1. Introduction

Water pollution is a critical environmental threat (Pimpunchat, Sweatman, Wake, Triampo, & Parshotam, 2009). Their close interaction with human activities such as agriculture, industry, transportation, and sewage discharges makes rivers as the main inland water bodies vulnerable to pollution. Because rivers play a significant ecological role, they are regarded as important indicators of the state of the environment. Population growth, industrialization, uncontrolled urbanization in developing countries, and other anthropogenic activities have led to ever-increasing river pollution (Su et al., 2010).

Total dissolved solids (TDSs), defined as any material in water that will pass through a filter with a pore size of 2 μm or smaller (Berdanier & Ziadat, 2006), is an important water pollution parameter. Most matter dissolved in fresh water consists of inorganic salts, small amounts of organic matter, and dissolved gases (Sawyer, McCarty, & Parkin, 2003). These usually include calcium, magnesium, sodium, potassium cations or carbonate, hydrogen carbonate, chloride, sulfate, and nitrate anions. Concentrations of TDS from natural sources vary from less than 30 mg/l to as much as 6000 mg/l, depending on the solubility of the minerals in different geological regions. Surface water with higher TDS levels are an important driver of the degradation of agricultural land. For example, irrigation with high TDS water is fatal to crop roots and can lead to soil salinization. Studies in Australia have shown that mortality from all categories of ischemic heart disease and acute myocardial infarction increased in a community with high TDS water (Meyers, 1975).

Recent multivariate statistical methods such as cluster analysis, discriminant analysis, and factor and regression analysis have been used to explore the spatial and temporal variations of surface water quality and identify the sources of pollution (Huang, Wang, Lou, Zhou, & Wu, 2010; Lindenschmit, 2006; Maillard & Pinheiro Santos, 2008; Pekey, Karakas, & Bakoglu, 2004; Su et al., 2010; Zhou, Huang, Guo, Zhang, & Hao, 2007), especially
of TDS (Brix, Gerdes, Curry, Kasper, & Grosell, 2010; Etemad-Shahidi, Afshar, Alika, & Moshfeghi, 2009; Magazinovic, Nicholson, Mulcahy, & Davey, 2004). Although these methods explore the main body of knowledge about water pollution, but water pollution is a complex problem composed of issues such as decentralization, modularization, poor structure, and weak predictability (Sokolova & Fernandez-Caballero, 2009). These elements suggest that the problem of water quality inherently includes high rates of imprecision, vagueness, and uncertainty. It is necessary to employ sophisticated data analysis and mining algorithms suitable to solve problems with uncertain and complicated properties.

Examination of the capabilities of the dominance-based rough set (RS) to represent ambiguities and to explore complex relationships between the water quality and environmental parameters is the main objective of this paper. Therefore, the variable consistency dominance-based rough set approach (VC-DRSA) was used for rule extraction and classification of TDS data, with available environmental data (NDVI, LST, precipitation, runoff, RWT) used as explanatory variables.

RS theory is a powerful and flexible mathematical tool for imprecision, vagueness, and uncertainty, first proposed by Pawlak (1982). This algorithm extracts predictive and useful knowledge in the form of rules from imprecise data. The philosophy of RS theory is based on a classification where any union of elementary sets is called a crisp (or precise) set. Granularity of knowledge can be achieved approximately, rather than precisely, defining notions within the available knowledge. This type of set is referred to as the rough set (Triantaphyllou & Felici, 2006). In RS theory, it is possible to associate every set \(X\) with two crisp sets (lower and upper approximations of \(X\)), thus, each vague concept is replaced with a pair of precise concepts. The lower approximation of a concept consists of all objects that definitely belong to that concept. The upper approximation of a concept consists of all objects that may possibly belong to the concept. Hence, a boundary region can be assumed between the lower and upper approximations of a concept. The greater the boundary region, the vaguer the concept. If, for example, the boundary region of a concept is empty, that concept is considered precise.

RS theory overlaps, to some extent, other theories dealing with uncertainty and vagueness, especially the Dempster–Shafer (DS) theory of belief functions (Gorsevski, Jankowski, & Karami, 2008; Slowinski & Stefanowski, 1989) and fuzzy set theory (Dubois & Prade, 1990, 1992; Wygrala, 1984). The difference between the DS and RS theories is that RS theory uses sets of lower and upper approximations to represent knowledge in data collection, while the DS theory uses belief functions represented by lower and upper probability functions (Gorsevski et al., 2008). The approximations for a given data set derived by RS theory are based solely on the data, while the approximations derived by the DS theory involve calculations of belief values using both subjective judgments and data (Dempster, 1967). The differences between the DS and fuzzy sets are long and detailed (see Yao, 1998). Widely-used discipline-specific applications include remote sensing, image and signal processing (Czyzewski, 2003; Czyzewski & Królkowski, 2001; Kostek, 1999, 2005; Peters, Han, & Ramanna, 2001; Tsunomto & Hirano, 2005; Wieczorkowska, Wrobleswski, Synak, & Slezak, 2003), urban planning (Chen, Hipel, Kilgour, & Zhu, 2009; Wang, Hashani, Wang, & Marceau, 2010), and GIScience (Duckham, Mason, Stell, & Worboys, 2001; Worboys, 1998).

In following sections of the paper, first a brief review of the VC-DRSA approach is provided followed by descriptions of the methodology and data, results, and conclusions of the study.

### 2. Methodology and data

#### 2.1. Dominance-based RS approach and VC-DRSA

In a dominance-based RS approach (DSRA), as in equivalence RS, an information table is introduced by the 4-tuple \(S = (U, Q, V, f)\) where \(U\) is a finite set of objects or observations (i.e., the universe); \(e \in D = \{q_1, q_2, \ldots, q_m\}\) is a finite set of attributes (including condition attribute set \(C\) and decision attribute set \(D)\); \(v_i\) is the domain of attribute \(q_i\); and \(V = \cup V_p\). Also, \(f: U \times Q \rightarrow V\) is a total function such that for each \(g \in Q\), \(x \in U\) (information function) (Zhai, Khoo, & Zhong, 2009a).

In this table of information, the values of each condition attribute are preference-ordered and inherently correlated. For example, an increase (or decrease) in a condition attribute value results in an upgrade (or downgrade) of the corresponding decision-class value. This is considered a key difference between the classical and dominance-based approaches. Usually, \(D\) has one member so that \(D = \{d\}\) and it partitions \(U\) into a finite number of classes, such as \(cl = \{cl_t | teT\}\), where \(T = \{1, \ldots, n\}\).

Classes in \(cl\) are ordered in an ascending sequence of class indices. For all \(r, x \in T\) and \(r > s\), the objects included in \(cl_r\) are preferable to those contained in \(cl_s\); therefore, for DSRA, the sets to be approximated are no longer single classes, but the upward and downward unions of the decision classes, respectively (Blaszczynski, Greco, & Slowinski, 2007).

The upward and downward unions of class \(cl_t\) are, respectively:

\[
cl_t^u = \bigcup_{s \geq t} cl_s,
\]

(1)

\[
cl_t^l = \bigcup_{s \leq t} cl_s,
\]

(2)

where \(t = 1, \ldots, n\).

The statement \(x \in cl_t^u\) reads “\(x\) belongs to at least class \(cl_t\);” while \(x \in cl_t^l\) reads “\(x\) belongs to at most class \(cl_t\).” Given a set of attributes \(p \subseteq c\) and \(x \in U\), the granules of knowledge used in DRSA for the approximation of the unions \(cl_t^u\) and \(cl_t^l\) are the open sets defined by the dominance cones with respect to \(x\) (Zhai, Khoo, & Zhong, 2009b). A set of objects dominating \(x\) and dominated by \(x\) are, respectively, called the P-dominating set and the P-dominated set:

\[
D^+_p(x) = \{y \in U : yD_px\}
\]

(3)

\[
D^-_p(x) = \{y \in U : xD_py\}
\]

(4)

Objects satisfying the dominance principle are called consistent and those that violate the dominance principle are called inconsistent. This inconsistency in the sense of the dominance principle is caused by the inclusion of object \(x\) in the upward union of classes \(cl_t^u\) for \(t = 2, \ldots, n\), given that set of criteria \(p \subseteq c\) is true and provided one of the following conditions hold:

- \(x\) belongs to class \(cl_t\) or better, but is p-dominated by object \(y\) belonging to a class worse than \(cl_t\); that is, \(x \in cl_t^u\) but \(D^+_p(x) \cap cl_t^l \neq \emptyset\).
- \(x\) belongs to a worse class than \(cl_t\), but p-dominates object \(y\) belonging to class \(cl_t\) or better; that is, \(x \notin cl_t^u\), but \(D^-_p(x) \cap cl_t^u \neq \emptyset\) (Zhai et al., 2009b).

The p-lower and p-upper approximations of \(cl_t^u\), \(t \in \{1, \ldots, n\}\) (denotation \(p(cl_t^u)\)) and \(p(cl_t^l)\), respectively, for \(p \subseteq c\) are:

\[
p(cl_t^u) = \{x \in U : yD^-_p x \subseteq (cl_t^u)\}
\]

(5)
\[ \hat{p}(c_t^p) = \bigcup_{x \in (c_t^p : D_p^t(x) = \{x \in U : D_p^t(x) \cap c_t^p \neq \emptyset\} \right] \]

The set of all objects belonging to \( c_t^p \) without ambiguity constitutes the \( \hat{p}(c_t^p) \) and the set of all objects that have the possibility of belonging to \( c_t^p \) forms the \( \tilde{p}(c_t^p) \). Analogously, \( p \)-lower and \( p \)-upper approximations of \( c_t^p \) can be defined as:

\[ \tilde{p}(c_t^p) = \{x \in U : y \in D_p^t x \subseteq (c_t^p)\} \]

\[ \hat{p}(c_t^p) = \bigcup_{x \in (c_t^p : D_p^t(x) = \{x \in U : D_p^t(x) \cap c_t^p \neq \emptyset\} \right] \]

All objects belonging to both \( c_t^{p,c} \) and \( c_t^p \) with some ambiguities, constitute the \( p \)-boundaries of \( c_t^{p,c} \) and \( c_t^p \), which are denoted by \( p_{\cap}(c_t^p) \) and \( p_{\cap}(c_t^{p,c}) \), respectively. They can be represented in terms of their upper and lower approximations as:

\[ p_{\cap}(c_t^p) = \tilde{p}(c_t^p) - \hat{p}(c_t^p) \]

\[ p_{\cap}(c_t^{p,c}) = \tilde{p}(c_t^{p,c}) - \hat{p}(c_t^{p,c}) \]

The accuracy of the approximation of \( c_t^{p,c} \) and \( c_t^p \) for all \( t = 1, \ldots, n \) and for any \( p \subseteq c \), respectively, is defined as:

\[ \alpha_p(c_t^{p,c}) = \frac{|\tilde{p}(c_t^{p,c})|}{|\tilde{p}(c_t^{p,c})|} \]

\[ \alpha_p(c_t^p) = \frac{|\tilde{p}(c_t^p)|}{|\tilde{p}(c_t^p)|} \]

The quality of approximation of classification \( c_t \) using criteria from set \( p \subseteq c \) can then be defined as:

\[ \gamma_p(c_t) = \frac{|U - \bigcup_{c \in (1, \ldots, n)} p_{\cap}(c_t^{p,c})|}{|U|} \]

\[ \gamma_p(c_t^p) = \frac{|U - \bigcup_{c \in (1, \ldots, n)} p_{\cap}(c_t^p)|}{|U|} \]

In DRSA, the lower approximation of a union of ordered classes contains only consistent objects; however, analysis of the large real-life data sets shows that application of DRSA results in larger differences between the lower and upper approximations of the decision classes for some multi-criteria classification problems. Consequently, since the decision rules induced by the lower approximations are supported by the few objects, they are thought to be weak (Błaszczynski et al., 2007).

Extensions of DRSA, such as VC-DRSA, have been proposed to relax the condition for inclusion of an object in the lower approximation. This approach also allows for some inconsistency in the lower approximations of sets controlled by the consistency level parameter. The extended lower approximation of a union of classes is defined as a set of objects for which the consistency measure satisfies a user-defined threshold value. For any \( q \in C, x \in U \) that belongs to \( c_t^{p,c} \) at consistency level \( 1 \in (0, 1] \), if \( x \in c_t^{p,c} \) and at least \( 1 \times 100\% \) of all objects \( y \in U \) dominating \( x \) with respect to \( p \) also belong to \( c_t^{p,c} \):

\[ \frac{|D_p^t(x) \cap c_t^{p,c}|}{|D_p^t(x)|} \geq l \]

where \(|\bullet|\) denotes the cardinality of a set. Level \( l \) is the consistency level, because it controls the degree of consistency between the objects quantified as belonging to \( c_t^{p,c} \) (Błaszczynski et al., 2007). Consistency level controls the strength of each rule that depends on the number of objects belonging to it.

In hydrology, especially in water quality problems, attributes are defined as preference-ordered scales; thus, using classical RS may not lead to exploration of the explicit relationships and knowledge between environmental variables and water quality parameters. Dominance-based RS (Greco, Matarazzo, & Slowinski, 2002) is an effective and reliable approach for solving this problem. Therefore VC-DRSA was used to explore the underlying knowledge related to data for TDS.

2.2. Study area

Jajrood River is located in the Latyan watershed north of Tehran. This watershed is located between 51.220–51.550° longitude and 35.450–36.500° latitude and is 790 km² in area (Fig. 2). Jajrood River is a main source of drinking water for Tehran. Because of this, many studies have addressed the problem of water supply in the area (e.g. Mahjouri & Kerachian, 2011; Mohammadkhian, Ahmadi, & Jafari, 2011; Nikoo, Kerachian, Malakpour-Estalaki, Bashi-Azghadi, & Azimi-Ghadikolaei, 2010; Padakhtii, Nabi Bidhendi, Torabian, Karbassi, & Yunesian, 2010; Razmkhah, Aribashamchi, & Torkian, 2010; Saeedi, Hosseinizadeh, & Rajabzadeh, 2011; Zeinivand & Smedt, 2009). The maximum elevation of Jajrood basin is 4375 m and the river is 1560 m above sea level. Total area of the sub-basin up to Latyan Dam is 7149 km², which is approximately 24.6% of the sub-basin area. The length of the main river from upslope to the Latyan Dam is about 38 km. Mean annual precipitation is about 573 mm, of which approximately 191 mm falls as snow. The mean yearly temperature is about 26°C with extremes of –8°C in January and 32°C in July.

Agriculture, including irrigated crops and orchards, are concentrated in the valleys and along the river, where the use of nitrogenous fertilizers is common. No wastewater and sewage treatment facilities exist in the area; municipal wastewater from cities, villages, restaurants, and industrial sites are mostly discharged into the river. In addition, the high population density and growth in Tehran have led to intensive land use and environmental impacts, particularly on water quality (Razmkhah et al., 2010). The location of Latyan basin and dam with the pollution monitoring sites are shown in Fig. 1.

2.3. Dataset

2.3.1. Pollution data

Eleven water quality parameters measured at 23 monitoring sites (Fig. 2) from 2001 to 2007 were acquired from the Water and Waste Organization of Tehran. These are chemical oxygen demand (COD), biological oxygen demand (BOD5), dissolved oxygen (DO), phosphate (PO4), bromide (Br), fluoride (F), total dissolved solids (TDS), sulfate (SO4), ammonia (NH3), nitrate (NO3), and nitrite (NO2). Statistical characteristics of these parameters for 2001–2007 are listed in Table 1.

2.3.2. Environmental variables

Selection of the environmental variables was limited by the availability of data; there is a shortage of reliable data for temporal variability of the land surfaces. To provide temporal information about surface vegetation cover and temperature, NDVI and LST products of the MODIS data were downloaded from the Archive Center of NASA (http://modis.gsfc.nasa.gov/data/dataprod/index.php). The study area is usually covered by the snow from January through March, so the NDVI and LST data are limited to the remaining nine months. Other relevant data with temporal...
dimensions, including the precipitation, runoff, and RWT, were processed and used as the environmental variables (Table 2).

### 2.4. Data preparation

The study area was divided into 23 sub-basins according to the 23 watershed areas corresponding to the 23 pollution monitoring sites. Five buffer areas (250, 500, 750, 1000, and 1500 m) from the tributaries of each sub-basin were defined to evaluate the effects of NDVI and LST on TDS. Mean NDVI and LST were extracted for each buffer zone. Monthly precipitation for each sub-basin area was calculated by interpolation from the existing point measurements of precipitation.

### Table 1

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Mean</th>
<th>Standard deviation</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>COD</td>
<td>3.13</td>
<td>1.33</td>
<td>0.80</td>
<td>15.00</td>
</tr>
<tr>
<td>BOD5</td>
<td>1.78</td>
<td>0.88</td>
<td>0.20</td>
<td>8.10</td>
</tr>
<tr>
<td>DO</td>
<td>8.30</td>
<td>1.09</td>
<td>4.70</td>
<td>11.60</td>
</tr>
<tr>
<td>PO4</td>
<td>0.05</td>
<td>0.05</td>
<td>0.00</td>
<td>0.60</td>
</tr>
<tr>
<td>Br</td>
<td>0.36</td>
<td>0.17</td>
<td>0.00</td>
<td>1.30</td>
</tr>
<tr>
<td>F</td>
<td>0.19</td>
<td>0.09</td>
<td>0.00</td>
<td>0.75</td>
</tr>
<tr>
<td>TDS</td>
<td>245.10</td>
<td>82.60</td>
<td>125.34</td>
<td>728.06</td>
</tr>
<tr>
<td>SO4</td>
<td>50.03</td>
<td>35.66</td>
<td>7.00</td>
<td>342.00</td>
</tr>
<tr>
<td>NH3</td>
<td>0.15</td>
<td>0.15</td>
<td>0.00</td>
<td>1.50</td>
</tr>
<tr>
<td>NO2</td>
<td>4.81</td>
<td>2.41</td>
<td>1.30</td>
<td>18.60</td>
</tr>
<tr>
<td>NO3</td>
<td>0.04</td>
<td>0.04</td>
<td>0.00</td>
<td>0.21</td>
</tr>
</tbody>
</table>

Table 2

<table>
<thead>
<tr>
<th>Statistics/variables</th>
<th>Means</th>
<th>Standard deviation</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>Annual Precipitation (mm)</td>
<td>31.40</td>
<td>36.08</td>
<td>0.00</td>
<td>197.46</td>
</tr>
<tr>
<td>Runoff (m³/s)</td>
<td>1.94</td>
<td>3.56</td>
<td>0.00</td>
<td>37.96</td>
</tr>
<tr>
<td>RWT (°C)</td>
<td>10.90</td>
<td>3.32</td>
<td>1.00</td>
<td>20.00</td>
</tr>
<tr>
<td>LST_buff_250</td>
<td>0.21</td>
<td>0.04</td>
<td>0.06</td>
<td>0.34</td>
</tr>
<tr>
<td>LST_buff_500</td>
<td>0.21</td>
<td>0.04</td>
<td>0.05</td>
<td>0.32</td>
</tr>
<tr>
<td>LST_buff_750</td>
<td>0.21</td>
<td>0.04</td>
<td>0.04</td>
<td>0.32</td>
</tr>
<tr>
<td>LST_buff_1000</td>
<td>0.22</td>
<td>0.03</td>
<td>0.04</td>
<td>0.30</td>
</tr>
<tr>
<td>LST_buff_1500</td>
<td>0.23</td>
<td>0.04</td>
<td>0.03</td>
<td>0.33</td>
</tr>
<tr>
<td>NDVI_buff_250</td>
<td>0.30</td>
<td>0.09</td>
<td>-0.03</td>
<td>0.57</td>
</tr>
<tr>
<td>NDVI_buff_500</td>
<td>0.28</td>
<td>0.08</td>
<td>-0.04</td>
<td>0.48</td>
</tr>
<tr>
<td>NDVI_buff_750</td>
<td>0.25</td>
<td>0.07</td>
<td>-0.05</td>
<td>0.45</td>
</tr>
<tr>
<td>NDVI_buff_1000</td>
<td>0.24</td>
<td>0.07</td>
<td>-0.05</td>
<td>0.45</td>
</tr>
<tr>
<td>NDVI_buff_1500</td>
<td>0.20</td>
<td>0.07</td>
<td>-0.06</td>
<td>0.40</td>
</tr>
</tbody>
</table>

LST_buff_... = land surface temperature with buffer... NDVI_buff_... = NDVI with buffer...
Fluctuations of local variations in TDS often lead to a decrease in performance of rule induction and classification. Applying a low-pass filter mitigates local variations and decreases the width of the boundary regions, considerably enhancing the results; hence, a moving average filter was applied to the dataset. Both raw and filtered datasets were used for rule induction to study the effects of filtering.

The TDS dataset was classified into the three qualitative classes: low (TDS < 200), moderate (300 > TDS > 200) and high (TDS > 300). JMAF software (Blaszyński, Greco, Matarazzo, Słowiński, & Szelag, 2009) was employed for analysis and extraction of rules from the data. JMAF is a RS data analysis framework written in Java based on the Java RS (jRS) library. The JMAF and jRS libraries (jRS and JMAF teams at Poznan University of Technology) implemented analytical methods provided by the dominance-based RS (http://www.cs.put.poznan.pl/jblaszyński/Site/jRS.html). Using the spatio-temporal similarities of the water pollution attributes, the 23 monitoring sites were clustered into the 8 homogeneous groups. VC-DRSA was applied to the sub-basins of groups 1 and 2.

3. Results and discussion

3.1. Application of VC-DRSA

Results of the VC-DRSA analysis are presented in four parts: the quality of approximation with and without the moving average filtering, rule generation, rule evaluation, and analysis of the significance of the criteria and attributes.

3.1.1. Quality of approximation

Results of VC-DRS analysis for filtered and unfiltered data of groups 1 and 2 are presented in Table 3. The filtered data in group 1 exhibited slightly higher quality with a consistency level of 1; however, a decrease in the consistency level from 1 to 0.9 or 0.8 resulted in lower-quality for filtered data than the raw data. The quality of the approximations for group 2 was not affected by the filtering and the highest quality of both the filtered and unfiltered data was 1.

Samples from group 1 were classified into the five temporally similar sub-groups and the approximation algorithms were applied to each sub-group separately. The results revealed that, for the three temporal sub-groups of group 1 (April–May, June–July, November–December) filtering increased the quality of the approximation. Nevertheless, the quality of the approximation was not much affected by filtering the August–September and October sub-groups (Fig. 2).

TDS values for August–September and October sub-groups were less affected by the local fluctuations than the other three sub-groups. Table 4 shows the accuracy of classification with three levels of consistency for the filtered and unfiltered datasets. TDS values for the August–September sub-group were divided into low, moderate, and high classes. The other four sub-groups were divided into the low and moderate classes.

The lowest-accuracy values mainly belonged to the moderate class of the April–May sub-group. This class showed the highest rate of accuracy enhancement from filtering. TDS values for the moderate class of this sub-group showed that most members were located in the boundary regions between low and moderate, which may be the main reason for the low accuracy of this class. Moving average filter in this class led to a movement of data towards the low-class data space, enhancing classification.

It is worth noting that the effects of filtering the temporal sub-groups with TDS values close to the boundaries (e.g., April–May) were considerable. For sub-groups that exhibited considerable differences between TDS values (e.g., August–September and October), the effects of filtering were negligible. This means that filtering can be considered a complementary operation that enhances consistency and reliability of the rules. However, relatively high accuracies for most classes and sub-groups indicate good relationships between the TDS classes and the environmental variables.

To better understand the effects of moving average filtering on the accuracy of results, boundary samples from the approximation of filtered and unfiltered datasets were examined. Results showed that moving average filtering narrowed boundary regions and significantly decreased the number of boundary samples. For example, results for April–May and November–December sub-groups are presented in Figs. 3 and 4. In all situations, the effects of filtering on the decrease of the boundary samples were more significant than the decrease in consistency levels.

<table>
<thead>
<tr>
<th>Table 3</th>
<th>Quality of approximation resulting from VC-DRSA on filtered and non-filtered dataset with consistency level (CL) of 1, 0.9 and 0.8.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group 1</td>
<td></td>
</tr>
<tr>
<td>Non filtered</td>
<td>Filtered</td>
</tr>
<tr>
<td>CL = 1</td>
<td>0.76</td>
</tr>
<tr>
<td>CL = 0.9</td>
<td>1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 4</th>
<th>Accuracy of classification resulting from VC-DRSA with CL of 1, 0.9, and 0.8.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Temporal period/ class</td>
<td>April–May</td>
</tr>
<tr>
<td>AM H(NF)</td>
<td>CL = 1</td>
</tr>
<tr>
<td>AM H(F)</td>
<td>0.78</td>
</tr>
<tr>
<td>AM M(NF)</td>
<td>0.19</td>
</tr>
<tr>
<td>AM M(F)</td>
<td>0.7</td>
</tr>
<tr>
<td>AM L(NF)</td>
<td>0.79</td>
</tr>
<tr>
<td>AM L(F)</td>
<td>0.93</td>
</tr>
</tbody>
</table>

AM = at most, AL = at least, H = high, M = moderate, L = low, F = filtered dataset, NF = no filtered dataset, (these abbreviations are reliable for Table 9, and Figs. 6 and 7).
3.1.2. Rule induction

The minimum cover rule induction procedure does not contain redundant rules and was used for the filtered and unfiltered data-sets for groups 1 and 2 and the five subgroups of group 1. Consistency levels were set to 1, 0.9, and 0.8. Rules were generated for each group, sub-group, and for 80% of samples that were randomly selected. The remaining 20% were used to evaluate the generated rules. By eliminating rules with confidence levels lower than 0.9...
than 0.9 and coverage factors lower than 0.3 in group 1 (239 samples), 13 of the 58 rules were left (Table 5). The reduced rule set contained 8, 4, and 1 rule for the low, moderate, and high classes, respectively. Table 5 shows that NDVI in buffers of 250, 1000, and 1500 m and RWT had the most important roles in rule generation, and precipitation and LST had the least contributions. The highest values for strength, coverage factor, and support were in the at least moderate class and the lowest values were in the at most moderate class.

Only four rules were generated for group 2 with 36 samples (Table 6). Of the 13 variables considered, only three variables (NDVI of buffers of 250 and 1000 m, and RWT) were used in the generated rules. Rule generation was carried out separately for each sub-group of group 1. The quantity and quality of the rules for each sub-group were considerably different, particularly the roles of the variables in the generated rules, although there were greater similarities between the adjacent sub-groups (Table 7).

### 3.1.3. Rule validation
Results of the evaluation of hit rates for the generated decision rules for groups 1 and 2 are presented in Fig. 5. It can be seen that the moving average filter created contrasting effects on the results. The filtered data exhibited better performance for group 2. For group 1, filtering resulted in a decrease in the accuracy of the rules. These differences may have resulted from the differences between the temporal resolutions of the two groups. Degradation of temporal precision as a result of filtering the monthly data in group 1 led to lower accuracy and filtering the five-month period data in group 2 led to higher accuracy. Nevertheless, changes in the consistency levels did not have much effect on the accuracy of the results for either group.

Generalization abilities of the rules of both groups were evaluated by classifying samples from one group using the rules of the other group. The rules of group 1 classified the filtered and unfiltered data of group 2 with three consistency levels (CL) of 1, 0.9 and 0.8.

### Table 7
Details of the generated minimum cover rules for August–September sub-group of group 1, with consistency level of 0.8 and confidence level greater than 0.9.

<table>
<thead>
<tr>
<th>Rule Num</th>
<th>Condition part</th>
<th>Decision part</th>
<th>Strength</th>
<th>Coverage factor</th>
<th>Support</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 (RWT ≤ 10.0)</td>
<td></td>
<td>(TDS ≤ low)</td>
<td>0.09</td>
<td>0.35</td>
<td>6</td>
</tr>
<tr>
<td>8</td>
<td>Runoff &gt; 0.04</td>
<td>(TDS ≤ moderate)</td>
<td>0.64</td>
<td>0.73</td>
<td>41</td>
</tr>
<tr>
<td>9</td>
<td>(NDVI_buff_500 ≤ 0.27) &amp; (NDVI_buff_1000 &gt; 0.19)</td>
<td>(TDS ≤ moderate)</td>
<td>0.56</td>
<td>0.64</td>
<td>36</td>
</tr>
<tr>
<td>10</td>
<td>(Runoff ≤ 0.03) &amp; (NDVI_buff_1000 ≤ 0.18)</td>
<td>(TDS &gt; high)</td>
<td>0.06</td>
<td>0.5</td>
<td>4</td>
</tr>
<tr>
<td>14</td>
<td>(RWT &gt; 11.0) &amp; (NDVI_buff_1000 ≤ 0.21)</td>
<td>(TDS &gt; moderate)</td>
<td>0.4</td>
<td>0.57</td>
<td>27</td>
</tr>
<tr>
<td>15</td>
<td>(RWT &gt; 13.0) &amp; (NDVI_buff_1500 ≤ 0.21)</td>
<td>(TDS &gt; moderate)</td>
<td>0.37</td>
<td>0.51</td>
<td>24</td>
</tr>
<tr>
<td>17</td>
<td>(RWT &gt; 12.0) &amp; (NDVI_buff_1000 ≤ 0.22) &amp; (NDVI_buff_500 ≤ 0.26) &amp; (NDVI_buff_1500 ≤ 0.21)</td>
<td>(TDS &gt; moderate)</td>
<td>0.5</td>
<td>0.68</td>
<td>32</td>
</tr>
<tr>
<td>20</td>
<td>(RWT &gt; 11.0) &amp; (NDVI_buff_500 &gt; 0.26) &amp; (NDVI_buff_1000 &lt; 0.23) &amp; (NDVI_buff_1500 &lt; 0.21)</td>
<td>(TDS &gt; moderate)</td>
<td>0.23</td>
<td>0.32</td>
<td>15</td>
</tr>
</tbody>
</table>

### Table 8
Accuracy (%) of rules for prediction of data of different sub-groups.

<table>
<thead>
<tr>
<th>Predictor/predicted</th>
<th>April–May</th>
<th>June–July</th>
<th>August–September</th>
<th>October</th>
<th>November–December</th>
</tr>
</thead>
<tbody>
<tr>
<td>April–May</td>
<td>83</td>
<td>63</td>
<td>27</td>
<td>33</td>
<td>62</td>
</tr>
<tr>
<td>June–July</td>
<td>85</td>
<td>77</td>
<td>27</td>
<td>81</td>
<td>68</td>
</tr>
<tr>
<td>August–September</td>
<td>90</td>
<td>65</td>
<td>83</td>
<td>81</td>
<td>78</td>
</tr>
<tr>
<td>October</td>
<td>25</td>
<td>37</td>
<td>64</td>
<td>58</td>
<td>78</td>
</tr>
<tr>
<td>November–December</td>
<td>19</td>
<td>37</td>
<td>64</td>
<td>81</td>
<td>75</td>
</tr>
</tbody>
</table>

![Fig. 5. Accuracy (%) of rules for filtered (F) and not filtered (NF) data of groups 1 (G1) and 2 (G2) with three consistency levels (CL) of 1, 0.9 and 0.8.](image-url)
tered samples of group 2 with accuracies of 69% and 94%, respectively. However, the rules of group 2 showed less accuracy at 37% and 23%, respectively, for filtered and unfiltered samples of group 1. The higher temporal detail in group 1 may have resulted in higher generalization of the extracted rules.

The generalization abilities of the rules of group 1 for each sub-group were evaluated to define similarities between the sub-groups of group 1 and the global trend of the region (Fig. 6). The results show that April–May and October are most similar to the global trend. June–July and August–September showed the lowest similarities.

Analysis and evaluation of the generalization abilities of the rules for each sub-group is useful for practical consideration of the temporal autocorrelation. Table 8 shows the results for evaluation and reveals that the August–September and June–July sub-groups have the highest prediction and generalization abilities of the sub-groups. October and April–May sub-groups showed the poorest generalization abilities. November–December was well predicted by the other sub-groups, and August–September was the least predicted. An examination of the temporal patterns of autocorrelation between the sub-groups shows that, for example, the November–December sub-group predicted the samples of October and August–September sub-groups more accurately than sub-groups with higher temporal distances (June–July and April–May). An evaluation of the rules of sub-groups and their corresponding test data show that April–May and August–September had the highest accuracies, and October had the lowest accuracy (Table 8).

4. Concluding remarks

In this paper, VC-DRSA data-mining algorithms were used for exploration of relationships between TDS data and the available environmental data (NDVI, LST, precipitation, runoff, RWT). VC-DRSA demonstrated that it can handle the attributes/criteria with or without preferences as an operational tool for rule induction for water quality studies when it is possible to relax the consistency level to less than 1. It is important to note that noise and local fluctuations in the environmental variables usually affect TDS analysis and rule induction. Moving average filtering smoothed the data, decreased the noise, and reduced the width of the boundary region between the lower and upper approximations. The results of this test demonstrated, particularly for overlapping classes, that smoothing with the simple moving average procedure can enhance classification and rule induction results considerably. The high accuracy levels resulting from rule validation showed that the rules induced from the selected criteria illuminated important relationships between pollution source indicators such as NDVI and the TDS data. This research indicated that the RS approach provides a reliable tool for the exploration of the complex rules and relationships inherent in the usually available water quality data.

References


