearance of vegetation;
- landscape fragmentation and decrease of connectivity between communities, which lead to erosion of biodiversity due to dispersal limitation;
- water deficiency, increasing the de-watering frequency and duration;
- fish farming, which ultimately leads to plant disappearance, as fishes uproot and graze plants.

Moreover, and in contrast to deeper lakes, the low depth of the water column, the absence of stratification, and the intense sediment-water interaction are the basis of several feedback loops that make shallow lakes particularly prone to explosive algae blooms that can rapidly wipe out most of the vegetation ([2]).

This paper proposes the employment of a data mining (DM) technique to aid the construction of a computer model of the fish-ponds network of the Dombes region (in France). The development of such computer model is the main task of a global project that links an ecology laboratory with a computer science laboratory. The goal is to study the dynamic interaction between phytoplankton and plants, and to assess the impact of agricultural and piscicultural practices on ecosystems, in order to stimulate good anthropogenic practices that maximize biodiversity, while maintaining food production.

2 From ecological data to a computer model

2.1 Available data

Data were collected in the Dombes region, in southeastern France, which is characterized by more than 1,000 man-made shallow lakes (average depth: 1m) organized in connected networks and with area varying from 1 to 100 ha. Nearly all these ponds have the same management policy, which consists in alternating fish farming and crop production. Thus, fish-ponds are regularly (every 5 to 7 years) emptied for one year, and the bottom is used for cereal cultivation. Naturally, this event leads to the disappearance of all aquatic vegetation. However, some stay dormant in the form of propagules during the drought period and can re-colonize the ponds when they are filled again.

In order to construct a computer model of the fish-ponds network, we gathered data from 90 ponds, for which 154 parameters were measured on a near-monthly basis from 2007 to 2010, and between April and October each year. In this work, we used between 23 and 68 of these parameters. They refer to:

- information regarding each pond and its management policies (surface, position in the pond network, fishing data, fertilization, etc);
- the established lake vegetation (in quadrats of 4 square meters), covering submerged and floating aquatic plants, for a total of 55 sampled aquatic plants;
- the levels of several chemicals (Nitrogen and various Nitrate compounds, Phosphorus, several Carbon compounds, pH, etc);
- the concentrations of up to 94 genera of bacteria and algae belonging to 7 families.

2.2 Modeling the equilibrium of plants and algae

Models regarding the dynamics of the transparency of shallow lakes are available from the literature (e.g., [3],[4]). Based on sets of partial differential equations, many of them promote the idea that the process of eutrophication occurs along a gradient that can be interrupted by catastrophic shifts, and that this process exhibits all the characteristics of non-linear dynamical systems (non-linear responses, regime shifts, basins of attraction, hysteresis, etc.). The final goal of our work is to explore this dynamical process. Such models, however, refer to at most 10 parameters: turbidity, re-suspended sediments, algae, water depth, nutrients, vegetation, waves, allelopathic substrates, fishes, and zooplankton. In our case, we have access to rich ecological data with more than 150 parameters that can be considered in the construction of our computer model. In order to identify the most pertinent ones, we want to extract information directly from the data, rather than rely on sole expert knowledge. To do so, we propose in this paper the employment of a data mining technique which is now described.

2.3 Data mining process: looking for co-variation patterns

Let D be our ecological database, where the attributes are different parameters described in the previous section and each record is composed by the corresponding parameter-values measured on a given month between 2007 and 2010, in a given pond.

The DM method used in this work is an extension of the one presented in [5]. Such method identifies sets of numerical attributes that behave similarly over the records of D. Generally speaking, the idea is the following: (1) first, the records are sorted in ascending order w.r.t every attribute: an index is associated with each attribute-value of each record to indicate its rank after sorting. That is, the smallest value of each attribute over all records gets index 1, and so on; (2) then, for a given set of attributes, we compute the ratio of the number of pairs of records for which all attributes in the set have the ranking index in the first record higher than in the second one or (3) some of them have the ranking index lower than in the second; (4) finally, when this ratio is significantly high, it indicates that the attributes in the set behave similarly and, therefore, represent an interesting pattern. The idea is to extract all such patterns with ratio higher than a given user-defined threshold, commonly referred to by the data miners as the "minimum support threshold" and, in this paper, the patterns are referred to as "co-variation patterns". An example of a co-variation pattern extracted from our data is "*The higher the level of total Nitrogen in the ponds, the higher the level of total Phosphorus*", which indicates that an increase of the parameter total Nitrogen causes an increase of total Phosphorus in a significant number of measurements (records). Another example pattern is "*The higher the level of total Nitrogen in the ponds, the lower the level of Nitrate*".

3 Results and Conclusion

Depending on the chosen minimum support threshold, the data mining process may generate many patterns describing positive or negative correlations in the analyzed data, and a difficult part of the process is to extract useful knowledge from these patterns. We are currently working on this interpretation process, but at the current time, however, they can be straightforwardly used to build a correlation hypergraph that explicitly describes the relationships between the parameters leading to indirect positive or negative feedback loops. As an example, one can observe in the "Bataillard" pond a positive correlation with a support of 0.53 (more than half of the records among all possible pairs, which is very high), between total Nitrogen, total Phosphorus and Calcium level, leading to a positive feedback between these 3 parameters. Once this mechanism has been identified, we may observe whether it is also present in other ponds, and also identify ponds that are "*the most representative*" of this process ("Épansardières", in our example), or those that "*differ the most*" ("Aubergères"). This gives us insights about specific local environmental parameters that could induce variations in a standard model.

Another example of an ecologically pertinent discovered relationship is the regulatory loop between the mineral nutrients (Phosphate, Nitrate and Ammonium), that can be either positive or negative, depending on the considered pond : increasing Phosphate leads to decreasing (or increasing) Ammonium, which in turn leads to increasing Nitrate in all the cases, which finally leads to decreasing (or increasing) Phosphate. This kind of feedback loop is particularly interesting, because Phosphate is the most important element in anthropogenic eutrophication.

When dealing with complex systems with non-linear multiple feedback loops, ecologists have to iteratively build, test, validate, and enhance their models until they describe the system of interest in its entirety. An example is the work presented in [3], which spans more than 11 years in the construction of an incrementally complex model of shallow lakes. Differently, we promote in this work the idea that the most pertinent parameters to be used by the model can be extracted directly from the available data, allowing us to build more accurate descriptions of the measured natural

dynamical systems. We showed that standard data mining techniques, such as mining co-variation patterns, may be of great help in the construction of more precise models. We are currently exploring the patterns obtained, in particular the construction of *association rules* (with left members implying a given right member) from the set of co-variation patterns, and on how to interpret these rules to improve model parametrization.

References

1. L. DeMeester, S. Declerck, R. Stocks, G. Louette, F. Van de Meutter, T. DeBie, E. Michels and L. Brendonck. *Ponds and pools as model systems in conservation biology, ecology and evolutionary biology*, Aquatic Conservation: Marine and Freshwater Ecosystems, 2005.
2. Scheffer M. *Critical Transitions in Nature and Society*, Princeton Studies in Complexity, Princeton University Press, 2009.
3. Scheffer M. Ecology Of Shallow Lakes, Population And Community Biology Series, Springer, 2005.
4. Egbert H. Van Nes and Marten Scheffer. *Shallow lakes theory revisited: various alternative regimes driven by climate, nutrients, depth and lake size*, Hydrobiologia, 2007.
5. Calders T, Goethals B and Jaroszewicz S. *Mining rank-correlated sets of numerical attributes*, In Proc. of ACM KDD'06, pages 96-105, 2006.