Ecological Informatics 24 (2014) 210–221

Contents lists available at ScienceDirect

Ecological Informatics

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

Discriminant temporal patterns for linking physico-chemistry and biology in hydro-ecosystem assessment

Mickaël Fabrègue [a,b,*], Agnès Braud [c], Sandra Bringay [d], Corinne Grac [e], Florence Le Ber [b], Danielle Levet [f], Maguelonne Teisseire [a]

[a] TETIS, IRSTEA, Montpellier, France
[b] ICube, University of Strasbourg/ENGEES, CNRS, Illkirch, France
[c] ICube, University of Strasbourg, CNRS, Illkirch, France
[d] LIRMM, Montpellier 3 University, CNRS, France
[e] LIVE, University of Strasbourg/ENGEES, CNRS, France
[f] AQUASCOP, Technopole d'Angers, Beaucouze, France

ARTICLE INFO

Article history:
Received 17 June 2014
Received in revised form 30 August 2014
Accepted 2 September 2014
Available online 16 September 2014

Keywords:
Data mining
Temporal patterns
Discriminant patterns
Hydro-ecology
River quality

ABSTRACT

We propose a new data mining process to extract original knowledge from hydro-ecological data, in order to help the identification of pollution sources. This approach is based (1) on a domain knowledge discretization (quality classes) of physico-chemical and biological parameters, and (2) on an extraction of temporal patterns used as discriminant features to link physico-chemistry with biology in river sampling sites. For each bio-index quality value, we obtained a set of significant discriminant features. We used them to identify the physico-chemical characteristics that impact on different biological dimensions according to their presence in extracted knowledge. The experiments meet with the domain knowledge and also highlight significant mismatches between physico-chemical and biological quality classes. Then, we discuss about the interest of using discriminant temporal patterns for the exploration and the analysis of temporal environmental data such as hydro-ecological databases.

1. Introduction

Identifying pollution sources in aquatic ecosystems is currently a major research area and remains a complex task. Many parameters are involved in the determination of aquatic ecosystems quality. These parameters are related to different aspects, such as biology, physico-chemistry and hydromorphology. The importance of having operational tools to help in the interpretation of complex information concerning the water quality of rivers and their functioning, as well as assessment of the effectiveness of ongoing action programs is underlined by international directives such as the European Water Framework Directive (E. Union, 2000). Therefore, it is important to propose new methods that take into account the complexity of the problem.

Measures of these different aspects are performed in river stations by several organizations, with specific research goals. Because the data collected by each actor of the domain have become substantial, it is important to design and implement a large, common and consistent database to aggregate these complementary data. In order to meet this issue, the French ANR[1] Fresqueau project[2] has begun in 2011. This project aims at collecting and unifying databases that are linked to the quality of water bodies. They include biological, physico-chemical and also hydro-morphological data. The result is a consistent spatio-temporal database that brings together information related to north-east and south-east French watersheds. It concerns 11,329 sampling sites spread over 161,100 km^2 that represent 29.45% of metropolitan France. These watersheds are grouped into two major hydrographic areas which are *Rhin-Meuse* (north-east), denoted as *RM* and *Rhône Méditerranée Corse* (south-east), denoted as *RMC*. Fig. 1 illustrates their respective geographic scopes. The dark gray area corresponds to *RM* while the black area corresponds to *RMC*. White line separations in the figure correspond to the different watershed delimitations.

Several dimensions of analysis have been collected. Fig. 2 illustrates the dimensions of analysis of the database: physico-chemistry, hydrobiology, climate, land use, hydrology and hydromorphology. The 11,329 sampling sites are described by these 6 dimensions. The objective is to provide researchers with a maximum amount of data to analyze. It aims at facilitating studies that focus on the relations between various environmental aspects, or the impact of one aspect on another one. Furthermore, some of these different environmental aspects involve a temporal dimension. For example, physico-chemical parameters may be sampled every two months in sampling sites. Considering the different hydro-ecological parameters with their temporal dimension allows the application of original methods. Indeed, some temporal

* Corresponding author.
E-mail address: mickael.fabregue@teledetection.fr (M. Fabrègue).
[1] French National Research Agency.
[2] http://engees-fresqueau.unistra.fr/.

http://dx.doi.org/10.1016/j.ecoinf.2014.09.003
1574-9541/

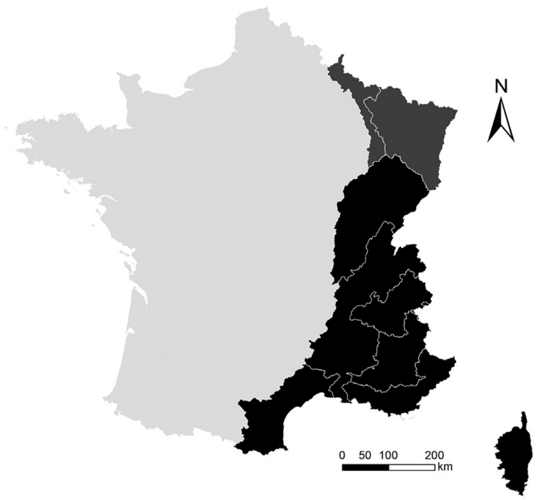

Fig. 1. French watersheds concerned by the Fresqueau database.

data mining approaches are well-adapted to tackle such issues. With specific data structures, they are able to process temporal information that describe environmental aspects.

This paper presents such a data mining method applied on hydrobiological and physico-chemical aspects. It addresses the following issue:

Can we temporally link sets of physico-chemical parameter values with bio-index values?

Identifying these links is important to evaluate more precisely the impact of physico-chemistry on biology. Finding temporally ordered sets of physico-chemical parameter values may help to highlight the synergy produced by their combination. The presented method proposes an original temporal pattern based approach, called discriminant closed partially ordered patterns, to obtain these correlations.

In Section 2, we present existing approaches from the literature. In Section 3, we describe our method divided into three parts:

1. Section 3.1 details the preprocess operations performed on the dataset, and the construction of quality class sub-datasets that correspond to each bio-index.
2. Mining discriminant partially ordered patterns is presented in Section 3.2.
3. The last part consists in selecting and reducing the discriminant partially ordered pattern result set (Section 3.3).

We then provide experimental results performed on the Fresqueau dataset (Section 4) and we finish with a discussion section (Section 5). Fig. 3 synthesizes this process, which is detailed in Section 3.

2. Related work

Several works investigated the task of mining hydrological data.

An important amount of studies focus on macro-invertebrate communities (D'heygere et al., 2003; Dakou et al., 2007; Dedecker et al., 2004; Goethals et al., 2007). For example, Dakou et al. (2007) used decision tree models in order to predict the habitat suitability of some macro-invertebrate taxa in river Axios (Greece). Authors considered physico-chemical and structural characteristics of the river. With the same goal, the efficiency of artificial neural networks in predicting macro-invertebrate taxa in Zwalm (Belgium) river has been shown by Dedecker et al. (2004).

The impact of hydrologic alterations on fish communities in Illinois River has been identified by Yang et al. (2008). Based on 32 indicators of hydrologic alteration, authors highlight the most ecologically relevant indicators by using a genetic programming approach.

Some other authors focused on flora instead of fauna. The first comprehensive checklist of diatoms (948 taxa) with ecological indicator values for pH, salinity, nitrogen uptake metabolism, saprobity, trophic state and moisture was presented by Van Dam et al. (1994). Recently, the physico-chemical impact on diatom communities has been studied by Kocev et al. (2010). They used a multi-target regression trees approach and identified a significant impact of metallic ions and nutrients on diatoms. Recknagel et al. (2013) analyzed phytoplankton phyla populations in Lake Kinneret (Israel), by using a hybrid evolutionary algorithm. Authors showed that considering both physico-chemical and biological variables in models provides the best results in the prediction of population dynamics. Likewise, Bertaux et al. (2009) rely on Formal Concept Analysis to study biological traits of macrophytes taxa in Rhin-Meuse watershed. The goal is to link environmental variables with biological trait granularity in order to identify groups of taxa adapted to a particular environmental context.

State-of-the-art methods show the importance of considering and combining biological and physico-chemical variables in order to find relevant knowledge. Nevertheless, none of these studies has taken into account the temporal aspect based on temporal pattern mining approaches, which is relevant to analyze pollution dynamics. The approach presented in this study is well-adapted to temporal datasets

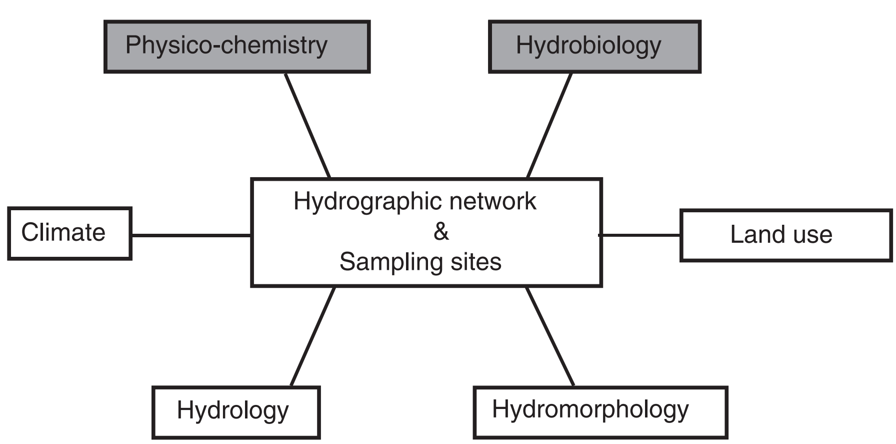

Fig. 2. Categories of data in the Fresqueau database.