A Context-Dependent Knowledge Model for Evaluation of Regional Environment

S. Kawano [a,1], V.N. Huynh [a,*], M. Ryoke [b], Y. Nakamori [a]

[a] *School of Knowledge Science*
Japan Advanced Institute of Science and Technology
Tatsunokuchi, Ishikawa, 923-1292, JAPAN

[b] *Graduate School of Business Sciences*
University of Tsukuba
Otsuka 3-29-1, Bunkyo, Tokyo 112-0012, Japan

Abstract

In this paper we develop a rule-based model for evaluation of regional environment based on both hard and soft data, where by hard data we mean the statistical measurements while by soft data we mean subjective appreciation of human beings of environmental issues. As people's feeling strongly depends on the social and economical characteristics of administrative regions where they live, we firstly use the hard data concerning these characteristics to do clustering in order to obtain clusters corresponding to regions with the homogeneous social and economical characteristics relatively. We then use the soft data, with helping of data mining techniques, to develop rule-based models which show association between evaluated items of residents in the clusters. Finally, a relationship between hard data and soft data through an integrated model will be explored. It is shown that the soft data is rather reliable and we should integrate subjective knowledge learnt from soft data into modelling of environmental issues.

Key words: Environmental modelling, fuzzy clustering, data mining, optimal rule, context-dependent knowledge model

* Corresponding author. Fax: 81-761-51-1149
Email address: huynh@jaist.ac.jp (V.N. Huynh).
[1] Current address: NORD Institute for Society and Environment, Forest Tower 7F, 1-9-10, Kyobashi, Chuo-ku, Tokyo, 104-0031, Japan.

1 Introduction

Traditionally, environmental models use mathematical equations to represent the interconnections in environmental systems (Ford, 1999). These models which correspond to time change and have dynamical features are built mainly based on statistical data collected from many sources. Namely, when considering the problem of building environmental models we usually use the numerical measurements which will be called hard data in this paper. On the other hand, environmental models do concern with social and economical elements. This makes them also become human-centered systems with more complexities and difficulties to deal with. Under such an observation, some approaches with the emphasis on intellectual intuition of human beings to complex systems analysis have been proposed and, simultaneously applied to environmental problems (Kainuma et al., 1990; Nakamori and Sawaragi, 1997, 2000). Recently, Nakamori (2000) has proposed a knowledge science based systems methodology to deal with environmental issues for regions. Within this framework, all kinds of knowledge should be incorporated into system models. Particularly, the knowledge discovered from soft data, where by soft data we mean the feeling of human beings of environmental issues, should be combined with knowledge learnt from hard data.

In the framework of a research program, we have collected two categories of data concerning environmental factors, here we confined ourselves the consideration to only Kaga area, a south one of Ishikawa prefecture, Japan. The first one called hard data consists of statistical data which are kept as archives at Ishikawa Statistical Information Division (1999). The second one called soft data that reflects local inhabitants' assessments on environmental factors by a questionnaire survey (Section 2).

This paper aims at developing a rule-based model for evaluation of regional environment based on both hard and soft data. Firstly, based on the observation that inhabitants' feeling strongly depends on the social and economical characteristics of administrative regions where they live, we have used hard data concerning these characteristics to do clustering so that each cluster consists of residential areas having homogeneous social and economical characteristics relatively. Then, the soft data are used to extract rules that show the association between partial evaluations and total evaluation of inhabitants in each cluster on an environmental factor such as, in our case, the quality of water sources. Previously, we have borrowed the idea from Apriori algorithm (Agrawal and Srikant, 1994) in mining association rules for extracting "semi-linguistic" rules from soft data for each cluster (Kawano et al., 2001). However, to obtain rules with a support constraint, we transformed original soft data into another form using the ordinal scale defined on the data. Particularly, original soft data in the one-to-five scale are transformed into the

3-scale. This causes a loss of original information and, consequently, attenuates the interpretability of the obtained rules. We would like to emphasize that, by mining maximal rules [2] according to the partial order defined in terms of support and confidence values of rules in this paper, we support an interactive phase in which the modeller can browse the optimal rules according to any of several goodness metrics known in the literature. This is useful as the subjective appreciation of human beings on environmental issues does not have a high consolidation due to vagueness (imprecision and conflict) of linguistic knowledge, it is impossible to expect for getting a highly significant model that correctly characterizes environmental issues for each cluster.

Finally, after analyzing and selecting the soft data based representative model for each cluster, we develop an integrated model for evaluation of regional environment, which permits us to incorporate linguistic knowledge learnt from soft data into the model with a closely connectedness to hard data.

The paper is organized as follows. In the next section, we first introduce a questionnaire survey that has been used to collect the soft data in our research, and then, thanks to a clustering technique for statistical data, we identify clusters so that each cluster consists of administrative regions having the homogeneous social and economical characteristics relatively. In Section 3, after briefly recalling some preliminary notions on optimized rule mining, we develop rule-based models which show the association between evaluated items of residents in the clusters. The relationship between hard data and soft data through an integrated model will be explored in Section 4. Finally, some concluding remarks will be given in the Section 5.

2 Soft Data and Clustering Regions of Ishikawa Prefecture

2.1 Questionnaire survey for collecting soft data

As mentioned in the Introduction, environmental models are related to mainly social problems, and moreover, effects of human activities on the environment have been increasing. The environmental issues have been influencing on daily life of residents, and people also have an appreciation of and behavior on these issues. It should be very useful if local authorities can capture subjective appreciation of residents on the environment so that they can have suitable policies for adjusting conception of residents on the environmental issues and/or improving in environmental quality. This motivates us to incorporate subjective appreciation of residents into our studies of regional environment.

[2] called optimal rules in Data mining, see, e.g. Bayardo and Agrawal (1999)

The main problem now is how to represent the subjective appreciation of residents, that we call soft data, so that the data correctly reflect aspects of environment to some extent. In our case, we have consulted ideas from residents in an area near Kahokugata, which is a highly polluted lake. On the base of opinions collected through a preliminary survey, we have designed a questionnaire survey on environmental problems for residents in Kaga area, a south area in Ishikawa prefecture where our university is located. We sent questionnaires to 3,000 people in Kaga area, and among them 900 people sent us their answers.

The survey consists of questions about water, air, wastes and environmental policies. The following are questions which is related to water quality used in our survey. Where words that are italicized will be used as abbreviations of questions in the following sections.

[a] Questions about waterside which is the nearest to each resident's place.

(1) Are there any *creatures* in the waterside?
(2) Can you *play* in the water (ex. swimming, fishing, boating, etc.)?
(3) Can you eat *fish* which you catch in this water?
(4) Do you think the water is *brown*?
(5) Can you do *barbecue* or *camp* in the waterside?
(6) Do you think there are pollutant *sources* near the water?
(7) Are there any *plants* in the waterside?
(8) How do you evaluate the water quality now?

Answers are given in the one-to-five scale, where, for example, the meaning of values 1∼5 corresponding to answers for the questions [a-1]–[a-8] is given in the Table 1. Furthermore, we have used item [a-8] as *total evaluation* and [a-1]–[a-7] as *partial evaluations*.

2.2 Clustering regions

Kaga area where we carried out a questionnaire survey is mixed some regions which have different residential environment. Then, we divide towns, cities and villages in Ishikawa prefecture and construct environmental evaluation models for every region. To be concrete, we do clustering regions by using hard data which show social and geographical characters, and extract rules described with soft data for each cluster. The reason why we do this work is that we think residents' evaluations for environment are similar within regions where social and geographical characters are similar. If we can relate the property of regions with soft data, it is possible to guess the environmental condition in the regions by residents' evaluation. It suggests that soft data can be applied to environmental evaluation as data which complement hard data.

As is well-known, the objective of clustering is to divide a set of data objects into clusters such that objects within the same cluster have a high degree of similarity, whilst objects belonging to different clusters have a high degree of dissimilarity. Clustering techniques have been applied effectively in pattern recognition, modelling among others. The hard clustering is described by a conventional crisp membership function. This function assigns each object to one and only one of the clusters, with a degree of membership equals to one. However, the boundaries between the clusters are not often well defined and this description doesn't reflect the reality. Fuzzy clustering is meant to deal with the problem of (not well-defined) vague boundaries between clusters where the requirement of a crisp partition of the data is replaced by a weaker requirement of a fuzzy partition. Generally, the clustering techniques can be divided into hierarchical and nonhierarchical or partitioning methods. In our research, two algorithms in fuzzy clustering, namely the algorithm developed in Nakamori and Ryoke (1994) that is based on Ward's method (Ward, 1963) and the fuzzy c-means algorithm (Bezdek, 1981), have been used as representatives of these methods. A final selection of clustering results will be made by modellers.

The hard data using for cluster analysis consist of statistical data which are collected by Ishikawa Statistical Information Division (1999). First we calculate the correlation matrix and delete one of the two attributes whose correlation coefficient is greater than 0.8. Finally we have selected 13 attributes by this way. Among them are the rate of diffusion of sewerage, the amount of burnable and recyclable wastes, the proportion of forested land, the proportion of field and rice field, the number of cars, the road length and the work force of primary industry, and so on.

The result by the Ward method is selected because the geological property is considered well (Figure 1). In this result, regions which belong to the same cluster are located near by. It was difficult to find some meanings or knowledge by other clustering results. A suitable result should be selected by a modeler's subjectivity finally. The clustering result consists of six clusters corresponding to regions in Ishikawa prefecture as depicted in Figure 1. Figure 2 shows the membership functions for population density and diffusion rate of sewerage for instance.

3 Rule-Based Models for Soft Data

Before extracting linguistic knowledge in the form of a rule-based model from soft data, and for the sake of self-contained presentation, in the following subsection we would like to recall briefly some necessary notions in the optimized rule mining from data. More details could be referred to (Bayardo and

Agrawal, 1999; Bayardo et al., 2000) and the references therein.

3.1 Optimized rule mining

During the last decades, knowledge discovery and data mining emerged as a rapidly growing interdisciplinary field which merges together databases, statistics, machine learning and related areas in order to extract useful knowledge from data. Recently, due to the large quantity of data related to environmental issues, data mining techniques have been also used to learn transparent understandable rules from data in environmental studies (e.g. Brian, 1998; Nakamori and Ryoke, 1994). These rules summarize the data and provide insights into underlying trends and system behavior. This gives us the capability of querying the knowledge base regarding implicitly stored information of specific interest without referring to the original data.

Finding patterns in databases is the fundamental operation behind several common data-mining tasks including association rule mining. Recently, there are numerous proposals for mining rules from data. Among them, optimized rule mining aims at identifying only the most interesting, or *optimal*, rules according to some interestingness metric. It is difficult to come up with a single metric that quantifies the *interestingness* or *goodness* of a rule, and, consequently, several different metrics have been proposed and used. Among them are *confidence, support, gain, chi-squared value, entropy gain, gini, laplace, lift,* and *conviction.* Bayardo and Agrawal (1999) defined a partial order on rules in terms of both rule support and confidence. It has been shown that the set of rules that are maximal according to this partial order includes all rules that are optimal according to any of the above metrics (Bayardo and Agrawal, 1999). The problem of mining optimized rules is stated as follows.

The input to the problem of mining optimized rules is a quintuple

$$\langle U, D, \leq, C, N \rangle,$$

where

- U is a finite set of Boolean predicates called *conditions*, which are applied on elements in D,
- D called a data-set is a finite set of records,
- $\leq$ is an ordered relation on rules, where a rule is an expression of the form $A \rightarrow C$, for $A \subseteq U$, A is called the rule antecedent and C the rule consequent,
- C is a condition specifying the rule consequent,
- N is a set of constraints on rules, where each constraint is understood as a Boolean predicate applied on rules such as the minimum support constraint.

Now the problem is to find a rule $r = A \to C$ such that r satisfies the input constraints and there exists no set $A' \subseteq U$ such that the rule $r' = A' \to C$ satisfies the input constraints and $r < r'$. Notice that when the subset $A \subseteq U$ occurs in a rule $A \to C$, it means a conjunction of all conditions in A.

We now define the support and confidence values of rules. The *support* of a condition A, denoted by $\sup(A)$, is equal to the number of records in the data set for which A evaluates to true. The support of a rule $A \to C$, denoted similarly as $\sup(A \to C)$, is equal to the number of records in the data set for which both A and C evaluate to true. The *confidence* of a rule $A \to C$, denoted as $\mathrm{conf}(A \to C)$, is then defined as follows

$$\mathrm{conf}(A \to C) = \frac{\sup(A \to C)}{\sup(A)}.$$

Based on both support and confidence values, a partial order on rules, denoted as $\leq_{sc}$, is defined as follows. Given rules r_1 and r_2, $r_1 <_{sc} r_2$ if and only if:

- $\sup(r_1) \leq \sup(r_2)$ and $\mathrm{conf}(r_1) < \mathrm{conf}(r_2)$, or
- $\sup(r_1) < \sup(r_2)$ and $\mathrm{conf}(r_1) \leq \mathrm{conf}(r_2)$.

In addition, $r_1 =_{sc} r_2$ if and only if $\sup(r_1) = \sup(r_2)$ and $\mathrm{conf}(r_1) = \mathrm{conf}(r_2)$.

In the following by *SC-optimal rules* we mean the rules that are maximal according to the partial order $\leq_{sc}$ defined above. In the next subsection, we use the Dense-Miner algorithm developed in (Bayardo et al., 2000) to identify SC-optimal rules from the soft data.

3.2 Linguistic knowledge learned from soft data

In this subsection, we apply the technique proposed by Bayardo and Agrawal (1999); Bayardo et al. (2000) to obtain the most interesting rules from the soft data, which expresses the association between partial evaluations and total evaluation of residents on environmental factors in each cluster. To this end, each question in [a-1]–[a-7] is treated as an attribute named by the italic word and the question [a-8] is designed as the class column, where, for the sake of simplicity, we only use the soft data related to water quality for illustration. As such the answer of a resident to questionnaire is considered as a data record in the data set of all residents. Particularly, we have data sets, each for one cluster, with the following set of attributes

$$\mathbb{A} = \langle creature, play, fish, brown, camp, source, plant, total\ evaluation \rangle$$

which correspond to questions [a-1]–[a-8]. Each attribute gets a value in the range of $V = \{1, 2, 3, 4, 5\}$.

We now address the specific problem of mining optimal rules under the partial order $\leq_{sc}$ within the soft data, which are obviously treated as categorically valued data due to the nature of this kind of data. Where, each *condition* in U is simply a test of whether the given input record contains a particular attribute/value pair, excluding values from the designated class column, i.e. *total evaluation*. Values from the class column *total evaluation* are used as consequents. Namely,

$$U = \{(a, v) \mid a \in \mathbb{A} \setminus \{total\ evaluation\}, v \in V\},$$

and for each $v \in V$, the pair $(total\ evaluation, v)$ is designed as a consequent.

Under such a formulation, we apply the Dense-Miner algorithm developed by Bayardo et al. (2000) to extract SC-optimal rules from the soft data for each cluster. By representing each rule by the pair of its support and confidence values, SC-optimal rules for clusters 1–5 are depicted in Figure 3. Intuitively, for each cluster the set of its SC-optimal rules defines a *support-confidence border* above which no rule that satisfies the input constraints can fall. We have not extracted SC-optimal rules for cluster 6 as only few residents in this cluster sent back us their answers to questionnaire.

We would like to recall that the soft data collected through a questionnaire survey reflects the subjective appreciation of human beings on environmental issues, which usually does not have a high consolidation due to vagueness (imprecision and conflict) of linguistic knowledge. Therefore, it is practically impossible to expect for getting a highly significant model that correctly characterizes environmental issues for each cluster. Furthermore, it is also difficult to decide a single goodness metric for constructing the best rule-based model from the soft data for each cluster. Fortunately, by mining SC-optimal rules, we support an interactive phase in which the modeller can browse the optimal rules according to each of several goodness metrics such as *support, confidence, conviction, lift, laplace, gain*, and *Piatesky-Shapiro's measure* (Piatesky-Shapiro, 1991).

In other words, the set of SC-optimal rules includes all rules that are maximal according to each of several metrics mentioned above. Consequently, with SC-optimal rules we can obtain some alternative models according to different goodness metrics so that we can analyze and select. For example, we have used the metric *laplace* on rules and the criterion on the length of IF-part of rules to select the representative model for each cluster, where the *laplace function*, which is commonly used to rank rules for classification purposes, is defined for each rule r as follows

$$\text{laplace}(r) = \frac{\sup(r) + 1}{\sup(r)/\text{conf}(r) + k}$$

The constant k is an integer greater than 1 and usually set to the number of

classes when building a classification model. In our case, we set k to 5, and the obtained models are presented in Tables 2–6.

To evaluate significance of the obtained models, let us give some explanations on rules. Recall that attributes *creature, play, fish, brown, camp, source, plant, total evaluation* respectively stand for questions [a-1] to [a-8] in the questionnaire for evaluating water quality, and the meaning of values is given in Table 1. For the sake of simplicity, we present below the meaning of the model for evaluating water quality of Kanazawa region (cluster 1). The meanings of the remaining models are interpreted in an analogous way.

- R_1^1: If people do not eat fishes they catch in the river and do not make barbecue in the waterside and there are pollutant sources near the waterside then the water quality is very bad.
- R_2^1: If people do not play in the waterside and do not eat fishes they catch in the river and it seems to have pollutant sources near the waterside then the water quality is bad.
- R_3^1: If there are many creatures in the waterside and the water is not so brown then the water quality is medium.
- R_4^1: If the water is clear and people can make barbecue in the waterside and there are many plants in the waterside then the water quality is good.
- R_5^1: If there are many creatures in the waterside and the water is clear and there is no any pollutant source near the waterside then the water quality is very good.

By analyzing the obtained models, it is of interest to see that residents in cluster 1 (Kanazawa region with 5 rivers) pay more attention to the influence of pollutant sources to the quality of water sources. Whilst the model for cluster 1 is a satisfactory one, and consistently reflects the dependence of water quality evaluation on related partial evaluations in subjective appreciation of residents, the model for cluster 2 is not so satisfactory intuitively. This can be interpreted as the region Matto in cluster 2 has many sources of industrial wastes, however, that are processed good and do not influence the water source from the nearby river. The model for cluster 3 is also a satisfactory one, where the cluster consists of two separate regions including Komatsu and Tsubata. The regions both have many water sources (rivers and lakes). As clusters 4 and 5 also have only few data records (under 30 records), it is difficult to get significant models. It should be emphasized that the more residents in each cluster sent us their answers, the more satisfactory the obtained model is. Especially, cluster 1 and cluster 3 have about 500 and 240 data records respectively.

4 An Integrated Model for Evaluation of Regional Environment

In Subsection 2.1 we have used the hard data to do clustering so that each obtained cluster consists of administrative regions having the homogeneous social and economical characteristics relatively. Then the soft data are collected for each cluster, and rule-based models learned from the soft data have been developed in Section 2. Consequently, these models depend upon the clustering result of hard data. In this section we will develop an integrated model for the prediction of evaluating regional environment based on both hard data and soft data. To this end, a structure of models that mimics the one of fuzzy models (Sugeno and Kang, 1988) will be proposed.

It has been known that fuzzy models have advantages of excellent capability to describe a given system and intuitive persuasion toward human operators over linear models. Recently, the fuzzy modelling technique has been applied to identification of fuzzy prediction models in environmental studies, e.g. Nakamori and Ryoke (1994); Ryoke et al. (2000).

As proposed by Takagi and Sugeno (1985), a fuzzy model is a nonlinear model consisting of a number of rules as follows [3]

$$\text{Rule } R^k: \begin{cases} \text{If } u_1 \text{ is } A_1^k, \text{ and } u_2 \text{ is } A_2^k, \text{ and } \ldots, \\ \text{then } y = c_0^k \ + \ \sum_{i=1}^m c_i^k \ x_i. \end{cases} \tag{1}$$

Here, y is the output variable, $u_1, u_2, \ldots$ in the conditional part are called premise variables, and variables $x_i\,(i = 1, \ldots, m)$ in the linear equation of the concluding part are called consequence variables. $A_1^k, A_2^k, \ldots$ are fuzzy sets with membership functions $A_1^k(u_1), A_2^k(u_2), \ldots$, and the coeficients $c_i^k (i = 0, 1, \ldots, m)$ of the linear equation are called consequence parameters.

In this paper, the membership function of A_j^k, for $j = 1, 2, \ldots$ in each rule R_k are defined as follows

$$A_j^k(u_j) = \begin{cases} \exp\Big\{-\dfrac{(u_j - q_{j2}^k)^2}{2(t_{j1}^k)^2(q_{j1}^k - q_{j2}^k)^2}\Big\}, & u_j \le q_{j2}^k \\ \exp\Big\{-\dfrac{(u_j - q_{j2}^k)^2}{2(t_{j2}^k)^2(q_{j3}^k - q_{j2}^k)^2}\Big\}, & u_j \ge q_{j2}^k \end{cases} \tag{2}$$

where parameters $q_{j1}^k, q_{j2}^k, q_{j3}^k$ are determined through the clustering process, and $t_{j1}^k, t_{j2}^k (> 0)$ are tuning parameters with the unit default taken similarly as in Nakamori and Ryoke (1994). Note that if two of these parameters are

[3] See also Nakamori and Ryoke (1994)

equal, we give one of them a small fluctuation to keep the restriction that $q_{j1}^k < q_{j2}^k < q_{j3}^k$. For the sake of simplicity, in the sequel we will denote $A_j^k(u_j)$ by $(q_{j1}^k, q_{j2}^k, q_{j3}^k)$, and A_j^k will be linguistically named as *about* q_{j2}^k.

As is well-known, there are many studies regarding fuzzy modelling mainly based on the pattern-recognition technique and system programming theory among others. In order to build a fuzzy model, Sugeno and Kang (1988) proposed an iterative algorithm that takes into account both the following problems at the same time:

- selection of consequence variables and identification of consequence parameters,
- selection of premise variables and identification of membership functions of fuzzy sets A_j^k, for all k and j.

Keeping these in mind, we now develop integrated models for the prediction of evaluating water quality in each cluster based on both hard data and soft data.

For the problem of selecting of premise variables and identification of membership functions, although we have used a total of 13 attributes concerning the hard data to do clustering, however, we only select the rate of diffusion of sewerage and the population density as the premise variables, and their membership functions are identified through the clustering process. The reason for this choice is that, in our opinion, the residuary attributes mainly depend upon these two attributes. The attributes of partial evaluations on water quality are considered as the consequence variables while the *total evaluation* is considered as the output variable. Furthermore, the soft data based models developed in Section 3 are also incorporated into integrated models. Thus the soft data that supported these models are utilized for regression analysis to identify consequence coefficients. It is of interest to note that the soft data based models depend on the clustering result and the choice of an interestingness metric on rules. Consequently, the integrated models are context-dependent. With the soft data based models obtained in the preceding section, in the following we show integrated models of evaluating water quality in the clusters 1 to 3. Because the cluster 4 to 6 have a few test subjects, it is difficult to get significant models.

Tables 7 to 9 respectively present parameters of membership functions in clusters 1 to 3, where, for short, *population* means the population density and *sewerage* means the rate of diffusion of sewerage. Further, the intervals [min,max] in these Tables denote the supports of respective membership functions.

Under such a construction, the integrated models are represented as in Tables 10–12. Where the linear regression models in the consequence parts cover

the whole data that support the soft data based models respectively. In the equations, the t-ratio of regression coefficients of partial variables with total evaluation are also presented.

We would like to finish this section by showing the relationship between the subjective appreciation of human beings of water quality and BOD (Biochemical Oxygen Demand; mg/l) values as intuitively justification for our research purpose. Put concretely, BOD is the amount of oxygen necessary for bacteria to decompose contaminants chemically into harmless matter. The higher the BOD value is, the more polluted the river becomes. In fact, if the subjective appreciation of human beings does not express environment correctly to some extent, and the soft data based models are not satisfactory intuitively, regression models by soft data do not have meanings. The following presents BOD values corresponding to the evaluation of residents of water quality in each cluster.

[Cluster1]
$0.7 \leq BOD \leq 1.0 \rightarrow$ Evaluation 1-2 (Morimoto River)
$0.7 \leq BOD \leq 2.0 \rightarrow$ Evaluation 2-4
$1.0 \leq BOD \leq 2.0 \rightarrow$ Evaluation 1-2 (Near Kahokugata)
$2.0 \leq BOD \rightarrow$ Evaluation 1-2

[Cluster2]
$BOD \leq 0.6 \rightarrow$ Evaluation 3-5 (in Mikawa)
$1.9 \leq BOD \leq 4.0 \rightarrow$ Evaluation 1-3 (in Nonoichi)
$2.0 \leq BOD \leq 3.7 \rightarrow$ Evaluation 1-3 (in Matto)

[Cluster3]
$0.5 \leq BOD \leq 0.8 \rightarrow$ Evaluation 3-4 (lower reaches)
$0.7 \leq BOD \leq 1.0 \rightarrow$ Evaluation 1-3 (near lakes)
$1.0 \leq BOD \leq 4.1 \rightarrow$ Evaluation 1-2 (near lakes)

[Cluster4]
$0.8 \leq BOD \leq 2.2 \rightarrow$ Evaluation 1-3 (in Unoke)
$0.5 \leq BOD \leq 0.6 \rightarrow$ Evaluation 3-5 (in Kawachi and Torigoe)

[Cluster5]
$2.3 \leq BOD \leq 3.1 \rightarrow$ Evaluation 1-3

[Cluster6]
$0.5 \leq BOD \leq 1.0 \rightarrow$ Evaluation 3-5

By making a correspondence between BOD indices and the meaning of evaluation values (question a-8) described in Table 1, we see that the soft data is rather reliable. Therefore, it would be useful to incorporate the knowledge

learnt from soft data into prediction models for evaluation of regional environment.

5 Conclusion

In this paper, we have developed integrated models for evaluation of regional environment based on both hard and soft data. Since the soft data based models depend on the clustering result and the choice of an goodness metric on rules, the integrated models are context-dependent. It is of interest to note that by mining the SC-optimal rules from the soft data, it supports us an interactive phase in which we can browse the optimal rule according to several goodness metrics. As such modellers have a chance to select the best among models available after making further analyses. We would like to emphasize that one can obtain the soft data based models by a conventional statistical analysis. However, in that case we can get only one model for each cluster, and it is difficult to get high-precision models by our data.

It should be noted that the increasing concern with the effects of human activities on the environment has led to a growing interest in sustainable development that is actually a need of introducing new systems methodologies beyond the traditional ones. In connection with this, Nakamori (2000) have developed a new systems methodology called *i*-system with the emphasis to knowledge creation. By this work we have presented a method for integrating hard and soft data in the knowledge creation process in development of the *i*-system.

Acknowledgements

The authors would like to thank Dr. Roberto Bayardo at the IBM Almaden Research Center for providing the "Dense-Miner" algorithm for constraint-based rule mining on-line, which was applied to our data sets to obtain the rule base used in the paper. Many thanks also go to referees for their constructive comments and valuable suggestions.

References

Agrawal,R., Srikant,R., 1994. Fast algorithms for mining association rules. In: Bocca, J.B., Jarke, M., Zaniolo, C. (Eds.), *Proceedings of 20th Int'l Conference on Very Large Data Bases.* Morgan Kaufmann, pp. 487–499.

Bayardo Jr., R.J., Agrawal, R., 1999. Mining the most interesting rules. In: *Proceedings of the Fifth ACM SIGKDD Int'l Conference on Knowledge Discovery and Data Mining*, pp. 145–154.

Bayardo Jr., R.J., Agrawal, R., Gunopulos, D., 2000. Constraint-based rule mining in large, dense databases. *Data Mining & Knowledge Discovery* **4** (2/3), 217–240.

Bezdek, J.C., 1981. *Patern Recognition with Fuzzy Objective Function Algorithms.* Plenum Press, New York.

Brian, J., 1998. Mining air pollution data. In: *3rd ERCIM Environmental Modelling Group Workshop on Air Pollution Modelling and Measuring Processes*, April 20–21, Madrid, Spain.

Ford, A., 1999. *Modeling the Environment: An Introduction to System Dynamics Modeling of Environmental Systems.* Island Press, Washington DC.

Ishikawa Statistical Information Division. *Ishikawa City and Town Pandect.* Ishikawa Statistical Information Division, Japan.

Kainuma, M., Nakamori, Y., Morita, T., 1990. Integrated Decision Support System for Environmental Planning. *IEEE Transactions on Systems, Man and Cybernetics* **20** (4), 777–790.

Kawano, S., Shakato, M., Ryoke, M., Nakamori, Y., 2001. Soft data analysis for evaluating regional environment. In: *Proceedings of Joint 9th IFSA World Congress and 20th NAFIPS International Conference.* Vancouver, Canada, pp. 116–121.

Nakamori, Y., Ryoke, M., 1994. Identification of fuzzy prediction models through hyperellipsoidal clustering. *IEEE Transactions on Systems, Man and Cybernetics* **24** (8), 1153–1173.

Nakamori, Y., Sawaragi, Y., 1997. Methodology and Systems for Environmental Decision Support. *Annual Reviews in Control* **20**, 143–154.

Nakamori, Y., Sawaragi, Y., 2000. Complex Systems Analysis and Environmental Modeling. *European Journal of Operational Research* **122** (2), 178–189.

Nakamori, Y., 2000. Knowledge management system toward sustainable society. In: *Proceedings of KSS'2000: "Knowledge and Systems Sciences: Challenges to Complexity"*, JAIST, Ishikawa, Japan, September 25–27, pp. 57–64.

Piatesky-Shapiro, G., 1991. Discovery, Analysis, and Presentation of Strong Rules. Chapter 13 of *Knowledge Discovery in Databases*, AAAI/MIT Press.

Ryoke, M., Nakamori, Y., Heyes, C., Makowski, M., Schöpp, W., 2000. A simplified ozone model based on fuzzy rules generation, *European Journal of Operational Research* **122** (2), 440–451.

Sugeno, M., Kang, G.T., 1988. Structure identification of fuzzy model. *Fuzzy Sets and Systems* **28**, 15–33.

Takagi, T., Sugeno, M., 1985. Fuzzy identification of systems and its applications to modeling and control. *IEEE Transactions on Systems, Man and Cybernetics* **15** (1), 116–132.

Ward Jr., J.H., 1963. Hierarchical Grouping to Optimize an Objective Func-

tion. *Journal of American Statistical Association* **58**, 236–244.

Figures and Tables

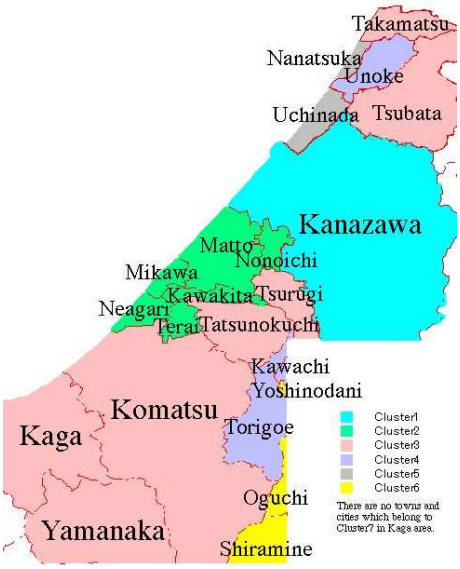
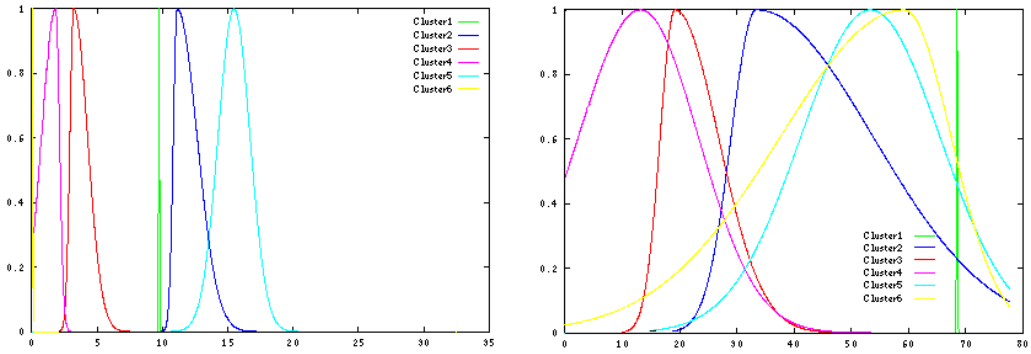
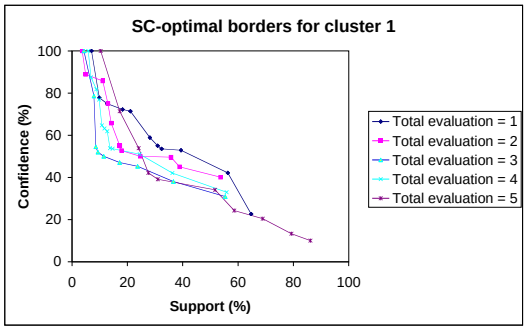


Fig. 1. The result of clustering of Kaga area

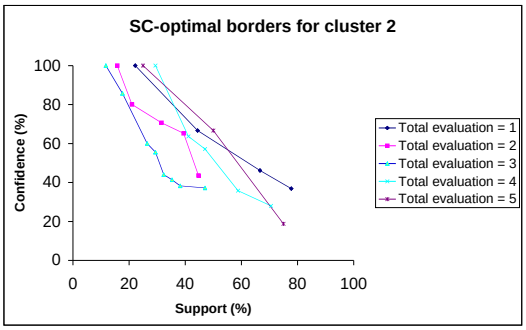


a) Population b) Sewerage

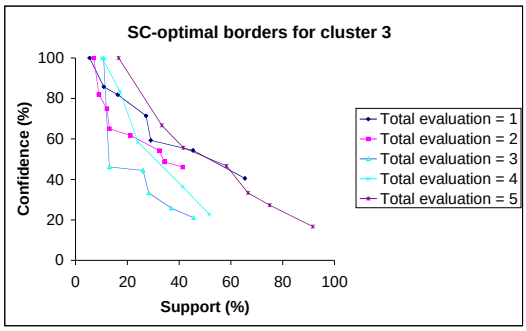
Fig. 2. The membership functions of population and diffusion rate of sewerage



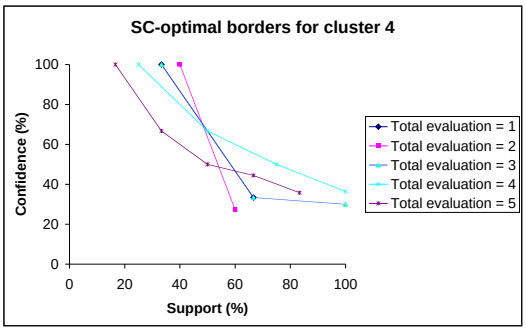
a)



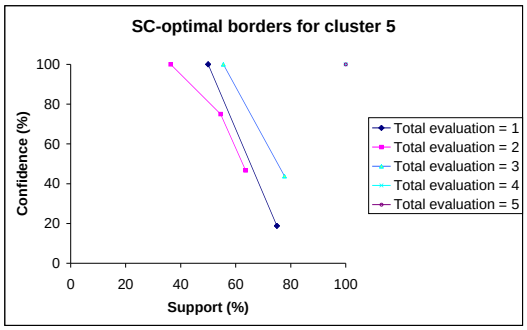
b)



c)



d)



e)

Fig. 3. SC-optimal rules

Table 1
The meaning of values of variables

Question	Value	Meaning	Question	Value	Meaning
[a-1]	1	no with a high confidence	[a-8]	1	very bad
≀	2	no with a low confidence		2	bad
[a-7]	3	neutral		3	medium
	4	yes with a low confidence		4	good
	5	yes with a high confidence		5	very good

Table 2
A rule-based model for evaluating water quality of cluster 1

ID-Rule	IF	THEN
R_1^1	$fish = 1$ and $camp = 1$ and $source = 5$	$total\ evaluation = 1$
R_2^1	$play = 1$ and $fish = 1$ and $source = 4$	$total\ evaluation = 2$
R_3^1	$creature = 5$ and $brown = 2$	$total\ evaluation = 3$
R_4^1	$brown = 1$ and $camp = 5$ and $plant = 5$	$total\ evaluation = 4$
R_5^1	$creature = 5$ and $brown = 1$ and $source = 1$	$total\ evaluation = 5$

Table 3
A rule-based model for evaluating water quality of cluster 2

ID-Rule	IF	THEN
R_1^2	$play = 1$ and $fish = 1$ and $plant = 5$	$total\ evaluation = 1$
R_2^2	$brown = 4$ and $source = 4$	$total\ evaluation = 2$
R_3^2	$fish = 5$ and $brown = 1$	$total\ evaluation = 3$
R_4^2	$fish = 5$ and $camp = 5$ and $source = 4$	$total\ evaluation = 4$
R_5^2	$creature = 5$ and $brown = 1$ and $source = 4$	$total\ evaluation = 5$

Table 4
A rule-based model for evaluating water quality of cluster 3

ID-Rule	IF	THEN
R_1^3	$brown = 5$ and $camp = 1$ and $source = 5$	$total\ evaluation = 1$
R_2^3	$brown = 2$ and $source = 4$	$total\ evaluation = 2$
R_3^3	$creature = 5$ and $brown = 1$	$total\ evaluation = 3$
R_4^3	$creature = 4$ and $source = 2$ and $plant = 4$	$total\ evaluation = 4$
R_5^3	$creature = 5$ and $fish = 5$ and $brown = 1$ and $plant = 5$	$total\ evaluation = 5$

Table 5
A rule-based model for evaluating water quality of cluster 4

ID-Rule	IF	THEN
R_1^4	$play = 4$	$total\ evaluation = 1$
R_2^4	$brown = 1$ and $camp = 5$	$total\ evaluation = 2$
R_3^4	$creature = 5$ and $play = 5$ and $plant = 5$	$total\ evaluation = 3$
R_4^4	$creature = 5$ and $fish = 5$ and $camp = 5$ and $plant = 5$	$total\ evaluation = 4$
R_5^4	$fish = 5$ and $brown = 1$ and $plant = 5$	$total\ evaluation = 5$

Table 6
A rule-based model for evaluating water quality of cluster 5

ID-Rule	IF	THEN
R_1^5	$play = 5$ and $camp = 2$	$total\ evaluation = 1$
R_2^5	$brown = 4$	$total\ evaluation = 2$
R_3^5	$source = 2$	$total\ evaluation = 3$
R_4^5	$brown = 5$ and $camp = 5$ and $plant = 5$	$total\ evaluation = 4$
R_5^5	$source = 1$	$total\ evaluation = 5$

Table 7
Parameters of membership functions in cluster 1

	min	q_{j1}^1	q_{j2}^1	q_{j3}^1	max
population ($j = 1$)	9.74	9.74	9.74	9.74	9.74
sewerage ($j = 2$)	68.5	68.5	68.5	68.5	68.5

Table 8
Parameters of membership functions in cluster 2

	min	q_{j1}^2	q_{j2}^2	q_{j3}^2	max
population $(j = 1)$	3.06	10.8	11.1	12.7	32.5
sewerage $(j = 2)$	.00	29.0	33.4	53.9	76.2

Table 9
Parameters of membership functions in cluster 3

	min	q_{j1}^3	q_{j2}^3	q_{j3}^3	max
population ($j = 1$)	.692	2.91	3.20	4.26	6.02
sewerage ($j = 2$)	1.50	16.6	19.2	26.8	33.3

Table 10
An integrated model for evaluating water quality of cluster 1

IF	THEN	
population is *about* 9.74 and	*total evaluation* = 3.778	$+0.178 \times fish$
sewerage is *about* 68.5 and		(t-ratio = +3.207)
rules R_1^1–R_5^1 are true		$-0.222 \times brown$
		(t-ratio = -4.120)
		$-0.458 \times source$
		(t-ratio = -9.187)

Table 11
An integrated model for evaluating water quality of cluster 2

IF	THEN	
population is *about* 11.1 and	*total evaluation* = 1.045	$+0.241 \times fish$
sewerage is *about* 33.4 and		(t-ratio = +2.337)
rules R_1^2–R_5^2 are true		$-0.538 \times brown$
		(t-ratio = -3.499)
		$+0.495 \times source$
		(t-ratio = +2.866)

Table 12
An integrated model for evaluating water quality of cluster 3

IF	THEN	
population is *about* 3.20 and	*total evaluation* = 3.823	$-0.441 \times brown$
sewerage is *about* 19.2 and		(t-ratio = -3.298)
rules R_1^3–R_5^3 are true		$-0.233 \times source$
		(t-ratio = -2.371)
		$+0.149 \times plant$
		(t-ratio = +1.561)