ORIGINAL ARTICLE

Application of *fuzzy logic and data mining* techniques as tools for qualitative interpretation of acid mine drainage processes

**J. Aroba · J. A. Grande · J. M. Andújar ·
M. L. de la Torre · J. C. Riquelme**

Abstract In this article, a set of clustering algorithms based on Fuzzy Logic and Data Mining are applied, allowing to obtain data in the form of linguistic rules and charts about the behaviour of the Tinto and Odiel river estuary (SW Spain) affected by Acid Mine Drainage (AMD). In order to provide researchers with no skills on data mining techniques an easy and intuitive interpretation, we have developed a computer tool based on fuzzy logic that allows immediate qualitative analysis of the data contained in a data from the estuary water chemical analyses, and serves as a contrast to functioning models previously proposed with classical statistics.

Keywords AMD · Fuzzy logic · Data mining · Tinto and Odiel rivers · Heavy metals · Spain

Introduction

In this article, we apply a new computer tool: Predictive Fuzzy Rules Generator (PreFuRGe) (Aroba 2003),

J. A. Grande · M. L. de la Torre (✉)
Grupo de Recursos y Calidad del Agua,
Escuela Politécnica Superior. Universidad de Huelva,
21819 La Rábida-Huelva, Spain
e-mail: mltorre@uhu.es

J. Aroba · J. M. Andújar
Grupo de Control y Robótica,
Escuela Politécnica Superior. Universidad de Huelva,
21819 La Rábida-Huelva, Spain

J. C. Riquelme
Departamento de Lenguajes y Sistemas Informáticos,
Escuela Técnica Superior de Ingeniería Informática.
Universidad de Sevilla, 41012 Sevilla, Spain

that allows qualitative interpretation of data recorded in a database relative to the chemistry of water. Specifically, we aim at finding information, in principle hidden and not likely to be detected by means of classical statistical techniques, that can help characterising and interpreting discharge-rainfall-dissolution processes that occur in the estuary of the Tinto and Odiel rivers (SW Spain) affected by Acid Mine Drainage (AMD) processes and by the presence of an industrial site discharging effluents to the estuary (Fig. 1).

Contamination of fluvial origin transported by run-offs coming from the existing mining operations is known as AMD, being one of the most serious types of water contamination because of its nature, extent and solving difficulty (Azcue 1999), as well as remediation costs (Commonwealth of Pennsylvania 1994). Rivers affected by this kind of contamination are characterized by their acidity, as well as by high sulphate and heavy metal content in their waters and the metal content of their sediments (USEPA 1994). Damages range from sublethal alterations for some individuals of the affected ecosystems in cases of slight pollution, with associated problems of bioaccumulation and biomagnification (Nebel and Wrigth 1999), to disappearance of the fluvial fauna, and loss of water resources, as water becomes useless for human, farming or industrial consume (Sáinz et al. 2004).
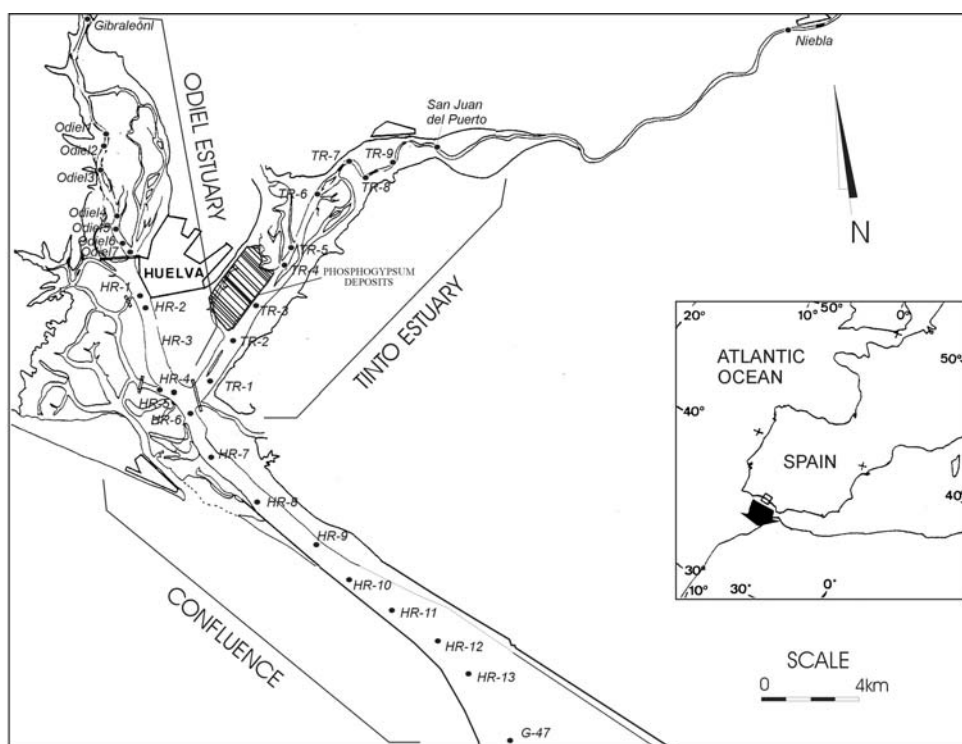
This kind of contamination originates when a sulphurous mineral gets in contact with oxygen and atmospheric humidity (Sáinz 1999). A complex device begins then on the mineral surface, starting with oxidation of sulphides, which are very insoluble, and then transforming them into sulphates with production of acid. The kinetics of this oxidation is very slow, between 1.08×10^{-15} and 1.8×10^{-14} mol/(cm^2 s), and may

Fig. 1 Location map

increase its speed up to 100× by the presence of ferric ion (Dogan 1999) and by the action of catalysing bacteria (Nicholson 1994). Together with pyrite oxidation, secondary reactions occur among products of the previous reactions and the remaining minerals present in the environment (Förstner and Wittmann 1983), the final result being a set of soluble contaminants deposited on the mineral, which are then dissolved and dragged by rain- or runoff water. This originates a contaminant liquid flow that carries acidity, sulphates and heavy metals to watercourses. The rate expressions for the oxidation of FeS_2 (s) and ferrous iron are precisely described in Younger et al. (2002).

Although the phenomenon of sulphide oxidation is a natural fact, production rates caused by mining allow us to distinguish between the natural geochemical process, of temporal patterns common in geology, and AMD, an anthropogenic process caused both by cropped out mineral mass amounts and by contact surface increase due to mining activities and granulometry decrease (EMCBC 1996).

Estimations of damage remediation costs range from two to five billion dollars for Canada (Feasby et al. 1997). This same amount is estimated for remediation costs just for the State of Pennsylvania (Commonwealth of Pennsylvania 1994). World remediation costs are estimated to be over ten billion dollars (Weatherell et al. 1997), although nowadays in the USA, the mining sector spends more than a million

dollars daily on the treatment of these acidic effluents because, according to the US Bureau of Mines, the past mining activity had already affected more than 20,000 km of water courses in the USA (Sáinz 1999).

There are other techniques which are applied to the kind of problems dealt with in this work, such as N-Land (Ward et al. 1994), Multidimensional Scaling (Ong et al. 2004) or the best known and used Self-organizing Maps (SOMs) (Lu and Lo 2002). The self-organizing map is a neutral network model and algorithm that implements a characteristic nonlinear projection from the high-dimensional space onto a low-dimensional array of neurons (Lu and Lo 2002). Nevertheless, from our point of view, the fuzzy logic-based techniques we use, and more specifically, PreFuRGe, have the same advantages as SOMs techniques: it is easy to understand and it classifies very well; it also removes SOMs' possible disadvantages, such as very expensive computational calculations, defective work with scarce or uncertain data, artificial definition of the mesh that can influence the cluster formation, and the need of building many maps to find out a good final map, etc.

General setting

The estuary of the Tinto and Odiel Rivers is on a mesotidal (mean tidal range of 2.10 m) mixed-energy coast (Borrego 1992). The tidal wave moves along the Odiel

estuary following a hypersynchronic model (Borrego et al. 1993), ranging within the estuary from 4 m during spring tides to 0.5 m during extreme neap tides. The volumes of water that come in and out of the estuary from the open sea (tidal prism) during a tidal half-cycle (6 h) range from 3,734 m^3 during a mean neap tide (1 m of tidal range) to 81.76 Hm^3 in a mean spring tide (3 m of tidal range) (Borrego 1992 in Grande et al. 2000).

The volume of freshwater inflow from the Tinto and Odiel Rivers to the inner zone of the estuary reflects a significant seasonal and year-to-year variation. Therefore, from 1960 to 1996, the average monthly inflow of river water was 49.8 Hm^3, and the average yearly inflow was 598 Hm^3. The marked seasonality of this inflow is due to a rainy season from October to March when the inflow may reach 100 Hm^3 per month and a dry period (from May to September) with average monthly volumes of less than 5 Hm^3 (Borrego 1992 in Grande et al. 2003a).

The variation of the volumes of water from the regime of river-water inflow and from the tidal prism gives rise to different types of mixing models within the estuary. During the rainy season (with an average volume of flow of 21 $m^3 s^{-1}$), the mixing conditions within the estuary can be defined as "partial stratification" (following the criteria of Simmons 1955) for any tidal situation. However, during the dry months (volume of flow less than 6 $m^3 s^{-1}$), the situation can be described as a "good mixing" (Borrego 1992 in Grande et al. 2000).

Major industrial activities are located on the western bank of the Odiel estuary. In addition, there are large waste deposits of phosphogypsum on the eastern bank of the Tinto near the junction of the two rivers. These deposits are drained by a small tributary which also carries effluent from nearby sewage treatment plants (Elbaz-poulichet et al. 1999a). The phosphogypsum wastes (about 10^{10} kg) cover an area of approximately 4×10^6 m^2 (Travesi et al. 1997 in Elbaz-Poulichet et al. 2000).

AMD processes undergone by the drainage network in the regional environment and its impact on the estuary have been widely described by different authors: Borrego (1992), Elbaz-Poulichet and Leblanc (1996), (1999a, b, 2000), Elbaz-Poulichet and Dupuy (1999), Braungardt et al. (1998), Cruzado et al. (1998), Sáinz (1999), Davis et al. (2000), Grande et al. (2000, 2003a, b, c, d, e), and Sáinz et al. (2002, 2003).

Objectives and methods

The main objective of the present work is the characterisation, by means of fuzzy logic and data mining

techniques, of discharge/rainfall/dissolution processes in an estuarine environment affected by AMD processes and enduring at the same time the discharge of industrial effluents from a chemical industrial site, both phenomena subject to tidal influence. The obtained results can serve as a validation to the functioning models proposed by Borrego et al. (2002), Grande et al. (2000), (2003a, c), and Sáinz et al. (2002, 2003), and as a contrast to numerous works on the estuary by other authors.

Sampling and analytical methodology

The sampling campaign was carried out in June 1997. Sampling stations or points were established (Fig. 1), where a sample of surface water (5 l per station) was taken using a Niskin sampling bottle. pH was measured with a TURO 130 sampling probe.

Water samples were filtered immediately after collection, stored at 4°C in the dark, and analysed with standard methods within 1 or 2 days. Nutrients were determined colorimetrically, and major ions and metal concentrations were determined by atomic absorption spectroscopy (Perkin-Elmer 3110) after total digestion in a mixture of HF–HCl_4–HNO_3. Table 1 shows the results obtained for each sample for the different variables.

Fuzzy logic and data mining

Fuzzy logic (Zadeh 1965) works with reasoning rules very close to the human way of thinking, which is approximate and intuitive. The main characteristic of fuzzy logic is that it allows us to define values without specifying a precise value, something which is not possible with classical logic, upon which computer development has been based so far. In classical logic, the membership to one class or set is binary, i.e., one is either member or not, so that only two precise values are worked with (1 and 0, yes or no). Thus, if "very low pH" is defined for some samples, it is evident that a sample with pH 2 belongs to the cluster and another one with pH 6 does not, but how do we classify a sample with pH 4.99? It is precisely in the answer to this kind of question where classical logic shows its limitations to us.

Fuzzy logic allows us to associate each sample with a certain degree of fulfilment of the "very low pH" prototype. This grade is called "membership grade" $\mu_{VLpH}(x)$ of the element $x \in X$ to the set "very low pH". The set X is called universe of discourse—range of values—of the variable x. The range of μ_{VLpH} ranges

Table 1 Data from chemical analyses (concentration in mg/l)

| Sample	$SO_4^=$	SiO_2	PO_4^{3-}	Na^+	NO_3^-	K^+	Mg^{2+}	Ca^{2+}	Cl^+	pH	Li	Cu	Zn	As	Rb	Sr
hr3	1.86	0.16	0.04	7.42	0.54	493	1.77	273	1.59	7.32	0.12	0.35	0.55	0.04	0.08	5.52
g47	1.66	0.05	0.01	8.28	11.3	406	1.38	421	1.6	7.6	0.13	0.33	0	0	0.11	7.37
hr5bis	4.23	0.31	0.05	7.69	10.5	371	1.61	366	1.83	7.2	0.11	0.33	0.39	0.03	0.08	5.11
hr3bis	3.16	1.89	0.04	10.5	1.26	162	0.6	166	1.86	7.2	0.05	0.3	0.14	0.02	0.04	2.39
Odiel5	1.64	12.4	0	10.1	3.83	177	0.83	208	0.93	4.2	0.07	1.38	3.49	0	0.05	0.53
tr3	2.59	13.2	0.32	7.86	0.66	340	1.53	432	1.83	3.1	0.09	1.97	4.02	0.21	0.08	5.48
g47tris	2.69	0.05	0.01	8.25	5.76	388	1.36	341	1.93	8	0.12	0.33	0	0	0.1	7.22
tr9	2.31	5.8	0.01	7.72	6.6	184	1.39	301	0.62	2.51	0.09	6.63	16.8	0.02	0.05	3.54
hr4	1.97	0.05	0	8.3	6.34	377	1.35	251	1.65	7.32	0.12	0.34	0.37	0.03	0.09	5.58
tr7	2.31	8.32	0.02	4.05	7.66	126	0.53	239	0.94	2.87	0.08	8.36	20.7	0	0.05	2.47
hr1bis	2.7	0.05	0.13	8.61	1.75	370	1.32	390	1.68	7.4	0.15	0.46	0.66	0.04	0.13	7.41
Odiel4	1.57	5.41	0	3.12	2.44	193	0.97	218	0.85	4	0.07	1.69	4.09	0	0.06	3.19
hr13bis	2.56	0.05	0.03	7.82	0.59	394	1.13	263	1.95	8	0.1	0.32	0	0	0.08	5.38
hr4bis	1.73	0.54	0.07	9.29	4.45	343	1.2	368	1.97	7.2	0.12	0.34	0.32	0.03	0.08	5.56
tr8	1.14	19.3	0.04	3.41	6.38	137	0.36	246	1.52	2	0.04	2.05	10.2	0	0.04	1.7
hr9bis	2.28	0.33	0.03	6.42	5.87	371	1.61	366	1.95	7.85	0.1	0.31	0.05	0.01	0.08	5.06
hr12	1	0.05	0	8.54	4.32	334	1.53	342	1.72	8.08	0.1	0.33	0.21	0.02	0.08	5.48
hr2bis	1.58	1.14	0.06	7.99	0.42	338	0.91	485	2.3	7.5	0.12	0.32	0.39	0.02	0.09	5.33
tr5	1.43	17.9	0.07	1.02	6.72	182	0.84	256	1.28	2.73	0.1	4.7	10.7	0.15	0.07	4.9
tr4	1.35	22	0.33	8.59	1.8	317	1.44	428	1.29	2.65	0.1	3.19	6.98	0.26	0.08	5.28
hr1	0.68	1.97	0.01	5.56	3.45	387	1.82	463	1.98	6.8	0.1	0.31	0.43	0.03	0.09	5.33
hr6bis	1.3	0.98	0.04	6.52	0.61	359	1.3	523	1.93	7.5	0.09	0.24	0.2	0.02	0.06	3.39
hr8	1.1	0.05	0.03	7.34	1.76	384	1.52	404	1.52	7.81	0.09	0.28	0.08	0.02	0.07	4.6
hr13	1.08	0.05	0	7.82	0.59	394	1.13	263	1.95	8.21	0.1	0.32	0	0	0.09	5.36
Odiel1	0.59	4.4	0	2.61	1.65	126	0.85	60.2	0.27	3.21	0.04	2.84	7.85	0	0.01	0.95
tr1	3.92	4.39	0.11	4.1	7.39	247	1.12	273	1.93	5.8	0.11	0.66	1.59	0.07	0.09	6.31
hr5	1.53	0.05	0.02	7.69	9.54	405	1.39	428	1.51	7.6	0.11	0.32	0.08	0.02	0.08	5.48
hr7bis	2.15	0.05	0.03	8.3	0.52	355	1.27	267	1.93	7.6	0.09	0.31	0.19	0.03	0.07	4.81
hr11bis	2.87	0.57	0.04	8.61	0.87	410	1.41	401	2.17	8	0.1	0.29	0.05	0.01	0.08	4.71
hr7	1.64	2.07	0	9.12	0.76	414	1.45	150	1.44	7.45	0.1	0.32	0.26	0.03	0.08	4.78
San Juan	1.23	0	0.1	5.67	2.84	119	0.21	32.7	0.07	2.4	0.07	10.9	31.4	0.06	0.01	0.18
tr2	1.23	9.95	0.17	8.75	8.56	208	0.74	434	1.72	3.66	0.09	1.43	2.62	0.1	0.08	5.21
hr6	1.88	3.82	0	10.3	0.52	366	1.37	314	1.42	7.53	0.1	0.33	0.36	0.03	0.08	4.79
hr2	1.22	3.77	0.05	7.99	0.42	338	0.91	485	2.3	7.35	0.13	0.38	0.61	0.04	0.09	6.01
hr8bis	1.93	0.05	0.04	10.6	1.53	406	1.38	421	1.77	7.6	0.09	0.28	0.08	0.02	0.08	4.91
hr9	1.37	1.84	0	6.42	12.9	371	1.61	366	1.95	7.8	0.1	0.31	0.25	0.03	0.08	5.22
Odiel7	1.83	7.21	0	7.63	10.8	260	0.84	228	0.96	6.5	0.11	0.31	0.35	0.03	0.1	6.04
Odiel3	0.94	6.51	0	3.69	1.98	126	0.42	156	0.52	3.6	0.06	2.31	5.93	0	0.03	2.21
Odiel6	2.51	3.88	0	6.75	6.76	214	0.65	284	1.22	5	0.09	0.98	2.51	0	0.06	4.37
Niebla	0.69	10.7	0.08	3.8	2.46	132	0.07	52	0.12	2.78	0.05	10.4	22.1	0	0	0.18
Gibraleón	0.47	0.05	0.13	4.76	4.04	131	0.17	73.9	0.18	3.35	0.02	2.04	5.73	0	0	0.04
hr10bis	2.57	2.24	0	9.51	7.54	410	1.87	425	2.02	7.9	0.1	0.3	0.03	0.01	0.08	5.22
tr6	2.28	20.7	0.1	6.73	8.34	253	1.14	401	1.26	2.8	0.1	4.51	10.4	0.14	0.08	5.08
hr12bis	2.9	0.86	0.04	6.79	11.2	405	1.48	485	1.86	8	0.09	0.26	0.17	0.01	0.07	4.41
hr10	1.18	0.05	0	9.15	6.34	136	0.51	198	1.47	7.77	0.1	0.35	0.25	0.03	0.08	5.48
hr11	1.55	0.05	0.01	9.27	1.12	422	1.46	459	1.72	7.98	0.1	0.32	0.49	0.02	0.08	5.16

from 0 to 1, each value representing the absolute non-membership or membership to the set, respectively.

The membership grade may be represented by a function (von Altrock 1995). Figure 2 shows an example of membership function. Note, e.g., that a pH of 4.5 and another one of 4.6 are evaluated differently, but only by means of a slight change, and not by means of a threshold.

Fuzzy sets are a generalization of traditional sets. The $\mu_{\text{VLpH}}\ (X) = 0$ and $\mu_{\text{VLpH}}\ (X) = 1$ cases which would correspond to conventional sets, are just special cases of fuzzy sets. The use of fuzzy sets defined by means of membership functions in logic expressions is called *fuzzy logic*. In these expressions, the membership grade of a set is the degree of certainty of the sentence. For example, in Fig. 2, the expression "the pH of the sample is very low", would be true in a grade of 0.67 for a sample with pH 4.5.

The geometric form of membership functions is totally arbitrary, but in general, simple geometry and
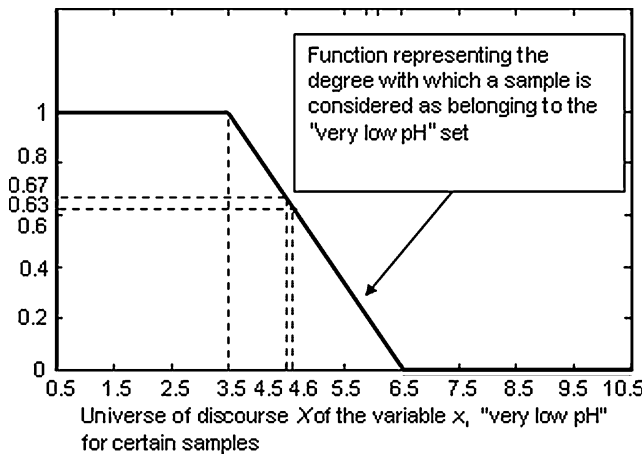
Fig. 2 Example of membership function for the fuzzy set "very low pH"

known equation functions, such as trapeziums, triangles or sigmoids, are used.

Once all variables involved in the problem are coded to the qualitative domain by means of membership functions, it is possible to write a set of rules representing the relation between input and output variables. These rules present the format *if-then*, and are made up of an antecedent and a consequent; the fulfilment of the antecedent implies the conclusion. From the standpoint of knowledge representation, a fuzzy rule *if-then* is a structure for representing imprecise knowledge. The main characteristic implied by the reasoning based on this type of rules is its ability to represent partial coincidence, which allows a fuzzy rule to provide inference even when the condition is satisfied only partially (Yen and Langari 1999). The following implications allow us to illustrate briefly these logic interferences:

If x is A then y is C (1)

If x is A and z is B then y is C (2)

The first rule has a single antecedent, i.e., of the type "if the variable x is a member of class A". However, the second rule has a compound antecedent–compound antecedents are logical combinations of single antecedents.

The process of extracting knowledge from a database is called KDD (*Knowledge Discovery in Databases*). This process is made up of several stages ranging from data preparation to achievement of results (Fallad and Uthurusamy 1996; Zaïane 1999). One of these stages is called *data mining* and can be defined as the non-trivial process of extracting implicit, a priori unknown useful information from the stored data (Holsheimer and Siebes 1994; Kruse et al. 1999).

The computer tool: Predictive Fuzzy Rules Generator (PreFuRGe) (Aroba 2003)

Classical clustering algorithms generate a partition of the population in a way that each case is assigned to a cluster. These algorithms use the so-called "rigid partition" derived from the classical sets theory: the elements of the partition matrix obtained from the data matrix can only contain values 0 or 1; with zero indicating null membership and one indicating whole membership. That is, the elements must fulfill:

(a) $0 \leq \mu_{ik} \leq 1, \quad 1 \leq i \leq c, 1 \leq k \leq n$

(b) $\sum_{i=1}^{c} \mu_{ik} = 1, \quad 1 \leq k \leq n$ (3)

(c) $0 \leq \sum_{k=1}^{n} \mu_{ik} \leq n, \quad 1 \leq i \leq C$

Fuzzy partition is a generalization of the previous one, so that it holds the same conditions and restraints for its elements, except that in this case real values between zero and one are allowed (partial membership grade). Therefore, samples may belong to more than one group, so that the selecting and clustering capacity of the samples increases. From this, we can deduce that the elements of a fuzzy partition fulfill the conditions given in (3), except that now condition (a) will be written as:

$$\mu_{ik} \in [0,1], \quad 1 \leq i \leq c, \quad 1 \leq k \leq n \quad (4)$$

The best-known general-purpose fuzzy clustering algorithm is the so-called *Fuzzy C-Means* (FCM) (Bezdeck 1981). It is based on the minimization of distances between two points (data) and the prototypes of cluster centres (*c-means*). For this purpose, the following cost function is used:

$$J(\mathbf{X}; U, V) = \sum_{i=1}^{c} \sum_{k=1}^{n} (\mu_{ik})^m \|\mathbf{x}_k - \mathbf{v}_i\|_A^2 \quad (5)$$

where U is a fuzzy partition matrix of $\mathbf{X}$, $V = [v_1, v_2, ..., v_c]$ is a vector of cluster center prototypes which must be determined and $m \in [1, \infty]$ is a weighting exponent which determines the degree of fuzziness of the resulting clusters. Finally,

$$D_{ikA}^2 = \|\mathbf{x}_k - \mathbf{v}_i\|_A^2 = (\mathbf{x}_k - \mathbf{v}_i)^T \mathbf{A}(\mathbf{x}_k - \mathbf{v}_i) \quad (6)$$

is the used norm for measuring distances (matrix $\mathbf{A}$ induces the rule to be used—provided that it is the unit

matrix, which is very frequent—, i.e., the Euclidean norm)

The described algorithm was used (Sugeno and Yasukawa 1993) to build a fuzzy model based on rules of the form:

$$R^l : \text{If } \mathbf{x} \in \mathbf{A}^l \text{ then } y \in B^l \qquad (7)$$

where $\mathbf{x} = (x_1, x_2, ..., x_n) \in \Re^n$ are input variables, $\mathbf{A} = A_1, A_2, ..., A_n$ are n fuzzy sets, $y \in \Re$ is the output variable and B is the fuzzy set for this variable.

The developed computer tool, PreFuRGe (Aroba 2003), is based on the previously described methodology (Sugeno and Yasukawa 1993) and represented by (7). This initial methodology has been adapted and improved in the following aspects:

1. It allows working with quantitative databases, with n input and m output parameters.
2. The different variables object of study can be weighted by assigning them weights for the calculation of distances between points of the space being partitioned.
3. The achieved fuzzy clusters are processed by another algorithm to obtain graphic rules trapeziums (Fig. 3).
4. An algorithm processes and solves cases of multiple projections in the input space (mounds).
5. The output provided in the original method has been improved with a graphic interface showing the graphic of the achieved rules.
6. An algorithm provides automatically the interpretation of the fuzzy graphic rules in natural language.

Figures 4 and 5 show two examples of rules generated by means of the tool PreFuRGe.

In the rule of Fig. 4, the fuzzy set assigned to each parameter is represented by a polyhedron. The parameter values are represented on the x axis of each fuzzy set, and the value of membership to a cluster on the y axis. This fuzzy rule would be interpreted as follows:
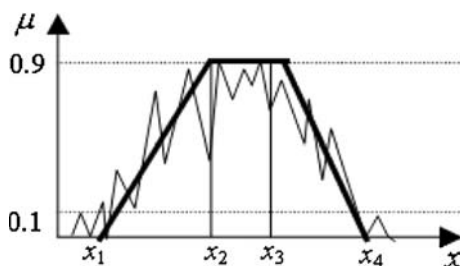
Fig. 3 Trapezium approximation of a fuzzy set

IF A1 is small and A2 is bigger or equal to average THEN S is very small.

When applying the fuzzy clustering algorithm (Aroba et al. 2001) to the generated databases, it is possible to obtain multiple projections in the input parameters (mountain). In the fuzzy rule of Fig. 5, a multiple projection (mountain) is represented in the input parameter A1. In this case, we observe how the parameter A1 can take different types of values for a certain kind of output. This fuzzy rule can be interpreted as follows:

IF A1 is *small* or *big* and A2 is *average* THEN S is *very small*.

Recently, one of the authors of this article has investigated the stability of fuzzy logic control systems, as well as their advanced industrials applications (Andújar et al. 2004; Andújar and Bravo 2005; Andújar and Barragán 2005).

Results

Table 1 shows the values taken by the concentrations of the studied variables and the geographical reference of the sample (Fig. 1).

Charts in Fig. 6 show, by means of fuzzy rules, pH behaviour as opposed to the rest of variables as a whole. The universe of discourse for each variable has as end values those maximums and minimums given in Table 1. So, in this case, 2 would be a very low pH and 8.21, a very high one. Thus, an average pH value is around 5. In order not to excessively divide the universe of discourse of pH, and perhaps make interpretation more difficult, only four intervals have been established for the consequent. Nevertheless, this consequent may project any (appropriate) value in the variables that make up the antecedent. The meaning of the charts is the following: when pH has values within the low to very low ranges (from around 3.5 to 2), average to low (from slightly over 5 to around 3.5), average to high (from over 5 to around 7) or high to very high (from over 7 to 8.21), the remaining variables considered as a whole take the values shown in Fig. 6.

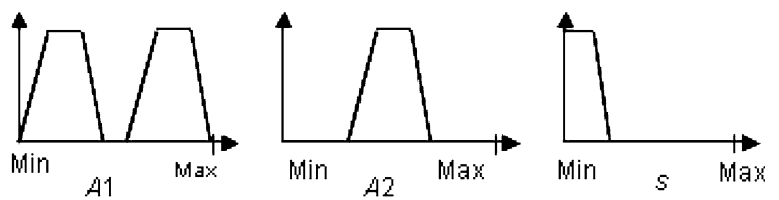From the individual examination of the charts, the following can be deduced:

High to very high pH values (Fig. 6a) are compatible with:

- SiO_2 and As, low to very low.
- PO_4, average-low to very low.
- SO_4, average to low.
- K and Sr, average to high.
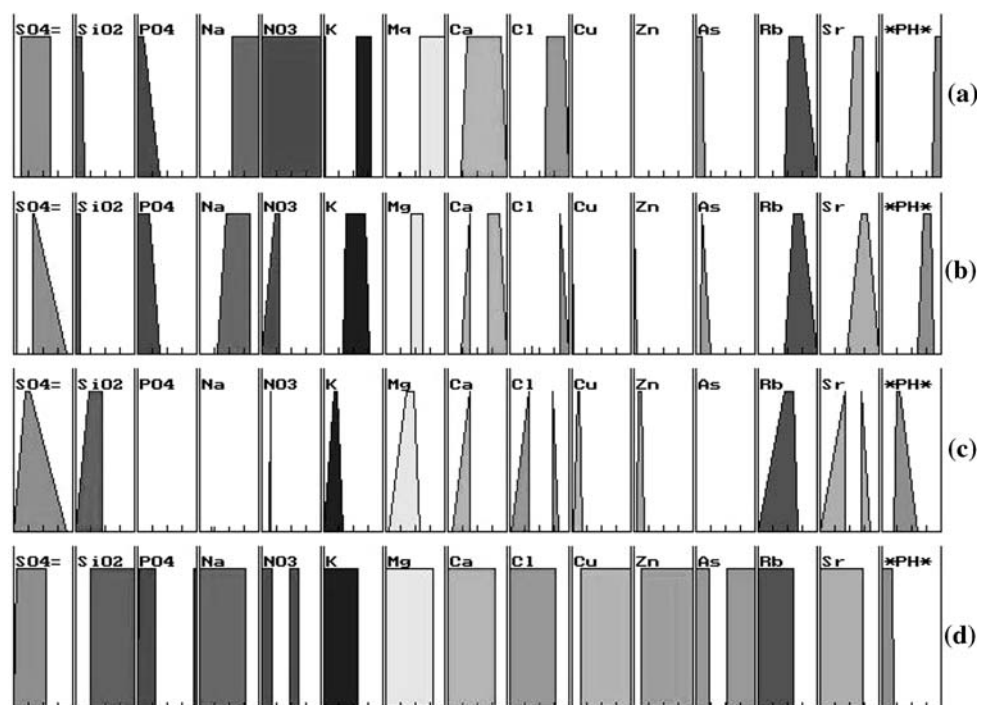- Ca, low to high, with a few points very high.

Fig. 4 Example of Fuzzy rule
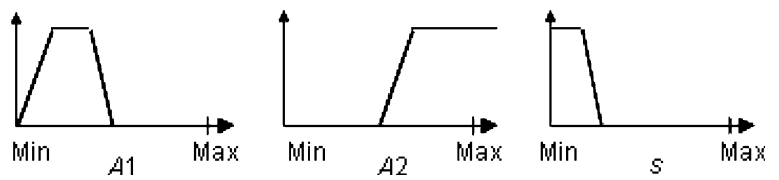
Fig. 5 Example of Fuzzy rule with multiple projection

Fig. 6 Graphic Fuzzy rules for qualitative behaviour of pH as opposed to the rest of variables

- Na, Cl and Mg, average to very high.
- Rb, average to high, with a few points very high.
- Cu and Zn, do not co-occur with this pH value.
- NO$_3$, may appear in any concentration (it does not provide significant information in that rule).

Average to high pH values (Fig. 6b) are compatible with:

- SiO$_2$, very low.
- NO$_3$, As, low.
- PO$_4$, low to very low.
- K, low to high, with a strong presence of average values.

- Mg, average.
- SO$_4$, low to high, the presence of points towards high values decreasing progressively.
- Na, low to high, with greater presence of points between average and high.
- Ca, average to low or high to very high, without average values and with a few points at end values.
- Rb and Sr, average to very high, with a strong presence of points at high values.
- Cl, high to very high, the presence of points towards very high values decreasing progressively.
- Cu and Zn, very low.

Average to low pH values (Fig. 6c) are compatible with:

- NO₃, low and few points highly concentrated at one value.
- Cu, Zn and K, low to very low.
- SiO₂, average to very low, the number of points decreasing towards the end value.
- Ca, low, the number of points decreasing towards the lower end value.
- Cl, low to very low or high, without average values and with few points at end values.
- Sr, average to very low or high, without average values and with few points at end values.
- Mg and Rb, average to very low values, with strong presence of points around average value and evident decrease towards the lower end.
- SO₄, very low to high, decreasing between end values.
- PO₄, As and Na, do not co-occur with this pH value.

Low to very low pH values (Fig. 6d) are compatible with:

- PO₄, low to very low or very high.
- NO₃, low to very low or slightly over average.
- SO₄ and K, average to very low.
- Na, Mg, Ca, Cl, Sr, high to very low.
- Rb, average to very high.
- Cu and Zn, except for very low values, any concentration is possible.
- SiO₂, low to very high.

As, any concentration except from average to low.

The joint examination of the charts allows us to observe the following behaviour patterns among variables:

As pH decreases, the range of concentrations of almost all elements increases, except for PO₄ y SO₄, which deserve a different reasoning, and NO₃, which has the opposite behaviour.

As pH increases, Cl, Na, K and Mg concentrations increase. Rb and Sr, showing similar behaviours with one another, must be included in this same group. Also Mg and Na present analogous responses with one another for end pH values.

Ca wide concentration ranges appear for very high or very low pH.

PO₄ concentrations differ just slightly from those of pH, except for average to low pH values, with which they are incompatible. In this same pH range, neither Na nor As are observed. It must be stressed that for this same pH range, there are very few points with highly defined low concentration of NO₃.

For non-end pH values, no big concentration ranges of any element exist.

Discussion

The clustering of responses from the different studied variables according to their capacity of co-occurrence with the remaining ones for concrete pH values is evident. The phenomenon is conditioned by the conjunction of different processes that control the chemistry of the estuary water. These processes have been widely described by different authors such as Borrego et al. (2002), Grande et al. (2000, 2003a, b, c), and Sáinz et al. (2002, 2003) and can be summarized as follows:

Fluvial contribution with acidic waters loaded with sulphates and dissolved metals from the spring area of the rivers affected by AMD processes as they cross the Iberian Pyrite Belt, where sulphate deposits have been operated for over 4,500 years BP. The result has been the formation of a fluvial system that is unique in the world. Sáinz et al. (2002) estimate maximum daily contribution as 1,481 kg of As, 470 kg of Pb and 170 kg of Cd.

Industrial effluents in the zone of the estuary coming from the chemical industrial complex with fertilizer factories and phosphogypsum deposits, cooper foundries, paper mills and others (Grande et al. 2003a).

Marine input as a result of the sea balance that takes remarkable values in this sector. According to Borrego (1992) and Grande et al. (2003c), the volume of freshwater inflow for the Tinto and Odiel rivers to the inner zone of the estuary reflects a significant seasonal and year-to-year variation. Therefore, from 1960 to 1996, the average monthly inflow of river water was 49.8 Hm³.

Taking into account that the data scanned here by fuzzy logic and data mining have been studied by applying different multivariate statistical techniques, such as factorial and cluster analysis (Grande et al. 2000, 2003c; Borrego et al. 2002), we will now compare the results in both cases, as a way of contrast for the validation of the method in question.

In principle, our results do not come into conflict with those of the referred authors. Nevertheless, in certain cases some differences can be established, which cannot be determined exactly with classical statistical tools. This is the case of stability fields of metals dissolved in water: while Pearson's (1901) correlation coefficient 'r' establishes the proximity between two variables, and therefore, in our case, informs us of the proximity of one variable with respect to another (i.e., whether one or another variable increases

or decreases depending on the other without specifying in which section of the variable path a greater or lesser correlation occurs), the computer tool PreFuRGe offers an overall view of the system's qualitative behaviour as contrasted with pH variation stimuli, showing the response of the set of variables, so that we obtain acidity range distribution of the co-occurrence regions for the studied parameters, also indicating the restriction degree of the variable with respect to the stimulus.

Similar responses of variables allow us to call common provenance phenomena as Grande et al. (2000, 2003c), Borrego et al. (2002), and Sáinz et al. (2002) propose. Thus, those variables that increase, in relation to the four considered pH sections (low to very low, average to low, average to high, and high to very high), towards the highest pH value, such as Cl, Na, K, Ca, Mg, Rb, Sr, inform us that the likelihood of their appearing dissolved in the estuarine environment increases with pH, i.e., in our case with tidal influence, whereby they can be considered variables of marine provenance.

The same approach is applied to the variables whose concentration ranges increase in the estuary as pH becomes lower, such as Cu, Zn, As and SiO_2. Here, we have solutes whose presence is favoured by acidity, as they are compatible only in highly acidic waters. Thus, they are expected to have a fluvial provenance associated to AMD or industrial processes.

PO_4 and SO_4 deserve a different treatment:

Since the PreFuRGe tool has been used for treating the data regardless of its geographical location, we understand that the triple provenance of phosphates—fluvial as a result of sulphide oxidation, marine and industrial associated to effluents from phosphogypsum deposits (Grande et al. 2003c)—justifies the limited definition of this salt's response relative to pH (Fig. 6). Measure points in the effluents from the phosphogypsum deposits that have extremely high sulphate concentration values (see Table 1) are disguised by the remaining data with much lower concentrations.

Phosphates show behaviour similar to that of As; in both cases, they present low to very low concentrations for average to very high pH values, and also in both cases they 'disappear' for average to low pH values. What is most significant of this behaviour is the fact that in low to very low pH sections we may find low to very low or very high concentrations, in the case of phosphate, and average to very high in the case of As. However, in these sections, none of the elements present average concentrations. This phenomenon, which has not been described before with statistical treatment, is clearly evident when scanning fuzzy rules. The explanation must be sought in the following phe-

nomenon: in the environment being described, the first input is acidic and fresh, whereby all metals will be dissolved in the water. As a result of the tidal clash, a periodic scavenging process occurs, originating in turn remarkable variations of the immobilized As. To this phenomenon widely described by Borrego (1992), the double provenance—on the one hand fluvial associated to AMD processes and on the other hand industrial associated to effluents from the phosphgypsum deposits, that can regulate the presence of dissolved As must be added. In this context, Grande et al. (2003c), after the correlate analysis of this same data, report that the examination of the correlation matrix shows the existence of high Pearson r values between pairs of variables, which can be interpreted as the result of a common provenance, since sampling has been carried out in a sector subject to periodic tidal influence and to fluvial input. In the referred work, after subjecting data to factorial analysis, variables are gathered round two factors (fluvial influence and tidal influence), whereby they result in Pearson's affinity clusters compatible with the fuzzy clustering analysis that we propose.

Acting again over the same data, Grande et al. (2003c), by means of classical cluster analysis, conclude about the existence of two groups of variables in the estuary. A first group with pH, SO_4 and typically marine indicators such as Na, K, Ca, Cl, Li, Rb and Sr, and a second group with PO_4, As, Zn, Cu, Mg, SiO_2, NO_3 as fluvial indicators. If we observe the behaviour of the charts provided in this work (Fig. 6), we can reach identical conclusions, but with the possibility of discriminating the functioning of the studied group of variables as a whole depending on the predefined pH range of our choice. This allows us also to know interdependency reasons and therefore the system's response to stimuli that modify concentrations of the dissolved elements. At the same time, the overall view of the rules relative to each pH range allows observation of similar behaviours (as is the case of Cu/Zn), (Rb/Sr), (PO_4/As), (Ca/Mg), that suggest phenomena of common provenance.

Conclusions

The aim of our work has been to present qualitative models, which allow us, in an easy, intuitive and at-a-glance way, and without the need of calculations or data processing, to have a clear idea of the physical processes that generate the data clusters shown by this computer tool.

The developed methodology allows us to establish cause–effect relationships, since the cause (fuzzy par-

tition pH) originates the effect (element concentrations and compounds), represented by the fuzzy clusters at the income. Of course, interpretation must be qualitative, i.e., as a human being would reason, so that no numeric values but predicates are used: high, low, medium, very high, very low, etc.

The application of fuzzy logic and data mining for characterizing hydrochemical processes in the same sector and from the same data confirm and enrich functioning models previously proposed for this sector by means of multivariate analysis.

While traditional tools of classical statistics widely used in this context are useful for defining proximity reasons among variables on the basis of Pearson's relations, the use of fuzzy-logic and data mining tools provides, in addition to easy handling of large data and easy understanding of charts, an improved definition of the variations originated by external stimuli on the whole set of variables.

The PreFuRGe tools allow high versatility, given that the configuration of the antecedent and the consequent of the fuzzy rule can be changed at will. In this work, pH has been considered as the only consequent, and the rest of elements as a whole have been considered the antecedent. However, other groupings could have also been scanned.

References

Andújar JM, Barragán AJ (2005) A methodology to design nonlinear fuzzy control systems. Fuzzy Sets Syst 154(2):157–181

Andújar JM, Bravo JM (2005) Multivariable fuzzy control applied to the physical-chemical treatment facility of a Cellulose factory. Fuzzy Sets Syst 150(3):475–492

Andújar JM, Bravo JM, Peregrín A (2004) Stability analysis and synthesis of multivariable fuzzy systems using interval arithmetic. Fuzzy Sets Syst 148(3):337–353

Aroba J, Ramos I, Riquelme JC (2001) Application of machine learning techniques to software project management. ICEIS 2001 (III International Conference on Enterprise Information Systems), Setubal (Portugal), p 433

Azcue JM (1999) Environmental impacts of mining activities. Springer, Heidelberg

Bezdek JC (1981) Pattern recognition with fuzzy objective function algorithm. Plenum, New York

Borrego J, Morales JA, Pendón JG (1993) Holocene filling of an Estuarine Lagoon along the Mesotidal Coast of Huelva: the Piedras River Mouth, southwestern Spain. J Coast Res 9:242–254

Borrego J, Morales JA, de la Torre ML, Grande JA (2002) Geochemical characteristics of heavy metal pollution in surface sediments of the Tinto and Odiel river estuary (Southwestern Spain). Environ Geol 41:785–796

Braungardt CB, Achterberg EP, Nimmo M (1998) Behavior of dissolved trace metals in the Rio Tinto/Rio Odiel Estuarine System. In: Morales JA, Borrego J (eds) European land–ocean interaction studies, Second Annual Scientific Conference, p 51 (abstracts)

Commonwealth of Pennsylvania (1994) Water quality assessment in Western Pennsylvania Watershed

Cruzado A, García HE, Velásquez ZR, Grimaldo NS, Bahamon N (1998) The Ria de Huelva (SW Spain). Hydrography and general characteristics. In: Morales JA, Borrego J (eds) European land–ocean interaction studies. Second Annual Scientific Conference, p 83 (abstracts)

Davis RA Jr, Welty AT, Borrego J, Morales JA, Pendón JG, Ryan JG (2000). Rio Tinto estuary (Spain): 5,000 years of pollution. Environ Geol 39:1107–1116

Dogan PA (1999) Characterization of mine waste for prediction of acid mine drainage. In: Azcue JM (ed) Environmental impacts of mining activities. Springer, Heidelberg, pp 19–38

Elbaz-Poulichet F, Dupuy C (1999) Behaviour of rare earth elements at the freshwater-seawater interface of two acid mine rivers: the Tinto and Odiel (Andalucía, Spain). Appl Geochem 14(8):1063–1072

Elbaz-Poulichet F, Leblanc M (1996) Transfer de métaux d'une province minière á l'océan par des fleuves acides (Rio Tinto, Espagne). CR Acd Sci Paris 322:1047–1052

Elbaz-Poulichet F, Braungardt CB, Achterberg EP, Morley NH, Cossa D (1999a) A synthesis of the results of TOROS and CANIGO projects on metal contamination in the Tinto-Odiel rivers (Southern Spain) and the Gulf of Cadiz. In: Pacyma JM, Kremer H, Pirrone N, Barthel K (eds) Socioeconomic aspects of fluxes of chemicals into marine environment, pp 1001–1109

Elbaz-Poulichet F, Morley NH, Cruzado A, Velásquez Z, Achterberg EP, Braungardt CB (1999b) Trace metal and nutrient distribution in an extremely low pH (2.5) river-estuarine system, the Ria of Huelva (South-West Spain). Sci Total Environ 227:73–83

Elbaz-Poulichet F, Dupuy C, Cruzado A, Velásquez Z, Achterberg E, Braungardt C (2000) Influence of sorption processes by iron oxides and algae fixation on arsenic and phosphate cycle in an acidic estuary (Tinto river, Spain). Water Resour 34(12):3222–3230

EMCBC (1996) The perpetual pollution machine. Acide mine drainage. B.C. Mining Control, Canada, pp 1–6

Fallad UM, Uthurusamy R (1996) Data mining y KDD. Commun ACM 39(11)

Feasby DG, Tremblay GA, Weatherell CJ (1997). A decade of technology improvement to the challenge of acid mine drainage- a Canadian perspective. In: Fourth international conference on acid rock drainage, vol 1, pp i–ix. Vancouver, Canada. Proceedings

Förstner U, Wittmann GTW (1983). Metal pollution in the aquatic environment. Springer, Heidelberg

Grande JA, Borrego J, Morales JA (2000) A study of heavy metal pollution in the Tinto-Odiel estuary in southwestern Spain using factor analysis. Environ Geol 39(10):1095–1101

Grande JA, Borrego J, de la Torre ML, Sáinz A (2003a) Application of cluster analysis to the geochemistry zonation of the estuary waters in the Tinto and Odiel rivers Huelva, Spain) Environ Geochem Hlth 25:233–246

Grande JA, de la Torre ML, Sáinz A (2003b) Odiel River: acid mine drainage and current characterization by means of univariate analysis. Environ Int 29:51–59

Grande JA, Borrego J, Morales JA, de la Torre ML (2003c) A description of how metal pollution occurs in the Tinto-Odiel rias (Huelva- Spain) through the application of cluster analysis. Mar Pollut Bull 46:475–480

Grande JA, Sáinz A, Beltrán R, González F, de la Torre ML (2003d) Caracterización de procesos de drenaje ácido de mina en la Faja Pirítica Ibérica sobre un embalse para abastecimiento público. Proceedings of VIII Congreso de Ingeniería Ambiental. Bilbao. Spain, pp 443–452

Grande JA, González F, Sáinz A, de la Torre ML, Beltrán R, Sánchez D (2003e) Aporte de metales pesados a la red fluvial procedentes de la actividad minera en el SW de España, vol 2. In: Proceedings of III Congreso Argentino de Hidrogeología. Rosario. Argentina, pp 331–343

Holsheimer M, Siebes A (1994) Data mining: the search for knowledge in Databases. Report CS-R9406, CWI Amsterdam

Kruse R, Klawonn F, Höppner F (1999) Fuzzy cluster analysis: methods for classification, data analysis and image recognising. Wiley, New York

Lu R, Lo S (2002) Diagnosing reservoir water quality using self-organizing maps and fuzzy theory. Water Res 36:2265–2274

Nebel BJ, Wrigth RT (1999) Ciencias ambientales. Ecología y desarrollo sostenible. Prentice Hall, México

Nicholson RV (1994) Iron-sulfide oxidation mechanism: laboratory studies. In: Jambor JL, Blowes DW (eds) The Environmental Geochemistry of Sulfide Mine-Wastes, Mineralogical Association of Canada Short Course Handbook. Canada, vol 22, pp 163–182

Ong C, Huang J, Tzeng G (2004) Multidimensional data in multidimensional scaling using the analytic network process. Pattern Recog Let 26(6):755–767

Pearson K (1901) On lines and planes of closets fit to systems of points in space. Philos Mag, ser. 6, 559–572

Sáinz A, Grande JA, de la Torre ML, Sánchez-Rodas D (2002) Characterisation of sequential leachate discharges of mining waste rock dumps in the Tinto and Odiel rivers. J Environ Manage 64(4):345–353

Sáinz A, Grande JA, de la Torre ML (2003) Analysis of the impact of local corrective measures on the input of contaminants from the Odiel river to the ria of the Huelva (Spain). Water Air Soil Poll 144:35–389

Sáinz A, Grande JA, de la Torre ML (2004) Characterisation of heavy metal discharge into the Ria of Huelva. Environ Int 30:557–566

Simmons HB (1955) Some effects of upland discharge on estuarine Hydraulic. Proc ASCE 81:1–20

Sugeno M, Yasukawa A (1993) A Fuzzy-Logic Based approach to qualitative Modeling. IEEE Trans Fuzzy Syst 1:7–31

Travesi A, Gasco C, Pozuelo M, Palomares J, García MR, Pérez del Villar L (1997) Distribution of natural radioactivity within an estuary affected by release from phosphate industry. In: Desmet G (ed) Freshwater and estuarine radioecology, Elsevier, Amsterdam, pp 267–279

USEPA (1994) Water quality standards handbook, 2nd edn. Washington DC, U.S. Environmental Protection Agency. Office of Water. EPA-823-B-94–005

Von Altrock C (1995) Fuzzy logic and neurofuzzy applications explained, Prentice-Hall, New York

Ward M, LeBlanc J, Tipnis R (1994) N-land: a graphical tool for exploring N-dimensional data. In: Computer graphics international conference. Melbourne, Australia

Weatherell CJ, Feasby DG, Tremblay GA (1997) The mine environment natural drainage program. In: Proceedings of the PMI 97, 28th Annual Seminars and Symposium, Chicago

Yen J, Langari R (1999) Fuzzy logic: intelligence, control and information, Prentice-Hall, New York

Younger P, Banwart S, Hedin R (2002) Mine water: hydrology, pollution, remediation. Kluwer, Dordrecht, Netherlands

Zadeh LA (1965) Fuzzy sets, information and control, vol 8, pp 338–353

Zaïane OR (1999) Principles of knowledge discovery in databases. CMPUT690, Department of Computing Science, University of Alberta

Sources of unpublished materials

Aroba J (2003) Advances en la toma de decisiones en proyectos de desarrollo de software, PhD thesis, University of Sevilla, Spain

Borrego J (1992) Sedimentología del estuario del Río Odiel, Huelva, S.O. España. PhD thesis, University of Sevilla, Spain

Sáinz A (1999) Estudio de la contaminación química de origen minero en el río Odiel. PhD thesis. University of Huelva, Spain