HydroQual: Visual Analysis of River Water Quality

Pierre Accorsi*
LIRMM
Univ. Montpellier 2

Mickäel Fabrègue†
IRSTEA Montpellier
Univ. Strasbourg/ENGEES

Arnaud Sallaberry*
LIRMM
Univ. Montpellier 3

Flavie Cernesson†
AgroParisTech Montpellier

Nathalie Lalande†
IRSTEA Montpellier

Agnès Braud‡
ICube
Univ. Strasbourg

Sandra Bringay*
LIRMM
Univ. Montpellier 3

Florence Le Ber‡
ICube
Univ. Strasbourg/ENGEES

Pascal Poncelet*
LIRMM
Univ. Montpellier 2

Maguelonne Teisseire†
IRSTEA Montpellier

ABSTRACT

Economic development based on industrialization, intensive agriculture expansion and population growth places greater pressure on water resources through increased water abstraction and water quality degradation [40]. River pollution is now a visible issue, with emblematic ecological disasters following industrial accidents such as the pollution of the Rhine river in 1986 [31]. River water quality is a pivotal public health and environmental issue that has prompted governments to plan initiatives for preserving or restoring aquatic ecosystems and water resources [56]. Water managers require operational tools to help interpret the complex range of information available on river water quality functioning. Tools based on statistical approaches often fail to resolve some tasks due to the sparse nature of the data. Here we describe HydroQual, a tool to facilitate visual analysis of river water quality. This tool combines spatiotemporal data mining and visualization techniques to perform tasks defined by water experts. We illustrate the approach with a case study that illustrates how the tool helps experts analyze water quality. We also perform a qualitative evaluation with these experts.

Keywords: Visual Analytics, Spatiotemporal Data Mining and Visualization, Water Quality

1 INTRODUCTION

Water is a vital component for all known forms of life. The World Health Organization (WHO) estimates that domestic uses (drinking, cooking, and hygiene) require a minimum of 20 litres of good quality water per day and per person. Water quality issues first arose in the nineteenth century when scientists revealed a link between water quality and health problems; e.g., J. Snow was the first to suggest that cholera is a waterborne epidemic [14, 48]. Scientists noticed that the severity of water quality degradation depends on the population growth and the development of cities and economic activities [40].

Water pollution is a major global problem that the United Nations Environment Program (UNEP) and WHO [56, 53] have highlighted, thus leading many countries to develop tailored water policies and programs for preserving and restoring water quality. Water policies are followed by the expansion and modification of water quality monitoring which in turn depends on the improvement of analytic methods and scientific knowledge. Early river monitoring of a few European rivers began in the late 19th century, but only five water quality descriptors were analyzed [40]. Nowadays, more than 150 descriptors (biological and chemical) are analyzed [40, 53]. Monitoring and associated databases primarily aim to characterize: (1) the overall conditions and trends concerning the river, and (2) the ability to control water for a given use or to assess the impact of human activities on water. This knowledge then leads to decision making on protection or restoration measures to take on the authorization or prohibition of the use of the resource. Large datasets are thus compiled by geolocalized river stations where sampling is based on sequences of biological indices and physicochemical parameters.

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 (WFD) [24]. Several tools based on GIS and including statistics and charts have already been proposed but, due the sparse and heterogeneous aspects of water quality data, statistical approaches overlook some properties. Data mining and information visualization techniques are necessary to overcome the shortcomings of these techniques and complete their results.

Visual analytics is the science of analytical reasoning facilitated by interactive visual interfaces [50]. More precisely, "visual analytics combines automated analysis techniques with interactive visualizations for effective understanding, reasoning and decision making on the basis of very large and complex datasets" [34]. Automated analysis techniques include statistics, mathematics, knowledge representation, management and discovery technologies.

In this paper, we focus on combining data mining techniques with visualization and interaction techniques (see Fig. 1). The tool is the result of 3 years of collaboration between domain experts (hydrologists, hydrobiologists, water managers) and computer scientists specialized in data mining and information visualization.

Our main contribution is the design of a visual interface to explore river water quality. The tool requirements also led us to make some contributions in data mining and information visualization. Here we: (1) propose a new metric to evaluate sequential dissimilarity, (2) present an optimized algorithm to extract temporal patterns, and (3) describe a new algorithm to visualize clusters.

The paper is organized as follows. Sec. 2 focuses on the domain problem characterization. Sec. 3 presents related work. Sec. 4 explains the data abstraction design. Sec. 5 describes the visual mappings and the interactive techniques of our tool. In Sec. 6, we evaluate our approach. We conclude in Sec. 7.

2 DOMAIN PROBLEM CHARACTERIZATION

In this section, two domain experts, who are co-authors of this paper, briefly introduce their domain (section 2.1) and then explain

*e-mail: firstname.lastname@lirmm.fr
†e-mail: firstname.lastname@teledetection.fr
‡e-mail: firstname.lastname@unistra.fr

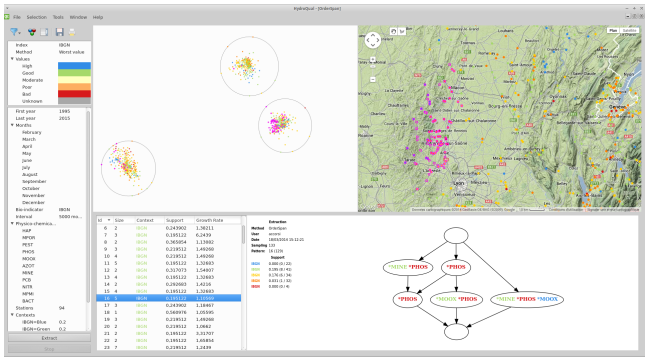

Figure 1: **HydroQual** is a tool that facilitates visual analysis of river water quality. The input dataset consists of sequences of biological indices and physicochemical values for several geolocalized sampling river stations. The **clustering view** (top left) shows stations grouped by their behavioral similarity. The **geographical view** (top right) shows the geolocation of the stations. By selecting a set of stations from these two views, users can extract and visualize temporal patterns regarding biological indices and physicochemical parameters in the **temporal patterns view** (bottom).

the problem (section 2.2). Later, in collaboration with 10 domain experts from universities and consultancy firms, they specify their needs (section 2.3.1) from which computer scientists deduce a list of requirements for the tool (section 2.3.2).

2.1 River Water Quality

Water quality is defined by the capacity of water to sustain various uses or processes and to sustain aquatic life [53]. Water has physical, chemical or biological characteristics that determine its quality. For example, toxic concentration limits are very strict and pathogens are not allowed in drinking water [25], some fishes live in very specific water temperature conditions [39], etc.

Surface waters and rivers are particularly vulnerable. For example, rivers are very sensitive to contamination by pathogens (bacteria, etc.), heavy metals (lead, etc.), and microorganic pollutants (hydrocarbon, etc.), organic matter degradation, and changes in the hydrological regime [53].

The greatest natural influences are geological, hydrological and climatic because they directly affect the quantity and quality of available water. Human activities also influence water resources. Rivers have always constituted an easily accessible water resource for numerous uses leading to consumption (drinking water, industrial water, irrigation water, etc.) or not (hydropower, transportation, fishing, recreational activities, etc.). Water resources are directly impacted by human activities (industrial waste or sewage treatment plants, agricultural nonpoint source pollution, etc.). Pollution can be diffuse or not, continuous or related to events such as the Rhine river pollution in 1986 after the Sandoz chemical spill at Basel (Switzerland) [31].

All the natural processes and human activities are intertwined. Identifying possible relations between biology (fishes, macroinvertebrates, diatoms, macrophytes), physicochemical parameters that represent general water characteristics, macropollutants or micropollutants, physical conditions (banks and riverbed state, hydrology, presence or absence of sill ...) or human pressures (land uses, industry or waste water plant output) needs to take into account many temporal and spatial scales [37].

For instance, when working on nutrients, which are related to water quality degradation, domain experts need to know (1) the local processes that control natural nitrogen, phosphorus and oxygen cycles; (2) the contributions related to human activities at different scales, i.e. atmospheric transport, point source pollution (urban sewage) and diffuse pollution (soil leaching) [51, 41]. They contribute to integrate nutrients issues in district or local water planning or reporting documents (taking into account national or supranational frameworks), to identify critical situations and to prioritize actions.

2.2 Identification of Current Domain Expert Problems

Domain experts are faced with (1) the technical advances in monitoring and databases, (2) the complexity of partially understood processes that govern water quality and river functioning and (3) the increasingly demanding challenges concerning sustainable management of water resources [53].

Monitoring water quality variables follows different standards (adapted metrology and pre-treatment) according to the nature of the variables and the objective of the water quality station (permanent monitoring, specific campaigns to improve knowledge, pollution crisis management). Standards changed in Europe with the implementation of WFD, and detection thresholds are more accurate, especially for micropollutants. Some water quality variables are not directly observed but stored as indices that summarize the information (biological data for instance) and provide qualitative information. Observations or indices are stored in numerous databases. So, databases provide experts with a mass of heterogeneous, sparse and irregular information from which he/she must extract knowledge essential for decision making. As the amount of data is huge, tools including data mining and visual analysis techniques are required for knowledge extraction. These tools must take into account the nature (observation, index, etc.), completeness and quality of the data (sampling river station data are more or less complete, some data are missing, others unreliable), their diverse shapes and origins, as well as the granularity of measurements (data on different spatial and temporal scales).

Firstly, domain experts need to easily access the data (user friendly queries), the treatment (quickness) and the results of a given analysis (effective and smart representation). It is very challenging with water quality data due to their specificities.

In this paper, we work primarily on **physicochemical parameters** and **biological indices**. The definition of the selected descriptors for this paper is provided in section 6.1. They are ranked in **five classes**: high (blue), good (green), moderate (yellow), poor (orange), bad (red) that give the status of the alteration according to the classification of the former French water quality assessment system, as used in [7] or [11]. As an example, we use IBGN, i.e. the French benthic macroinvertebrate biological index [1] based on the abundance and selective sensitivity of river benthic invertebrates to stresses. The obtained value is an integer between 0 and 20 [6]. High values correspond to a high quality status (blue class), which means good conditions for fauna development.

Secondly, domain experts need to respond to operational questions. As a consequence of 2.1, many dynamics are involved in water quality processes, and different dynamics can lead to the same situation. So, it is necessary to find relations between the parameters, e.g. between biological indices and physicochemical parameters. For instance, a high organic matter value, which is one of the signs of excess of feed, can explain low biological index values. In terms of sequences, a poor IBGN class can follow a moderate or poor organic matter class. But is it always the case? Is it the only observed relationship with such result? Do we observe same situations for a specific area (the process is generalized), for comparable contexts (the process is localized)? Do we observe anomalies on data, or outliers (and why)? Do we observe resilient processes (and why)? Can we identify new pressures? Can we quantify restoration actions' efficiency? Do we have more consistent relations, when we work with all the available data instead of a subset localized in a given basin?

2.3 Requirement Analysis

Domain experts currently use statistical methods or models that require a priori an extreme simplification of processes. Data exploration methods end by the interpretation and explanation of the processes. Nevertheless, it is necessary to define practical questions to guide the tool design process.

2.3.1 Expert-Oriented Needs

We identify five major types of questions expressed in general terms that integrate temporal and spatial scales issues and satisfy scientific and operational aspects for the expertise domain: (Q1) For a given goal, which dataset is relevant? (Q2) For a given study area, what river stations have behavioral similarities? (Q3) Conversely, for a given similar subset, where are the river stations being involved are located? (Q4) For a given river station and a given biological index, what is the worst qualitative value computed? (Q5) For a given set of river stations, what are the main temporal trends among the biological indices and physicochemical parameters?

2.3.2 Requirements for the tool

Starting from these questions, we identify seven requirements for our tool: **R1** - Select sub-dataset (years, months, biological indices and physicochemical parameters), for Q1. **R2** - Visualize/navigate through sampling river stations from their locations, for Q3 and Q4. **R3** - Visualize/navigate through sampling river stations from their behavioral similarity, for Q2 and Q4. **R4** - Select groups of sampling river stations from their locations, for Q3 and Q5. **R5** - Select groups of sampling river stations from their behavioral similarities, for Q2 and Q5. **R6** - Visualize biological indices classes, for Q4. **R7** - Extract and visualize temporal patterns of classes of biological indices and physicochemical parameters for Q5.

3 Related Work

Our visual analytics approach combines data mining and visualization techniques. We thus first focus on related work dealing with mining river water status data. We then present related work on visualization.

3.1 Mining Water Quality Data

Many parameters are involved in determination of the water ecosystem status. These parameters are related to different aspects, such as biology, physicochemistry and hydromorphology. Most studies focus on the impact of physicochemistry or hydromorphology on different biological dimensions.

Some research has applied data mining approaches to study the fauna aspect, represented by macroinvertebrates [29], fish communities [60], as well as the fauna dimension with diatoms [35] or phytoplankton populations [45]. State of art methods have revealed the importance of considering and combining biological and physicochemical variables in order to find relevant knowledge. However, none of these studies have taken into account the temporal aspect with pattern mining approaches, which are essential for analyzing pollution dynamics. The approach presented in this study is well adapted for temporal datasets with multiple variables, represented here by biology and physicochemistry. Furthermore, knowledge extracted via temporal pattern approaches are easy to analyze by domain experts. This paper uses such approaches in a method that we present in the section 4.

Currently, the most common pattern-based method used to explore temporal data is called sequential pattern mining. In the literature, this approach has been widely used in many applications such as analysis [28], classification [16] or prediction [54]. They were first introduced by [3] and are a temporal extension of association rules [2]. They were first developed to find correlations between supermarket products. Sequential patterns are relevant when there is an order among database events. This order is usually a temporal one. Let us consider an hypothetical temporal database and a sequential pattern, $\langle$*(Pollution)(Dead_animals)*$\rangle$: 30%, extracted from the database. This sequential pattern means that *Pollution* event is temporally followed by a *Dead_animals* event with a frequency of 30% in the database. Despite their advantages, sequential patterns often bring limited information since they only provide totally ordered information about data. To illustrate this, let us consider a second pattern $\langle$*(Pollution)(Dead_vegetation)*$\rangle$:30% discovered in the same database. It is possible to extract the two patterns exactly from the same set of transactions. They coexist in the database. The coexistence of sequential patterns is not taken into account with this method. However, this coexistence can be synthesized based on partial ordering. Figure 2 presents a so-called partially ordered pattern, denoted *po-pattern* [26], that combines the two previous sequential patterns.

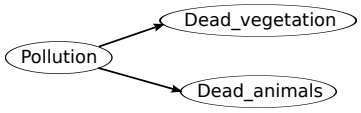

Figure 2: Example of a partially ordered pattern with a frequency of 30%

This pattern means that a *Pollution* event is followed by two events *Dead_animals* and *Dead_vegetation*, which are not ordered. Po-pattern approaches used in hydrobiological databases have some advantages: (1) they are well adapted to the temporal aspect of the database, (2) they provide more information on order among elements than sequential patterns, and (3) they are represented as a directed acyclic graph, which helps in understanding, which is important for domain experts.

3.2 Visualizing Water Quality Data

To our knowledge, the earliest work dealing with water quality visualization is a dot map of John Snow, who plotted the location of deaths from cholera in London for September 1854 [48]. The map clearly shows that deaths were mostly grouped around the Broad Street water pump, which was contaminated.

Several geographical information systems (GIS) have been developed for, or support, water quality analysis [15]. Their visualization is based on a map in which different parameters are plotted, such as areas, rivers, stations, measurements, etc. Some approaches are enriched with statistics and charts (e.g. see [9, 10] or ArcGIS Geostatistical Analyst[1]). Techniques based on more sophisticated plots [27, 33] or self-organizing maps [38] have also been studied. Boyer et al. [12] proposed various visualization approaches, with each visualization technique demonstrating a water quality trend in a specific way.

Statistical based visualization approaches help in extracting knowledge on water quality, but statistical analysis fails to capture the complexity of some processes. In this paper, we thus propose a visualization based on data mining techniques. These techniques help reveal different kinds of patterns to supplement the statistical analysis. To our knowledge, no visualization tool including data mining for water quality analysis has been proposed so far.

As this paper focuses on geospatial and temporal data only when applied to water quality, we do not give an overview of the related visualization techniques. For an introduction to geospatial data visualization, please refer to the 5th and 6th chapters of [55]. An introduction to temporal data visualization and a major overview of the techniques is given in [4]. Spatiotemporal data visualization has been studied in [5]. A tool combining data mining and visualization

[1] `http://www.esri.com/software/arcgis/extensions/geostatistical` [Online; accessed 24-July-2014]

techniques for spatiotemporal data has also been proposed, but the techniques do not meet our requirements [17].

Finally, sequential pattern visualization has been investigated in several studies (e.g. [57, 59, 47]), but to the best of our knowledge no studies have been specifically devoted to po-pattern visualization.

4 Data Abstraction

The data we explore are temporal sequences composed of biological indices and physicochemical parameters from river sampling stations. Thus, sequences are built by ordering measure samplings according to their timestamp. In relational database vocabulary, the physicochemical and biological variables are called items. An itemset IS is a non-ordered group of measures sampled at the same timestamp. A sequence $S = \langle IS_1 IS_2 \dots IS_{|S|} \rangle$ is a non-empty and ordered list of $|S|$ itemsets, i.e. groups of measures ordered according to their timestamp. Discussions with domain experts lead to the discretization of all variables from the dataset into five classes: Blue, Green, Yellow, Orange and Red, as previously explained in Sec. 2.2. To illustrate this, let $\langle$(AZOT,PHOS)(IBGN)(AZOT)$\rangle$ be the sequence of a river station. This means that, in this river, a green AZOT has been measured at the same time as a blue PHOS, followed by a blue IBGN, in turn followed by a red AZOT. Sec. 4.1 and 4.2 introduce two approaches based on temporal sequences.

4.1 Clusters of Sequences

To address the third requirement (**R3**), we need to group river stations according to their behavioral similarity, i.e. similarities in their corresponding sequences. All clustering techniques are based on dissimilarities between data features. Measuring the similarity between two itemset sequences is not a trivial task. Thus, we adapt one of the most famous dissimilarity measures used in the timeseries domain: Dynamic Time Warping [46]. We call our approach adapted to itemset sequences Dynamic Sequence Warping. Let $S = \langle IS_1 IS_2 \dots IS_{|S|} \rangle$ and $S' = \langle IS'_1 IS'_2 \dots IS_{|S'|} \rangle$ be two itemset sequences. Dynamic Sequence Warping on S and S', denoted $DSW(S,S')$ is defined as:

$$DSW(S,S') = \gamma(|S|,|S'|) \tag{1}$$

Where $\gamma(|S|,|S'|)$ is the optimal cumulative path recursively defined as:

$$\gamma(i,j) = \delta(IS_i, IS'_j) + \min \begin{cases} \gamma(\ i-1\ ,\ j-1\ ), \\ \gamma(\quad i\quad ,\ j-1\ ), \\ \gamma(\ i-1\ ,\quad j\quad) \end{cases} \tag{2}$$

It is inspired by the Dynamic Time Warping definition. The main difference concerns the dissimilarity measure used between sequence elements. In timeseries, elements are quantitative values but in itemset sequences elements are discrete variables. Then we use a Jaccard distance, denoted $\delta(IS_i, IS'_j)$ in the equation 2, which is very popular for determining how two sets of elements are different. In this paper, we use this dissimilarity measure between sequences in a hierarchical agglomerative clustering algorithm.

4.2 Closed Partially Ordered Patterns

The aim of this step is to extract po-patterns from a sequence database, where each sequence corresponds to the samples of a station (**R7**). In the following, the database in Table 1 is used to illustrate our method. We symbolize items with alphabet letters to facilitate the following definitions. In practice, items are physicochemical and biological variables. This database example contains three sequences composed of items a, b, d, e, f and g.

A po-pattern (partially ordered pattern) is a directed acyclic graph $G = (V,A)$, where V is the set of itemsets and A is the temporal order between these itemsets. The temporal order is transitive

Seq id	Sequence
S_1	$\langle (af)(d)(e)(a) \rangle$
S_2	$\langle (e)(abf)(g)(bde) \rangle$
S_3	$\langle (e)(a)(b)(g) \rangle$

Table 1: An example of a sequence database

for all $u,v \in V$, $u < v$ if there is a directed path from u to v. However, these elements are not comparable if there is no path from u to v or from v to u. Each path in the graph is a sequence. In our case, the order relation between vertices represents the temporality. Thus, in a po-pattern $G = (V,A)$, two vertices $u,v \in V$ such as $u < v$ means that the itemset u is temporaly followed by the itemset v. For example, in the po-pattern G_4 in Fig. 3.d, there is a path from (e) to (g) then $(e) < (g)$, i.e. (e) is followed by (g). Po-patterns are characterized by their support. The support is the number of sequences in the database that contain the po-pattern. A sequence S supports a po-pattern G, denoted $G \preceq_s S$, if all paths (sequences) in G are included in S. For example the sequence $\langle (a)(f) \rangle$ is included in the sequence $\langle (a)(c)(cf) \rangle$. Then the po-pattern G_1 in Fig. 3.a is supported by sequences S_1 and S_2 in the database, i.e. $G_1 \preceq_s S_1$, $G_1 \preceq_s S_2$ and support of G_1 is 2.

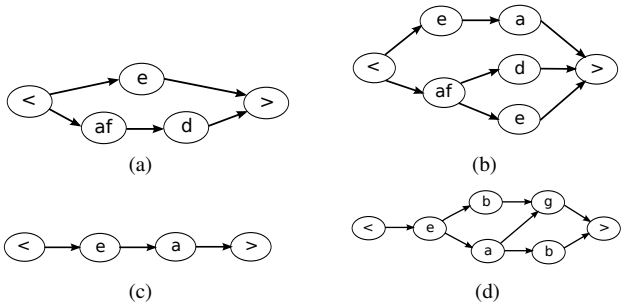

Figure 3: List of patterns extracted from Table 1: (a) G_1 on $\{S_1,S_2\}$. (b) G_2 on $\{S_1,S_2\}$. (c) G_3 on $\{S_1,S_2,S_3\}$. (d) G_4 on $\{S_2,S_3\}$.

In po-pattern mining, some po-patterns are redundant in the results. For example, po-patterns G_1 and G_2 in Fig. 3.a and 3.b are both supported by sequences S_1 and S_2 but we observe that for each path in po-pattern G_1, there is at least one path in po-pattern G_2 that supports it, but the reverse is not true. G_1 is considered as included in G_2 (G_2 is more specific than G_1) and is denoted $G_1 \preceq_g G_2$. Then G_1 is redundant w.r.t. G_2 since they are supported by the same set of sequences and G_2 contains all the information in G_1 as well as other information not in G_1. Furthermore, we observe that there is no other po-pattern G' such that $G_2 \preceq_g G'$ supported by S_1 and S_2. G_2 is then considered as a closed po-pattern. The complete set of closed po-patterns is a subset of the set of po-patterns without information loss, i.e. it is possible to retrieve the complete set of po-patterns with the complete set of closed po-patterns. In this paper, we focus on such patterns. Fig. 3.b, 3.c and 3.d give the complete set of closed po-patterns from the database in Table 1 with a support greater or equal to 2.

Extracting closed po-patterns is a complex task. We use the algorithm proposed in [26] that extracts the complete set of closed po-patterns with a support greater than a given a minimum support θ. The minimum support parameter is mandatory since the closed po-pattern search space is huge.

Closed po-pattern extraction follows a pattern-growth approach represented by a prefix-tree. Such a tree is a representation of the overall search space of the algorithm. Closed po-patterns are constructed by expanding frequent sequence prefixes [43] from the database (sequence prefixes with a support greater or equal to θ).

Fig. 4 shows the prefix-tree that covers the database in Table 1.

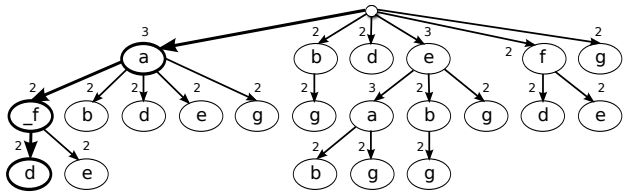

Figure 4: Overall search space

Labels in vertices represent frequent items and a path from the root to a leaf represent a frequent sequence prefix. For example the path on the left in the prefix-tree represents the sequence prefix $\langle(af)(d)\rangle$. Numbers under vertices are the support of the frequent sequence prefix represented by the node, i.e. the support of $\langle(af)(d)\rangle$ is equal to 2. We observe two kinds of items, normal items and items preceded by symbol '_'. They represent S-Extensions and I-Extensions, respectively. An I-Extension extends an itemset: sequence prefix $\langle(a)\rangle$ I-Extended with item d gives the sequence prefix $\langle(ad)\rangle$. An S-Extension extends the sequence: sequence prefix $\langle(a)\rangle$ S-Extended with item d gives the sequence prefix $\langle(a)(d)\rangle$. The set of closed po-patterns given by Fig. 3.b, 3.c and 3.d are extracted from this prefix-tree. Indeed, each pattern is deduced from a sub-prefix-tree of the overall prefix-tree. Fig. 5.a provides the sub-prefix-tree used to process closed po-pattern G_3. Thus, the algorithm is divided into two parts: (1) extracting the complete set of closed sub-prefix-trees and (2) each closed sub-prefix-tree is processed by starting from leaves to remove redundancy and merge equivalent sequence suffixes. For more details about the algorithm, please refer to [26].

Unfortunately, applying this approach generates many redundancies in some branches of sub-prefix-trees. For example, we highlight two redundant parts in Fig. 5.a. This means that branches $\langle(a)(g)\rangle$ and $\langle(a)(b)\rangle$ are uselessly explored since they are re-explored in branches $\langle(e)(a)(g)\rangle$ and $\langle(e)(a)(b)\rangle$. These redundancies make the algorithm time consuming and are unsuitable for a visualization tool that requires realtime interaction. We thus improve the algorithm to avoid these redundancies and improve the process. We propose one inspired by [58], in which the authors tackle this issue by detecting and merging branches that are equivalent (refer to [58] for details about the equivalence checking property). We adapt to closed po-pattern this technique that was initially proposed in closed sequential pattern mining since both approaches follow the same pattern-growth paradigm. We illustrate this process in Fig. 5.b where a redundant part of the tree is merged. This markedly improves the extraction process because redundant paths in sub-prefix-trees are directly merged instead of explored. Furthermore, it improves the second step based on frequent sequence suffixes since some branches have already been merged by the optimization.

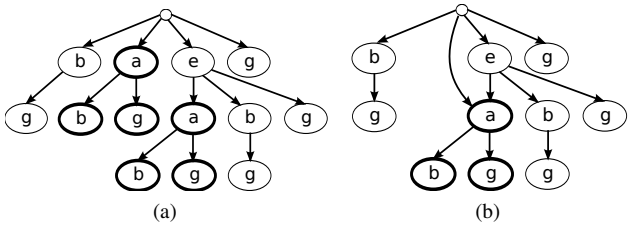

Figure 5: Example of the optimization: (a) Non-optimized sub-prefix-tree. (b) Optimized sub-prefix-tree.

In order to highlight the optimization gain, we process experiments on different groups of river station sequences. Fig 6 shows results obtained with the non-optimized and optimized algorithms. The protocol is: we randomly select 50 sequences from the initial dataset and extract closed po-patterns at various minimum support thresholds with both algorithms. We perform this operation 30 times and compute the average computation for each minimum support threshold. We apply this on various subsets of sequences to simulate the sequence selection and the pattern extraction proposed by HydroQual. We observe that the optimized version is more efficient than the non-optimized one, especially at low minimum support threshold. For example, at a minimum support of 0.32, the extraction time is 6.7 sec for the optimized version while it is 35.53 sec for the non-optimized one. The optimized version is 3.96 times faster at a minimum support of 0.35, and 5.3 times faster at a minimum support of 0.32. This optimization substantially improves the HydroQual interactivity since there is a greater computation gain at lower minimum support.

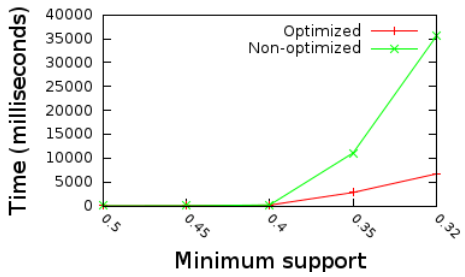

Figure 6: Non-optimized versus optimized algorithm

5 Visual Mappings and Interactive Functionality

Based on the data abstraction described in Sec. 4, we now focus on the different visualizations (Sec. 5.1 - 5.3) and interaction methods (Sec. 5.4) of HydroQual. This tool was developed for and in cooperation with domain experts. It is based on three coordinated views designed to fit the seven requirements described in Sec. 2.3.2.

5.1 Geographical View

In this section we describe the view of HydroQual that shows the geographical locations of the river stations (**R2**).

The user can switch between several background maps like the ones provided by Open Street Map or Google Map. In particular, the use of Google Maps physical background highlights topological characteristics of the area and was found very useful by the domain experts (see Fig. 1). Domain experts also lead us to include the Corinne Land Cover background [23], which shows the land use (forest, agricultural lands, urban and industrial areas, etc.). See Fig. 7.a for an example.

Additional layers are also available: watershed, hydroecological regions, administrative regions and hydrographical network. The user can select the layers he wants to see. They are then displayed on the foreground of the map (see Fig. 7.b for an example).

The final view also includes Leaflet's[2] zooming (mouse wheel) and moving function (mouse drag 'n drop) to ease exploration of the visualization. Fig. 7 shows the results obtained on a real example.

5.2 Clustering View

In this section, we describe the HydroQual view showing clusters of stations extracted as described in Sec. 4.1. This view helps domain experts identify which stations act in a similar way (**R3**).

[2] `http://leafletjs.com/` [Online; accessed 24-July-2014]

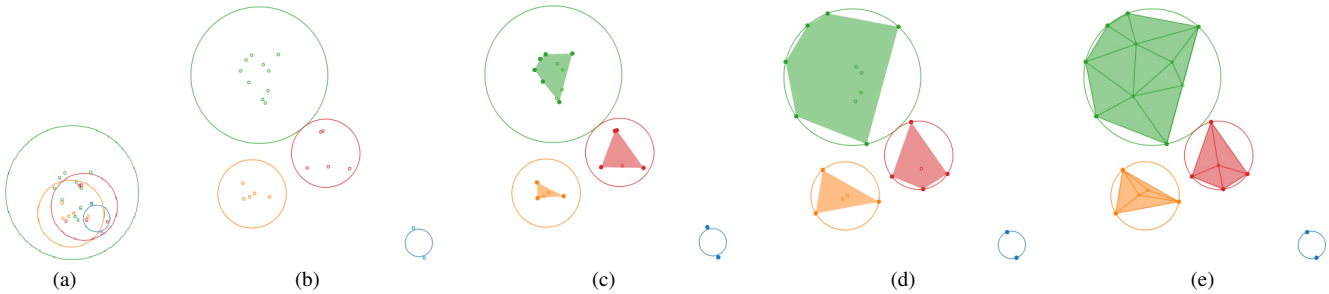

Figure 8: Example of the embedding process: (a) Initial placement with a MDS algorithm. (b) Cluster overlap removal. (c) Computation of the convex hull. (d) Moving of the stations of the convex hull to cluster circles. (e) Triangulation and rearrangement of the stations with Tutte's algorithm.

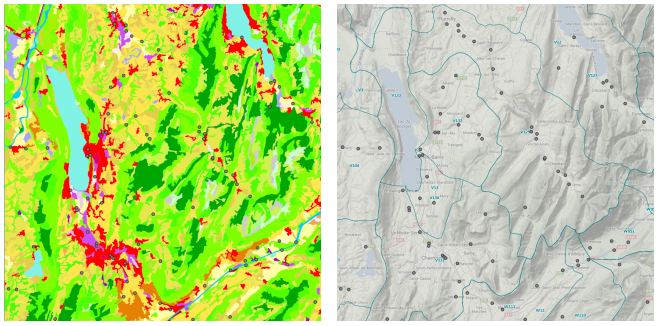

Figure 7: View of the sampling stations of the area of Aix-les-Bains, France. (a) Corinne Land Cover background. (b) Google Maps physical background with watershed contours.

We propose a four step technique for laying out stations and clusters: (1) we use the distance matrix to position stations with a multidimensional scaling algorithm, (2) we represent clusters as circles positioned at the barycenter of the stations they contain, (3) we use methods drawn from collision detection studies to seperate overlapping cluster circles and we also move the stations to keep their barycenters at the centers of the circles, and (4) we rearrange the stations for each cluster to better fit the areas demarcated by the circles.

5.2.1 Initial Placement of Stations (step 1)

Multidimensional scaling techniques [13] are aimed at modeling dissimilarity data among pairs of objects into points in a low-dimensional geometric space. These points can be represented graphically and visualized afterwards. For our purpose, points represent river stations and the dissimilarity measure is computed as described in Sec. 4.1. A non-metric MDS (Smacof, [18]) is performed. When provided a dissimilarity matrix and a desired destination dimension k (2 in our case), it fits stations to a k-dimension configuration by iteratively increasing a stress function. An initial configuration can be either randomized or given. A metric MDS [36] is often computed as an initial configuration. However, this technique is very time-consuming and performs poorly on large datasets. We therefore use the freefold embedding heuristic [44] to obtain the initial configuration.

5.2.2 Cluster Display (step 2)

We represent clusters as circles positioned at the barycenter of the stations they contain. Discussions with domain experts reveal that the main information to highlight is the cluster size, which is not easy to detect for very dense clusters (most of the stations overlap). We thus map the cluster size to the radius of the corresponding circle (see Fig. 8.a).

5.2.3 Clusters Overlap Removal (step 3)

The obtained configuration contains overlapping clusters which makes it hard for domain experts to distinguish clusters and view river stations as part of a cluster (see Fig. 8.a). Discussions with domain experts reveal that they are interested in visualizing distances between clusters and distances between pairwise nodes of the same clusters. Distances between nodes of different clusters are not as important for them. We thus move clusters to avoid overlap while preserving the relative distances of their centers. The method is based on collision detection between circles. Most collision detection approaches act in two phases to reduce computation time [22]. First, the broad phase is aimed at excluding a maximum of couples to be tested with space partitioning algorithms. We use a quadtree in our application. Then the narrow phase precisely determines if the remaining couples collide with each other and reveals their penetration vector. Since we represent a cluster with a circle, collision detection between two clusters as well as computation of the penetration vector is straightforward. We finally use the highest penetration vector among all couples of overlapping clusters to rearrange the circles while preserving their relative distances (see Fig. 8.b).

5.2.4 Final Station Placement (step 4)

We finally rearrange the river stations for each cluster to better fit the area defined by the circle. After this step, distances between stations of a cluster do not correspond to the same dissimilarity measures of the same distances in another clusters. Discussions with the domain experts reveal that it is more important to detect relative positions of couples of inner river stations than to preserve absolute river station distances within the whole configuration. Moreover, space distortion of inner river stations removes some overlaps and helps in visualizing stations.

We first compute a convex hull of the river stations for each cluster [30] (see Fig. 8.c). Then we move river stations on the hull to the circle (see Fig. 8.d). We compute a Delaunay triangulation among the river station positions. This gives us a graph where nodes on the outer face represent stations of the convex hull. We apply Tutte's algorithm on this graph [52]. It iteratively positions every river station that is not on the outer face at the barycenter of its neighbors. Our graph is triangulated and we fix its outer face. So the solution is unique and preserves the relative positions of river stations in the triangulation. Using Tutte's algorithm to reorganize the node positions has already been tested in another context [8]. Fig. 8.e shows the results obtained on an example.

The final view also includes a zooming (mouse wheel) and a

moving function (mouse drag 'n drop) to ease exploration of the visualization. Fig. 9 shows the results obtained on a real example.

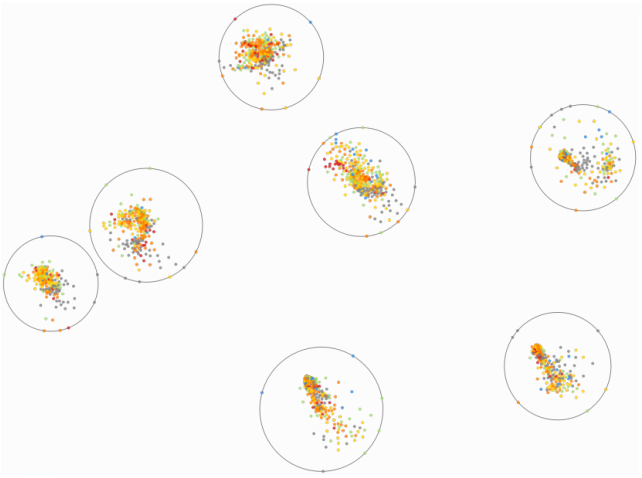

Figure 9: View of sampling stations grouped by their behavioral similarity.

5.3 Temporal Patterns View

In this section we describe the HydroQual view that shows closed po-patterns extracted as described in Sec. 4.2 (**R7**). According to domain expert requirements, this view requires a set of stations as input and a biological index. The biological index is selected by the user when he/she launches the extraction. Closed po-patterns extracted only deal with this biological index and the related physicochemical parameters. In the next section, we describe how stations are selected from the two previous views.

The temporal pattern view is divided into two panels (see bottom of Fig. 1). The first panel located on the left side shows the list of closed po-patterns extracted. Each element on the list shows the id of its associated closed po-pattern, its size (number of nodes), a class of the biological index selected and an index showing how the closed po-pattern is discriminant as compared to the other classes. The discriminant score is determined by using the growth rate interestingness measure [20] (see Growth Rate in Fig. 1). Briefly, it involves retrieving closed po-patterns that are specific to a considered biological index class. A click on an element displays the corresponding closed po-pattern on the right panel.

Since a closed po-pattern is a directed acyclic graph, it can be displayed using graph drawing techniques. We draw the graph in a layered fashion using Sugiyama's algorithm [49] to place emphasis on the sequential aspect. With this approach, each layer represents a different timestamp. We use the implementation of the algorithm provided by the Graphviz library [21].

As mentioned in Sec. 4, an itemset consists in a set of items. An item is a physicochemical parameter. It possesses a variable name and its associated class which belongs to a set of colors, as defined by domain experts. Depending on the domain experts requirements, we represent an item with its variable name in plain text, highlighted by the color corresponding to its value. Fig. 1 shows a closed po-pattern on a real example (see bottom).

The final view also includes a zooming (mouse wheel) and a moving function (mouse drag 'n drop) to facilitate exploration of the visualization.

5.4 Interaction

This section covers the issue of how the user interacts with the views. The user is first asked to select a dataset. This dataset can be furthermore filtered. The filter parameters include year interval, consecutive months, biological indices and physical parameters of interest (**R1**). Parameters selected are displayed on the left panel (see Fig. 1).

When launched with a specific button, the temporal pattern view computes and shows the closed po-patterns corresponding to a set of stations. This set has to be defined by the user from the clustering and/or the geographical views (**R4** and **R5**). There are two kinds of interaction technique for this purpose. One can select a station individually by clicking on it or he/she can select multiple stations at once by using a Lasso type tool. The selected stations are then highlighted by changing their border color to pink (instead of black). When the selection is changed in a given view, the other view is then updated so that the selection is the same in both views (see Fig. 1).

The color of a station corresponds to the worst value of a selected biological index found in the station sequence. The user can change the selected indicator using a dedicated button (**R6**).

As requested by domain experts, we implement copy, print and save features to allow further consultations of the measures and other station related data. We also keep logs of temporal patterns generated during a session in the left panel.

6 Demonstration with Feedback

Two domain experts who are co-authors of this paper first evaluate the visual analysis tool through a case study showing how the tool has been useful for them to find already known information and, more importantly, to discover new ones. We then present their feedback collected through a questionaire they had to fill.

6.1 Case Study

The objective of the case study is to illustrate one application of the use of HydroQual. This example highlights the relationship between values of a biological index and physicochemical parameters describing river water quality. The expectations cover different aspects: the ability to analyse all the available data, considering their spatial and temporal characteristics, to confirm some existing knowledge, to find new ones or anomalies, and if possible, to deduce some operational conclusions.

As a biological index, we chose to work with the French benthic macroinvertebrate index IBGN [1], based on the abundance and selective sensitivity of river benthic invertebrates to stresses, which are commonly monitored to assess organic pollution and hydromorphological degradation [42]. However, invertebrates are sensitive to multiple stressors, which is an obstacle to the identification of relationships between biological indices and physicochemical parameters [32].

We decided to select five groups of physicochemical parameters: four groups to account for interference by macropollutants (MOOX, AZOT, NITR and PHOS) and the last group to account for disturbances in the general water characteristics (in our case, we focus on mineral characteristics, MINE). First of all, we define these five groups of physicochemical parameters.

The MOOX group consists of eight physicochemical parameters and indicates the presence or absence of organic pollution which consumes oxygen in the river. Various origins can explain this pollution: mainly domestic and industrial (agrofood) wastewater, livestock wastewater, winegrowing effluents, natural plant debris, etc. The main impacts are (1) water deoxygenation, (2) the silting of river bottoms, and (3) macroinvertebrate habitats modifications. To reduce this disorder, water managers improve treatment of wastewater organic matter. **The AZOT group** concerns nitrogenous parameters except nitrates. It indicates nitrogen contamination in rivers derived mainly from domestic, industrial and livestock effluents. An excess of nitrogen contributes to create anoxic conditions and can thus have toxic effects on the ecosystem, including

on macroinvertebrates and fish fauna. The treatment of this kind of pollution is the same as for MOOX. **The NITR group** consists only of the nitrate concentration. Nitrates can be directly assimilated by plants. Combined with other nutrients, they favor algae and aquatic plant growth. Nitrates mainly indicate diffuse pollution from agricultural sources (fertilizer leaching during rainfall events). Atmospheric nitrates seeping into rivers during the rainy season resulting from agricultural spraying, industrial and road traffic emissions. An excess of nitrates causes eutrophication. Specific treatments have to be implemented to reduce nitrates in wastewater. To limit agricultural nitrates, different actions can be proposed: (1) limit the use of fertilizers, (2) promote an interface area and catch crops to limit nitrate fluxes into rivers. **The PHOS group** consists of total phosphorus and orthophosphate. Phosphorus is an essential element for proper development of organisms. Phosphorus fluxes are derived from domestic and industrial wastewater (mainly detergents), livestock building effluents, and soil erosion. This is the main factor limiting or not eutrophication. Phosphorus fluxes are mainly reduced by limiting the use of detergents that contain phosphorus, and limiting fertilizer applications, and specific treatments of livestock effluents. **The MINE group** concerns the conductivity and concentration of calcium, chlorides, sulfates, cations, sodium, and magnesium. MINE provides information on the water origin. River waters enriched in minerals and salts depend on the residence time in geological formations and the nature of soils leached during rainfall events. Minerals and salts are mainly of natural origin but can also be related to human activities (agricultural, industrial or domestic). Macroinvertebrates are very sensitive to water salinity (and hence conductivity) and the presence of minerals for their growth and development [19]. But there are thresholds (deficiency or excess) above which salts and minerals are harmful to fauna. Remediation actions are limited to setting up some specific wastewater treatments.

Then, we prepare the queries for the test set. The IBGN values and those of the five physicochemical parameter groups are ranked in five quality classes as described above (high/blue, good/green, moderate/yellow, poor/orange, bad/red). More precisely, we study the relationship between quality classes of IBGN and of physicochemical parameters. The tested dataset extends from 1 July 1999 to 31 December 2010. River stations are located in the east of France. We decided to define subsets of water quality stations used in the clustering view for a first diagnosis of the advantages of this ranking, seldom used by domain experts.

The cluster view shows eight clusters from all water quality stations available in the database. A comparison between clusters and station locations on the geographical view shows that each cluster regroups stations with behavioral similarity and with a geographical organization: each cluster consists of one or a few groups of neighboring stations plus points scattered throughout the entire study area. The clusters could be explained by the overall criteria of position (latitude), environment (climate, altitude), and the nature of human activities (rural, industrial plains, etc.). So we find a coherence at the district level.

We select a single cluster whose stations are exclusively located in the Rhone-Mediterranean and Corsica districts. These stations represent rivers located in rural foothills. The tested datasets consist of 957 sampling records across 287 water quality stations.

We launch closed po-pattern extraction on these stations with the IBGN biological index. HydroQual proposes 176 closed po-patterns for a support of 0.25. Thanks to our optimization, the pattern extraction is quasi-immediate.

We analyze the closed po-patterns for each IBGN class (Table 2). Two criteria, i.e. the support and growth rate, help domain experts to interpret the results. In our case, the interpretation is solid, especially since the support and growth rate values are the highest.

We identify for each class of IBGN, the most frequent po-patterns. Let us begin by the IBGN blue class. The IBGN blue class is mainly following the MINE blue class, and the MOOX blue (in one case green) class, in different combinations and different MINE-MOOX group temporal associations. The blue class for the MINE group and MOOX group is discriminating (high growth rate) compared to the other four IBGN classes. The IBGN green class is following the MINE green class. The IBGN yellow class is mainly following the yellow or orange classes for the PHOS group, which often appears in combination with the MOOX green class. Another kind of association relates to the yellow or green classes for the MINE group. The IBGN orange class is following primarily the PHOS red class that often appears in combination with the AZOT yellow class. Other associations relate yellow or orange classes for the MINE group, and green or yellow classes for the MOOX group. The IBGN red class is only represented by five samples and is following the MINE red class.

Class	Sampling	Closed po-patterns
Blue	274	13
Green	446	2
Yellow	170	110
Orange	62	52
Red	5	12

Table 2: Number of river stations and of closed po-patterns extracted for each class.

From this description, we highlight known information. Even at this stage, identification of known information with a new approach leads to a very interesting conclusion: excellent correspondence between a given IBGN class and its corresponding MINE class, i.e. the IBGN blue class follows a MINE blue class, a IBGN green class, a MINE green class, etc. Further, MINE classes are discriminating. Here we clearly find that the IBGN reflects the physical environmental conditions. In addition, we assume that the associations are different between blue and green classes on the one hand, yellow, orange and red on the other. This distinction is interesting because it allows us to clearly distinguish samples from rivers with good or very good water quality in terms of IBGN classes from the others. Thus, when the IBGN is in the blue class, the MINE blue class is associated (before, after or in the same time) with the MOOX blue class. This means that there is little or no organic pollution present. This is a typical situation for semi-natural areas with few human activities.

The analysis also provides new insight for rivers whose classes (yellow, orange and red) indicate poor water quality in terms of IBGN classes. The PHOS group always has a yellow class or worse and appears to be more limiting than the class for the MOOX group, which can be green. Here we trace pollution of human or animal origin. The PHOS-AZOT group association for the orange class indicates point source pollution due to the presence of livestock or wastewater treatment plants. Indeed, the predominance of the PHOS-AZOT group combination compared to the NITR group reveals a hierarchy of pollution sources: point pollution seems to influence water quality of the studied rivers more than diffuse pollution due to agricultural activities, even though it could be present in these areas. For these rivers, water managers have to improve wastewater treatment and limit detergent use.

It is encouraging for hydrologists to discover known relationships between biological index and physicochemical parameters in a qualitative way. It is also interesting to be able to propose some operational recommendations useful for water managers at district level who have to define priorities to restore water quality.

In conclusion, this first test demonstrates the consistency between the values of classes of the studied biological index and of the physicochemical parameters. Different closed po-patterns high-

light many forms of association, thus expressing the complexity of biophysical processes related to oxygen, nitrogen and phosphorus cycles. This type of analysis is not as directly possible by conventional statistical approaches.

6.2 User Feedback

We evaluated the benefits of our visual analysis tool with two academic domain experts. We first explained the data abstractions, the visual encoding schemes and the interaction features of the tool. Then we asked them to manipulate the prototype. Finally, we asked them to fill in a questionaire dealing with the aesthetics, visual design, interactions, learnability, performance, functionality and ability to help in extracting information.

They both found the tool aesthetically pleasing. On the visual design, they found it was tailored for their tasks. They particularly appreciated the use of colors in the different views. One also pointed out the quality of the mapping of stations and clusters in the clustering view. Concerning the interaction features, they both mentioned that they corresponded to their needs. They really appreciated the modes of selection in the clustering and geographical views. Both agreed that manipulating the tool is easy and intuitive. They estimated the learning time from 15 to 30 min. They were more reserved about the data abstraction. They explained that use of the tool requires an understanding of the structures extracted with data mining techniques (clusters and closed po-patterns), which is not trivial for non-specialists. Nevertheless, when they got familiar with them, they both agreed that these structures are necessary for their tasks, and are major improvements as compared to previous tools. Both domain experts highlighted the interactivity of the tool, appreciating especially that the running time of the closed po-pattern extraction never exceed a few seconds for current use. The domain experts used all the features implemented and thought they were all useful and tailored to their needs. This is not surprising because most of these features meet their initial requirements. Their main suggestion was to group highly similar closed po-patterns. Defining similarity between closed po-patterns and designing an adapted interactive visualization is beyond the scope of this paper, but we plan to do this in the near future. Finally, the tool helped the domain expert to find previously known information and new information (see examples in Sec. 6.1).

In conclusion, the domain experts were very enthusiastic regarding the prototype we gave them. They both mentioned that the new approaches developed are very promising to supplement their usual statistical analysis tools. They particularly appreciated the fact that they gained new insight to help them in addressing important questions. They also think that the tool can be a good support in discussions. They plan to use it to perform their daily analyses.

7 Conclusion and Future Work

We have described HydroQual, a visual analysis tool of river water quality. It combines spatiotemporal data mining and visualization techniques to perform tasks defined by domain experts. Besides the design of the overall process, the tool requirements led us to make some contributions concerning data mining and information visualization techniques: (1) we propose a new metric to evaluate sequential dissimilarity in Sec. 4.1, (2) we present an optimized algorithm to extract closed po-patterns in Sec. 4.2, (3) we describe a new algorithm to visualize clusters in Sec. 5.2.

We have illustrated the efficiency of the tool with a case study. The first results are very promising. Within a few hours, domain experts were able to find previously known information and new information. We also collected domain experts feedback that highlighted their enthusiasm.

As a future work, we plan to improve the temporal pattern tools. As mentioned by the domain experts, some closed po-patterns are highly similar. Finding both a method to group them and a visual encoding to show the overall representative patterns and explore groups in detail would help experts to extract information faster.

Acknowledgements

This project was funded through the French National Research Agency (ANR) project ANR 11 MONU 14 FRESQUEAU. We would like to thank Corinne Grac (LIVE), Xavier Dolques (ICUBE) and Danielle Levet (AQUASCOPE) for their assistance in collecting data and defining the requirements. We also would like to thank Hugo Alatrista Salas and Vijay Ingalalli (LIRMM) for their technical expertise.

References

[1] AFNOR. Qualité écologique des milieux aquatiques - Détermination de l'indice biologique global normalisé (ibgn). Technical Report NF T90-350, Association francaise de normalisation, 2004.

[2] R. Agrawal and R. Srikant. Fast algorithms for mining association rules in large databases. In *International Conference on Very Large Data Bases (VLDB'94)*, pages 487–499, 1994.

[3] R. Agrawal and R. Srikant. Mining sequential patterns. In *International Conference on Data Engineering (ICDE'95)*, pages 3–14, 1995.

[4] W. Aigner, S. Miksch, H. Schumann, and C. Tominski. *Visualization of Time-Oriented Data*. Springer, 2011.

[5] N. Andrienko, G. Andrienko, and P. Gatalsky. Exploratory spatio-temporal visualization: an analytical review. *Journal of Visual Languages and Computing*, 14(6):503–541, 2003.

[6] V. Archaimbault, , P. Usseglio-Polatera, J. Garric, J.-G. Wasson, and M. Babut. Assessing pollution of toxic sediment in streams using bioecological traits of benthic macroinvertebrates. *Freshwater Biology*, 55(7):1430–1446, 2010.

[7] V. Archaimbault, P. Usseglio-Polatera, and J.-P. Bossche. Functional differences among benthic macroinvertebrate communities in reference streams of same order in a given biogeographic area. *Hydrobiologia*, 551(1):171–182, 2005.

[8] D. Auber, C. Huet, A. Lambert, B. Renoust, A. Sallaberry, and A. Saulnier. GosperMap: Using a gosper curve for laying out hierarchical data. *IEEE Transactions on Visualization and Computer Graphics*, 11(19):1820–1832, 2013.

[9] R. Axler, N. Will, E. Ruzycki, J. Henneck, J. Olker, and J. Swintek. Minnesota lake water quality on-line database and visualization tools for exploratory trend analyses. Technical Report NRRI/TR-2009/28, University of Minnesota Duluth, 2009.

[10] S. T. Benedict, P. A. Conrads, T. D. Feaster, C. A. Journey, H. E. Golden, C. D. Knightes, G. M. Davis, and P. M. Bradley. Data visualization, time-series analysis, and mass-balance modeling of hydrologic and water-quality data for the McTier Creek Watershed, South Carolina, 20072009. Technical Report Open-File Report 20111209, U.S. Geological Survey, 2012.

[11] G. Billen, J. Garnier, J.-M. Mouchel, and M. Silvestre. The seine system: Introduction to a multidisciplinary approach of the functioning of a regional river system. *Science of The Total Environment*, 375(13):1–12, 2007.

[12] J. Boyer, P. Sterling, and R. Jones. Maximizing information from a water quality monitoring network through visualization techniques. *Estuarine, Coastal and Shelf Science*, 50(1):39–48, 2000.

[13] I. Brog and P. J. F. Groenen. *Modern multidimensional scaling : Theory and applications*. Springer-Verlag, 1997.

[14] W. H. Brooke and J. V. Liebig. *The Chemical Gatekeeper*. Cambridge University Press, 2002.

[15] D. A. Bruns and T. O. Sweet. Geospatial tools to support watershed environmental monitoring and reclamation: Assessing mining impacts on the Upper Susquehanna-Lackawanna American heritage river. In *Advanced Integration of Geospatial Technologies in Mining and Reclamation*, 2004.

[16] H. Cheng, X. Yan, J. Han, and P. S. Yu. Direct discriminative pattern mining for effective classification. In *International Conference on Data Engineering (ICDE'08)*, pages 169–178, 2008.

[17] P. Compieta, S. D. Martinoc, M. Bertolottoa, F. Ferruccic, and T. Kechadi. Exploratory spatio-temporal data mining and visualiza-

tion. *Journal of Visual Languages and Computing*, 18(3):255–279, 2007.

[18] J. de Leeuw. Applications of convex analysis to multidimensional scaling. In *Recent Developments in Statistics*, pages 133–145, 1977.

[19] R. G. Death and M. K. Joy. Invertebrate community structure in streams of the Manawatu-Wanganui region, New Zealand: the roles of catchment versus reach scale influences. *Freshwater Biology*, 49(8):982–997, 2004.

[20] G. Dong and J. Li. Efficient mining of emerging patterns: Discovering trends and differences. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'99)*, pages 43–52, 1999.

[21] J. Ellson, E. R. Gansner, E. Koutsofios, S. C. North, and G. Woodhull. Graphviz and dynagraph static and dynamic graph drawing tools. In *Graph Drawing Software*, pages 127–148. Springer-Verlag, 2003.

[22] C. Ericson. *Real-Time Collision Detection*. Morgan Kaufmann Publishers Inc., 2005.

[23] European Environment Agency. CLC2006 technical guidelines. Technical Report No 17/2007, Office for Official Publications of the European Communities, 2006.

[24] European Union. Directive 2000/60/ec of the European parliament and of the council of 23 october 2000 establishing a framework for community action in the field of water policy. *Official Journal*, OJ L 327:1–73, 2000.

[25] European Union. Council directive 98/83/ec of 3 november 1998 on the quality of water intended for human consumption. *Official Journal*, OJ L 330:1–32, 2009.

[26] M. Fabrègue, A. Braud, S. Bringay, F. Ber, and M. Teisseire. Orderspan: Mining closed partially ordered patterns. In *International Symposium on Intelligent Data Analysis (IDA'13)*, pages 186–197, 2013.

[27] A. B. Forgang, B. Hamann, and C. F. Cerco. Visualization of water quality data for the Chesapeake Bay. In *IEEE Symposium on Information Vizualization (InfoVis'96)*, pages 417–ff, 1996.

[28] L. Geng and H. J. Hamilton. Interestingness measures for data mining: A survey. *ACM Computing Survey*, 38, 2006.

[29] P. L. Goethals, A. Dedecker, W. Gabriels, S. Lek, and N. Pauw. Applications of artificial neural networks predicting macroinvertebrates in freshwaters. *Aquatic Ecology*, 41:491–508, 2007.

[30] R. L. Graham. An efficient algorithm for determining the convex hull of a finite planar set. *Information Processing Letters*, 1(4):132–133, 1972.

[31] H. Guttinger and W. Stumm. Ecotoxicology an analysis of the Rhine pollution caused by the Sandoz chemical accident, 1986. *Interdisciplinary Science Reviews*, 17(2):127–136, 1992.

[32] D. Hering, R. K. Johnson, S. Kramn, S. Schmutz, K. Szoszkiewicz, and P. F. M. Verdonschot. Assessment of European streams with diatoms, macrophytes, macroinvertebrates and fish: a comparative metric-based analysis of organism response to stress. *Freshwater Biology*, 51(9):1757–1785, 2006.

[33] S. S. P. Hye Won Lee, Kon Joon Bhang. Effective visualization for the spatiotemporal trend analysis of the water quality in the Nakdong river of Korea. *Ecological Informatics*, 5(3):281292, 2010.

[34] D. Keim, G. Andrienko, J.-D. Fekete, C. Görg, J. Kohlhammer, and G. Melancon. Visual analytics: Definition, process, and challenges. In *Information Visualization: Human-Centered Issues and Perspectives*, volume 4950 of *LNCS*, pages 154–175. Springer Berlin Heidelberg, 2008.

[35] D. Kocev, A. Naumoski, K. Mitreski, S. Krstić, and S. Džeroski. Learning habitat models for the diatom community in lake Prespa. *Ecological Modelling*, 221(2):330–337, 2010.

[36] J. B. Kruskal. Nonmetric multidimensional scaling: A numerical method. *Psychometrika*, 29:115129, 1964.

[37] N. Lalande, F. Cernesson, A. Decherf, and M.-G. Tournoud. Implementing the DPSIR framework to link water quality of rivers to land use: methodological issues and preliminary field test. *International Journal of River Basin Management*, pages 1–17, 2014.

[38] G. Lischeid. Non-linear visualization and analysis of large water quality data sets: a model-free basis for efficient monitoring and risk assessment. *Stochastic Environmental Research and Risk Assessment*, 23(7):977–990, 2009.

[39] M. H. Martin. The effects of temperature, river flow, and tidal cycles on the onset of glass eel and elver migration into fresh water in the American eel. *Journal of Fish Biology*, 46(5):891–902, 1995.

[40] M. Meybeck and R. Helmer. The quality of rivers: From pristine stage to global pollution. *Global and Planetary Change*, 1(4):283–309, 1989.

[41] C. Neal, M. Neal, L. Hill, and H. Wickham. The water quality of the river Thame in the Thames basin of south/south-eastern England. *Science of The Total Environment*, 360(13):254–271, 2006.

[42] T. Ofenbock, O. Moog, J. Gerritsen, and M. Barbour. A stressor specific multimetric approach for monitoring running waters in Austria using benthic macro-invertebrates. *Hydrobiologia*, 516(1-3):251–268, 2004.

[43] J. Pei, J. Han, B. Mortazavi-Asl, J. Wang, H. Pinto, Q. Chen, U. Dayal, and M.-C. Hsu. Mining sequential patterns by pattern-growth: The prefixspan approach. *IEEE Transactions on Knowledge and Data Engineering*, 16:1424–1440, 2004.

[44] N. B. Priyantha, H. Balakrishnan, E. D. Demaine, and S. J. Teller. Anchor-free distributed localization in sensor networks. In *International Conference on Embedded Networked Sensor Systems (SenSys'03)*, pages 340–341, 2003.

[45] F. Recknagel, I. Ostrovsky, H. Cao, T. Zohary, and X. Zhang. Ecological relationships, thresholds and time-lags determining phytoplankton community dynamics of lake Kinneret, Israel elucidated by evolutionary computation and wavelets. *Ecological Modelling*, 6:1–3, 2013.

[46] H. Sakoe and S. Chiba. A dynamic programming approach to continuous speech recognition. In *International Congress on Acoustics*, volume 3, pages 65–69, 1971.

[47] A. Sallaberry, N. Pecheur, S. Bringay, M. Roche, and M. Teisseire. Sequential patterns mining and gene sequence visualization to discover novelty from microarray data. *Journal of Biomedical Informatics*, 44:760–774, 2011.

[48] J. Snow. *On the Mode of Communication of Cholera*. John Churchill, 1855.

[49] K. Sugiyama, S. Tagawa, and M. Toda. Methods for visual understanding of hierarchical system structures. *IEEE Transactions on Systems, Man, and Cybernetics*, 11(2):109–125, 1981.

[50] J. J. Thomas and K. A. Cook. *Illuminating the Path*. IEEE Computer Society Press, 2005.

[51] M.-G. Tournoud, S. Payraudeau, F. Cernesson, and C. Salles. Origins and quantification of nitrogen inputs into a coastal lagoon: Application to the Thau lagoon (France). *Ecological Modelling*, 193(1-2):19 – 33, 2006.

[52] W. T. Tutte. How to draw a graph. *Proceedings of the London Mathematical Society*, 13:743–768, 1963.

[53] UNEP/WHO. *Water Quality Monitoring - A Practical Guide to the Design and Implementation of Freshwater Quality Studies and Monitoring Programmes*. UNEP/WHO publications, 1996.

[54] M. Wang, X.-q. Shang, and Z.-h. Li. Sequential pattern mining for protein function prediction. In *Advanced Data Mining and Applications (ADMA'08)*, volume 5139, pages 652–658. 2008.

[55] M. Ward, G. Grinstein, and D. Keim. *Interactive Data Visualization: Foundations, Techniques, and Applications*. A K Peters, Ltd., 2010.

[56] WHO/UNEP. *Water Pollution Control - A Guide to the Use of Water Quality Management Principles*. WHO publications, 1997.

[57] P. C. Wong, W. Cowley, H. Foote, E. Jurrus, and J. Thomas. Visualizing sequential patterns for text mining. In *IEEE Symposium on Information Vizualization (InfoVis'00)*, pages 105–111, 2000.

[58] X. Yan, J. Han, and R. Afshar. Clospan: Mining closed sequential patterns in large datasets. In *SIAM International Conference on Data Mining (SDM'03)*, pages 166–177, 2003.

[59] L. Yang. Visualizing frequent itemsets, association rules, and sequential patterns in parallel coordinates. In *Computational Science and Its Applications (ICCSA'03)*, volume 2667 of *LNCS*, pages 21–30. Springer Berlin Heidelberg, 2003.

[60] Y.-C. E. Yang, X. Cai, and E. E. Herricks. Identification of hydrologic indicators related to fish diversity and abundance: A data mining approach for fish community analysis. *Water Resources Research*, 44, 2008.