Analyzing water monitoring data with RCA-based approaches

Xavier Dolques[1], Agnès Braud[1], Corinne Grac[2,3], and Florence Le Ber[1]

(1) Université de Strasbourg, CNRS, ENGEES, ICube UMR 7357, F67000 Strasbourg
{dolques,agnes.braud}@unistra.fr florence.leber@engees.unistra.fr
(2) Université de Strasbourg, CNRS, LIVE UMR 7362, F67000 Strasbourg
(3) ENGEES, F67000 Strasbourg
corinne.grac@engees.unistra.fr

Abstract. This paper is a short feedback on a collaborative research work by computer scentists and hydroecologists. We have applied Relational Concept Analysis on complex data about running water characteristics (physical, biological and chemical parameters), to answer various questions. Two approaches are presented and discussed: the first one extracts patterns from temporal data, the second one extracts rules from a multi-relational dataset.

1 Introduction

For almost ten years, we have been working on watercourse monitoring data and, among other approaches, we have exploited Relational Concept Analysis (RCA) [8]. We have focused on temporal characteristics, that are inherent in monitoring data, but also on relations betweeen the different parts of the system being monitored (biological as well as physical and chemical characteristics of the watercourses). We have developed RCA-based approaches to extract temporal patterns from sequential data, or to extract rules from complex relational data, results being always assessed by domain experts. This paper is a short feedback on this collaborative research work. It is organized as follows. Section 2 describes the datasets, Sect. 3 gives a short explanation on RCA functioning. The approaches used are described in Sect. 4 and discussed in Sect. 5.

2 Data characteristics

Data were collected during the ANR Fresqueau project (2011-2015)[1] and organized into a database (60 million records and 20 gigabytes). The study area covers 161,100 km^2 in the east of France, for the 2000-2010 period. We collected five categories of water data from 21 different sources (mainly public data banks or research projects): 1) river quality data; 2) sampling site data; 3) hydrographic network data; 4) human activities data and 5) driving data.

[1] http://dataqual.engees.unistra.fr/fresqueau_presentation_gb

RealDataFCA'2021: Analyzing Real Data with Formal Concept Analysis, June 29, 2021, Strasbourg, France

River quality data are temporal data, and the most complex part of the data as detailed below. They are divided into three sub-categories: physical (e.g., the dimensions and shape of the river bed, characteristics of the substrate), physico-chemical (like pH, nitrate and phosphate, pesticides in the water or sediments) and biological data (i.e. lists of fauna and flora taxa, and metrics on these taxa, e.g., total abundance, diversity, and biological indices) for four groups: macroinvertebrates, diatoms, macrophytes and fishes. Biological indices are defined according to French standards (e.g., the French macroinvertebrate index, IBGN [1] for invertebrates). Taxa are associated to their life traits (like the type of respiration, or the habitat preference).

The other four categories of data are mostly geographic data. Sampling site data give the location of the sites where the river quality data were sampled and their main features. A sampling site is considered as a point. Hydrographic data comprise the different segments of running waters or waterbodies, the size of watersheds, administrative regions, and additional information on the hydrographic network. Human activity data allow to estimate anthropogenic pressures on running waters: land use, impediments to flow, location of discharges. Driving data concern forcing or context variables such as climate (for instance, average atmospheric temperature, precipitations), flows, geology or administrative information. They allow to characterize the environment of the running waters and sampling sites.

Building a database integrating data coming from different sources was not easy. For example, concerning taxa, although the data are based on a standard, their identifiers may evolve over time so that it is difficult to combine data from different years. Besides, geometric data may be imprecise: it may be hard to determine whether a sample site is placed on a watercourse or another one.

Then these data are highly heterogeneous in their values (quantitative continuous or discrete, semi-quantitative or qualitative), temporal variability (frequency and duration of sampling) and topological structure (with a geometry or not). They may be simple measures of a parameter (e.g., a pH value) or a complex index using different metrics (e.g., IBGN) or based on expert knowledge. For instance physico-chemical measures are collected 4 to 6 times per year, while biological samples are done at most once a year, and physical measures even more rarely. Moreover, some sample sites may require stronger monitoring, based on more parameters, in particular pesticides, so that some parameters are less abundant in the database. Let us also notice that depending on the area, taxa may differ.

We have proposed some approaches based on RCA in order to deal with some of these problems when analyzing the data, starting from questions asked by experts.

3 RCA basics

Relational Concept Analysis [8] is an extension of Formal Concept Analysis [6] which considers relational data, formalized within a *relational context family*

Table 1. Relational Context Family example.

object-attribute contexts

taxons	≤1 year	> 1 year
Athericidae	x	
Bithynia	x	x
Boreobdella		x

stations	small watercourse	fresh, running watercourse	phreatic stream
BREI0001	x	x	
BRUN001			x
FECH001		x	

object-object contexts

taxonPresence	Athericidae	Bithynia	Boreobdella
BREI0001			x
BRUN001	x	x	
FECH001	x		x

(RCF), $\mathbf{F} = \{\mathbf{K}, \mathbf{R}\}$ where $\mathbf{K}$ is a set of object-attribute contexts (each context corresponding to an object category) and $\mathbf{R}$ is a set of object-object contexts (relations between objects of various categories).

The principle of RCA consists in integrating object-object relations as new attributes (called *relational attributes*) in the formal contexts of $\mathbf{K}$ thanks to scaling quantifiers, such as the existential (*exist*) or universal strict (*exist+forall*) scaling quantifiers. It produces iteratively a set of concept lattices (one lattice per object category) interconnected through relational information. The concepts in a given lattice group objects according to the shared attributes and to the connections they have with objects of another category. The result is a family of concept lattices where concepts of a lattice are linked to concepts of other lattices. We illustrate this with a small example (from [4]). Table 1 introduces two object-attribute contexts, one about taxa and their life traits, one about river sites and their physical characteristics, and an object-object context linking taxa to the sites where they have been found.

First, the RCA process builds lattices on the two object-attribute contexts, `stations` and `taxons` (Fig. 1). Then the `stations` context is extended by relational attributes linking `stations` objects to `taxons` concepts, based on the `taxonPresence` context. For example, ∃`taxonPresence` : `Concept_2` means that at least one object of `Concept_2` in the `taxons` lattice is present on each `stations` object that owns this attribute. The lattice built on this extended context is shown in Fig. 1 (right). In this example, the process stops here since there is only one (one-way) relation linking the two contexts.

4 Questions and Approaches

For domain experts, the general question is to link physical and physico-chemical data to biological ones, the first ones giving an instantaneous information, the second one giving a long term integrative information. It covers more specific questions, for example, how can values of biological indices be explained by

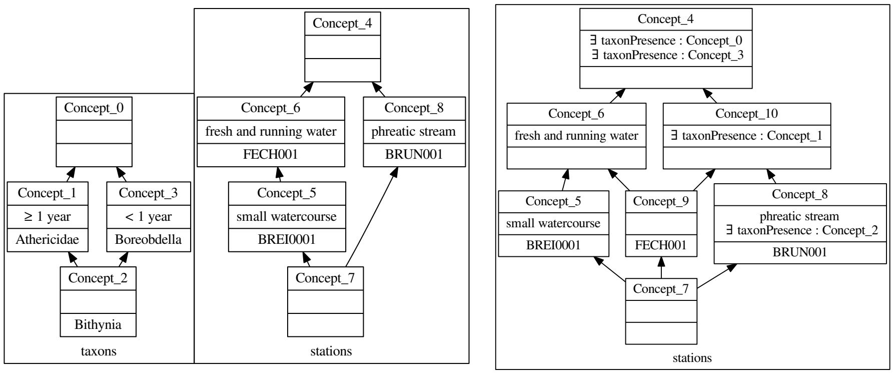

Fig. 1. Lattices of the two object-attribute contexts of Table 1 – initialization (left) and first step (right) of RCA – Only `stations` lattice changes

the preceding successive measures of physico-chemical parameters? What is the relation between the values of physico-chemical or physical parameters and the life traits of taxa living in a site? The first question raises the problem of dealing with temporality and it has been undertaken with a pattern mining approach [5] and then by RCA [9]. Moreover, working on biological quality requires to overcome the difficulty of analyzing sites with different taxa. This problem has been tackled by working with biological indices in the pattern mining approach, one of their aims being to make biological quality comparable between sites. For the second question, it has been tackled by working on life traits, and the question has been undertaken by a RCA-based rule mining approach [3].

Analyzing sequences of physico-chemical and biological measures. In [9], we focused on sequences of physico-chemical measures (6 measures per year) ending with biological samples (one per year). The selection and preprocessing of the data were done under the supervision of domain experts. Physico-chemical measures were discretized into qualitative scales. Biological samples were synthetized into qualitative indices (five levels from red to blue, corresponding to quality). The question was to explain the biological quality wrt the physico-chemical quality assessed during the last months. Sequences were encoded into an RCF, according to the schema shown in Fig. 2: each rectangle corresponds to an object-attribute context, while the arrows correspond to object-object contexts.

Then data are processed as follows. Firstly, RCA is applied to the RCF in order to obtain a family of concept lattices. Secondly, the interrelated concepts from the RCA result, are navigated to extract a set of sequential patterns for each concept of the `BioSamples` lattice. A pattern is actually a directed graph, where the various paths represent sequences of parameter values preceding a biological sample. Iceberg lattices [11] have been used to select patterns with the highest support (i.e. represented in many sample sites). Figure 3 shows an example of an extracted pattern: it summarizes a set of sequences of physico-chemical

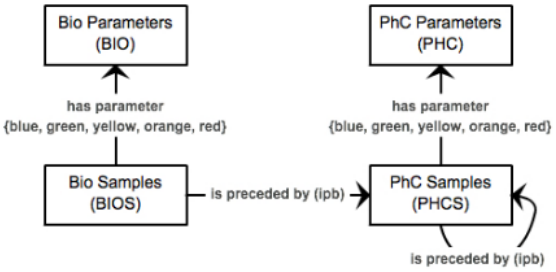

Fig. 2. Modeling sequences of physico-chemical and biological samples [9]

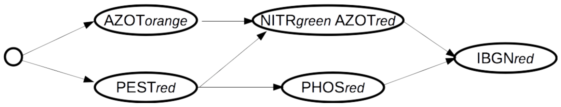

Fig. 3. An example of a sequential pattern

parameter values measured before a biological index (IBGN) with red value. The pattern is read as follows: in all sequences, an orange value for AZOT (nitrogen except nitrate) and a red value for PEST (pesticides) have been measured before a green value for NITR (nitrate) and a red value for AZOT occurring at the same moment. Also, a red value for PHOS (phosphorous) has been measured after the red value for PEST. According to expert domains, this pattern can be interpreted as follows: the quality values of physico-chemical parameters are consistent with the biological value, macroinvertebrates being sensitive to high rates of pesticides and ammonium.

Extracting rules linking taxa life traits and site physical and physico-chemical characteristics. In [3], we tried to connect physical characteristics of sample sites, physico-chemical parameters, and the traits of taxa living in sample sites. We therefore developed an RCA-based method using AOC-posets to deal with large datasets. This approach allowed to provide a reasonable number of concepts and to extract meaningful implication rules (association rules whose confidence is 1). In order to offer more flexibility on the quantification of taxa on sample sites than with the existing *exist* and *forall* scaling operators, new scaling operators were defined and experimented, providing different semantics for the rules. Data were modeled as shown in Fig. 4. Detailed temporal information (physico-chemical parameters) was aggregated into annual values and then discretized into five levels. Taxa numbers were also discretized (taxa weakly to highly represented). An example of rule is given below, where a percent-quantifier is used [3]: such a quantifier allows to build relational attributes with for example, the form $\exists\forall_{>n\%}r(C)$; an object owning this attribute is linked to at least $n\%$ of C objects

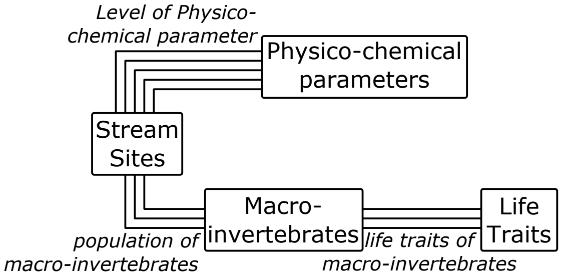
Level of Physico-chemical parameter
Physico-chemical parameters
Stream Sites
Macro-invertebrates
Life Traits
population of macro-invertebrates
life traits of macro-invertebrates

Fig. 4. Modeling the links between physico-chemical parameters and the life traits of taxa (macro-invertebrates) living in sample sites [3]

with the r relation [2].

$$\begin{array}{ll} S_{>50\%} & \text{high\_population}(\exists\text{strong\_affinity}(\texttt{slow current})) \\ \rightarrow & \exists\text{bad\_state}(\texttt{hydrology}) \end{array}$$

This rule means that a sample site having more than half of its highly represented taxons preferring slow current, has a bad hydrological state. According to the experts, it corresponds to small disconnected phreatic streams, with almost no current, that are specific of the Alsace plain.

5 Discussion and Conclusion

In the following we describe some problems we have faced with the data and novelties that stemmed from these works. Note that all these experiments led to new proposals to make the use of RCA easier [10].

First of all, ecosystems are very complex entities, influenced by many parameters. We have tried to integrate many of these parameters in our database, but handling all of them in a single analysis is not really possible, and even for experts to fully understand such results involving many information types. We have thus worked on subproblems, handling a smaller but consistent part of parameters, based on an expert question. Also, data like biological indices already integrate several parameters in one measure. Their definition has been proposed by groups of specialists and it is pertinent to use them when studying water quality, in particular to overcome the difference of taxa between areas, but at the same time they are not raw data and represent a kind of bias.

Regarding the pattern mining approach, given the sequence set of biological and physico-chemical samples, we found hierarchies of multilevel cpo-patterns that summarize the impact of physico-chemistry to biology. The hierarchical representation allows to enhance the analysis of the extracted set of sequential patterns. Nevertheless, there are too many patterns, and relevant interestingness measures must be chosen. Another problem is due to the irregular distribution of biological index values, leading to more or less frequent, complex, and informative patterns for the different values. Regarding the rule mining approach:

using AOC-posets causes loss of concepts, and some interesting rules will thus not be found. Other techniques should be explored for processing the complexity of this relational dataset.

With respect to classical approaches in the hydroecological domain, our approaches are original since most work are based on statistical analysis or supervised machine learning methods (see e.g., [7, 12]).

References

1. AFNOR: Qualité de l'eau : détermination de l'Indice Biologique Global Normalisé (IBGN). NF T90-350 (1992)
2. Braud, A., Dolques, X., Huchard, M., Le Ber, F.: Generalization effect of quantifiers in a classification based on relational concept analysis. Knowl.-Based Syst. **160**, 119–135 (2018)
3. Dolques, X., Le Ber, F., Huchard, M., Grac, C.: Performance-friendly rule extraction in large water data-sets with AOC posets and relational concept analysis. Int. J. General Systems **45**(2), 187–210 (2016)
4. Dolques, X., Le Ber, F., Huchard, M., Nebut, C.: Relational Concept Analysis for Relational Data Exploration. Adv. in Know. Disc. and Manag. **5**, 55–77 (2016)
5. Fabrègue, M., Braud, A., Bringay, S., Grac, C., Le Ber, F., Levet, D., Teisseire, M.: Discriminant temporal patterns for linking physico-chemistry and biology in hydro-ecosystem assessment. Ecological Informatics **24**, 210–221 (2014)
6. Ganter, B., Wille, R.: Formal Concept Analysis: Mathematical Foundations. Springer Verlag (1999)
7. Goethals, P.L., Dedecker, A., Gabriels, W., Lek, S., Pauw, N.: Applications of artificial neural networks predicting macroinvertebrates in freshwaters. Aquatic Ecology **41**(3), 491–508 (2007)
8. Hacene, M.R., Huchard, M., Napoli, A., Valtchev, P.: Relational concept analysis: mining concept lattices from multi-relational data. Ann. Math. Artif. Intell. **67**(1), 81–108 (2013)
9. Nica, C., Braud, A., Dolques, X., Le Ber, F., Huchard, M.: Exploring temporal data using relational concept analysis – an application to hydroecology. In: Concept lattices and their applications (CLA 2016). pp. 1–13. CEUR Workshop Proc. (2016)
10. Ouzerdine, A., Braud, A., Dolques, X., Huchard, M., Le Ber, F.: Adjusting the exploration flow in Relational Concept Analysis – An experience on a watercourse quality dataset. Adv. in Know. Disc. and Manag. **9** (2021)
11. Stumme, G., Taouil, R., Bastide, Y., Pasquier, N., Lakhal, L.: Computing iceberg concept lattices with Titanic. Data & Know. Eng. **42**(2), 189 – 222 (2002)
12. Villeneuve, B., Souchon, Y., Usseglio-Polatera, P., Ferréol, M., Valette, L.: Can we predict biological condition of stream ecosystems? a multi-stressors approach linking three biological indices to physico-chemistry, hydromorphology and land use. Ecol. Indic. **48**, 88–98 (2015)