Available online at www.sciencedirect.com

SCIENCE DIRECT®

Ecological Modelling 174 (2004) 421–431

ECOLOGICAL MODELLING

www.elsevier.com/locate/ecolmodel

A data-mining approach to determine the spatio-temporal relationship between environmental factors and fish distribution

Fenzhen Su [a,*], Chenghu Zhou [a], Vincent Lyne [b], Yunyan Du [a], Wenzhong Shi [c]

[a] *LREIS, Chinese Academy of Sciences, Institute of Geographical Science and Resources Research, Building 917, Datun Anwai, Beijng 10010, China*

[b] *Marine Research, Commonwealth Scientific and Industrial Research Organisation, Parkes, Australia*

[c] *Department of Land-Surveying and Geo-Informatics, Hong Kong Polytechnic University, Hong Kong, China*

Received 12 August 2002; received in revised form 30 August 2003; accepted 27 October 2003

Abstract

The interaction between environmental factors and the spatiotemporal dynamics of living organism is an important aspect in ecology. We describe here a data-mining approach—the spatiotemporal assignment mining model (STAMM)—to extract the spatiotemporal pattern, or assignment of environmental factors, which control the distribution of a living organism. In STAMM, the spatiotemporal assignment of environmental factors is expressed via neighbourhood rules which will reflect the fuzzy or uncertain prior knowledge about the relationship. The values of cells or points in the neighbourhood and the relationships are used to construct a decision table. Indices expressing the probabilities of the ecological association rules are recursively processed in order to determine the spatiotemporal assignment. These rules are objective assessments of our prior knowledge and they refine our knowledge and understanding of the ecosystem. As a case study, we used this model to study the temperature pattern which controls the assembling of fish in the Dasha area of the Yellow Sea in China.

Keywords: Geographical Information System(GIS); Spatiotemporal assignment; Ecological association rule; Fish assembling; Fish distribution

1. Introduction

The regional spatio-temporal distribution of populations of organisms is intimately controlled by local interactions between organisms and their surrounding environmental characteristics (Brosse et al., 1999). Modelling and simulation are useful tools to characterise key aspects of the dynamics of ecosystems but the capabilities of these models depend on the available modelling techniques and computing power (Giske et al., 1998). Many existing models on the interaction between the environmental factors and the organism do not explicitly consider the spatial aspects (Kracker, 1999; Mackinson, 2000; Meaden, 2000). However recently a number of modelling approaches have utilised spatial information to analyse the distribution, or the movement, of living marine organisms (Hays et al., 2001; Letcher and Rice, 1997; Meaden, 2000; Solanki et al., 2001). Brown et al. (2000), Gertseva and Gertsev (2002), Ji and Jeske (2000).

Some new techniques, or models, have been developed to analyse the spatial influence from the environment factors, for example, ANN (Brosse et al., 1999), IBM (Bian, 2003; Fang et al., 1999), ES (Mackinson, 2000), and so on. With the development of spatial information technology, Geographical Information System (GIS) has been used to study

* Corresponding author. Fax: +86-10-64889630.
E-mail address: sufz@lreis.ac.cn (F. Su).

0304-3800/$ – see front matter © 2003 Elsevier B.V. All rights reserved.
doi:10.1016/j.ecolmodel.2003.10.006

the influence of environmental factors (Du et al., 2001; FAO-CRODT-ORSTOM, 1995; Meaden, 2000; Somers and Long, 1994; Su et al., 2000, 2001).

Some of these research approaches analysed qualitatively the relationship between the distribution of environmental factors and that of the living organism. Others used the overlay function of the GIS to evaluate the suitability index of each point or area based on the hypothesis the organism in one point or one area is just influenced by factors in the local point or area. However, these models don't consider the spatial relationship or assignment among the environmental factors.

Living organisms are affected by local environmental factors as well as factors around their neighbourhood. The neighbourhood factors may influence an organism independently, but the more complex issue is that the influence is through spatial patterns and spatial relationships. In this paper, the influence of spatial structure, or the spatial assignment of the environmental factors, will be considered using as an example the spatio-temporal influence of sea-surface temperature (SST) on the distribution of pelagic fish.

Fig. 1 shows that Spanish mackerel will assemble in part of the area where the value of SST is in the range of 9.3–10.8 °C in the south of Yellow Sea in the first ten days in April in some years (Wei, 1988). However, a simple linear regression between SST and the density of fish (Fig. 2) cannot explain why there is no fish in part of the area where SST is in the range of 9.3–10.8 °C (Fig. 1) and why there is no

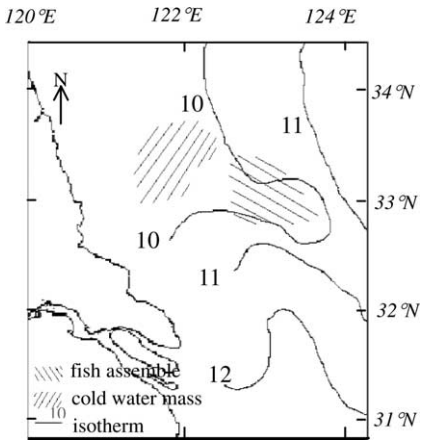

Fig. 1. Fishing areas and SST.

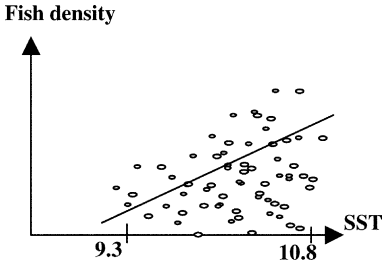

Fig. 2. Regression between SST and fish density.

assembling in some years although the temperature is in the same range. A possible reason is that according to anecdotal experience, the distribution of fish in this area is determined by the presence of cold water mass in the northwest in the period (Fig. 1). This cold water mass will block the migration of fish in the period (Hu, 1995; Wei, 1988). Thus, the fish distribution relies on the status of SST in this area and that to its north so the regional spatial pattern of SST is important in explaining distributions of fish.

Many studies used correlation analysis or multiple least-square regression (MLR) to parametrise the relationship between the environmental factors and the spatial occupation of living organisms (Binns and Eiserman, 1979; Ricker, 1975). The linear MLR method is now a statistical tool that is used routinely in ecology, but suffers from some drawbacks in that the relationships between variables found in environmental sciences are often nonlinear (James and McCulloch, 1990).

In our case, it is difficult to take into account the proper spatial relationship of the cold water mass with a linear regression method. Nonlinearities between the cold water area and the fish distribution arise because of the Boolean relationship where the fish go north when the cold water mass does not form. Hence, it is difficult, with a quantitative analysis method, to extract the spatial structure of environmental factors which determine the distributional behavior of fish. Integrated information techniques are clearly required in this field.

At the same time, life cycle processes affect the spatio-temporal response of fish to the environmental variations. For example, outside of the migration period of the fish, there is no fish assembling when the time is not in the period of migration of the fish although the SST distributes as Fig. 1. That means

that both spatial assignment with its corresponding time determines the aggregation pattern of the fish. So, our approach in investigating this problem was to set up a spatiotemporal assignment mining model (STAMM) to extract the spatiotemporal assignments that "restrict" the distribution of fish. This assignment and its corresponding distribution of fish can be expressed as rules which can be used as knowledge in a decision support system or an expert system. We define this kind of rule as an ecological association rule.

Much prior knowledge about the ecosystem, which maybe fuzzy or uncertain, can be used to guide the formulation of the analysis by suggesting types of relationships and associations which might be operating. Analysis of the data and the relationships in STAMM provides more exact and quantitative knowledge about the associations which may be operating within the ecosystem. This is one of the main aims of ecological modelling (Jorgensen, 1997).

2. STAMM

2.1. Ecological association rules

In deriving the ecological association rules, our modelling approach will adopt some techniques used in data warehouse mining or knowledge discovery. Spatial association means the relationship among objects that are distributed in space (Fayyad et al., 1996; Koperski and Han, 1995). That is to say, the spatial attribution, or scale of an object, can be inferred from other objects that are related with it in space.

In this paper, we develop modelling approaches to examine spatiotemporal associations. Spatiotemporal association means that one spatiotemporal event can be inferred by another spatiotemporal event. Once the attributes in a spatial structure at a special time get their values, we say that a spatiotemporal event has happened. In this paper, the spatiotemporal event is an ecological event and the spatiotemporal association is an ecological association. For example, an event occurs when fish assemble when the SST at position A is 5 °C with 10 °C at position B at 3 o'clock. If assembling of the fish is controlled by the SST of A and B at 3 o'clock, then an ecological association rule can be expressed as (T, 3 o'clock) $\wedge$ (A, 5) $\wedge$ (B, 10) $\rightarrow$ (fish, assembling).

2.2. Spatiotemporal assignment

As mentioned above, the behavior of the living organism is influenced by local environment factors, and also by the pattern of the factors around the local place. The pattern includes the distribution of objects with their spatial relationships. The spatial relationships include those associated with topology, distance, direction and the relative value of attributes. The pattern will change with time. We call the pattern with its corresponding time as spatiotemporal assignment.

In order to express the spatial pattern, we divide the 2-dimension space as a grid and we set the state of the living organism in one cell as the dependent variable of interest. We define this cell as the focus. Our prior qualitative or fuzzy knowledge are used to guide the selection of appropriate explanatory factors within some areas or points around the focus which we expect will affect the state of the focus. These areas are defined as the neighbour, and the focus together with the neighbours makes up the neighbourhood. As Fig. 3 shows, the shaded cell is the focus and the blank cells or the points are the neighbours.

The shape of the neighbourhood can be a rectangle, a circle, an annulus or a slice of a circle, Neighbourhoods can also be created by an irregular grid size. We can extract one or some factors from the neighbourhood which might effect the behavior of the state of the living organism at the focus. For example, temperature and chlorophyll will affect the behavior of fish, thus, we can get their values and their relative values from each of the cells or points in the neighbourhood. Or, if we just want to explore the relationship between the temperature and the neighbour variables, we can extract the temperature and its relative value only.

The behavior of the living organism at the focus varies with the values of factors in the neighbourhood through time. The spatiotemporal assignment is the quantitative expression of the spatial pattern of factors, at the corresponding time, that is relevant to the state of the focus. We define the behavior of the living organism with its spatiotemporal assignment as an ecological event (EE). All the factors, or their spatial relative relationship, the time and the behavior at the focus are defined as attributes of the ecological event.

The neighbourhood is defined by the qualitative or fuzzy knowledge or prior experience. That means some factors in some cells may not influence the

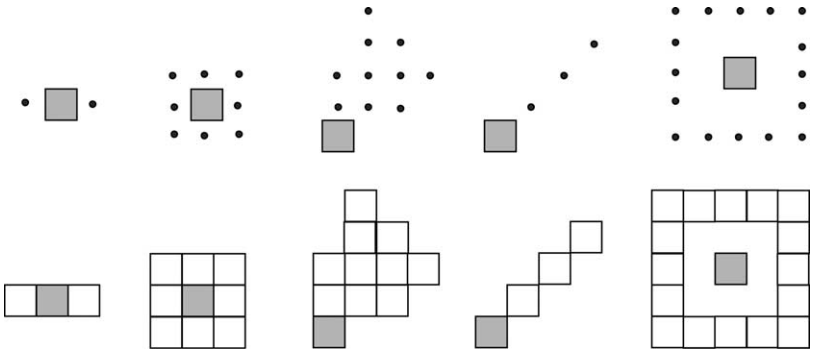

Fig. 3. The structure of neighbours.

behavior at the focus. These factors or the cells can be unselected by STAMM through the algorithm in Section 2.5.

2.3. From ecological event to decision table

In order to extract the spatiotemporal assignment, a decision table should be constructed. As an example, we divide the research space as a raster as in Fig. 4e and the cells are identified by an ID, as shown in Fig. 4e. We assume, from prior knowledge, the behavior of the living organism is influenced by the state in its-four neighbour cells. So we set the neighbourhood as the cells which are identified by letters in Fig. 4a, where cell F is the focus.

We assume, from prior knowledge, that the behavior of the living organism is influenced by factors *A* and *B*. The numbers in Fig. 4b are the values of factor *A* at time T_1, numbers in Fig. 4c are the values of factor *B* at T_1, the number or letter in Fig. 4d is the behavior of the living organism in the focus cell at time T_1.

The numbers in Fig. 4f are the values of factor *A* at time T_2, the numbers in Fig. 4g are the values of factor *B* at T_2, the number or letter in Fig. 4h is the behavior of the living organism in the focus cell at time T_2.

The decision table is constructed by overlaying the neighbourhood on the analysis area to obtain the ecological event, and the attributes of the event are identified (ID) and tabulated as in Tables 1 or 2. For the particular example in Table 2, the attribute for the focus, denoted D in the table, is a binary presence or absence (Y or N) of the decision attribute. Each row in Table 2 is derived by successively moving the focus across the cells of the analysis grid and recomputing

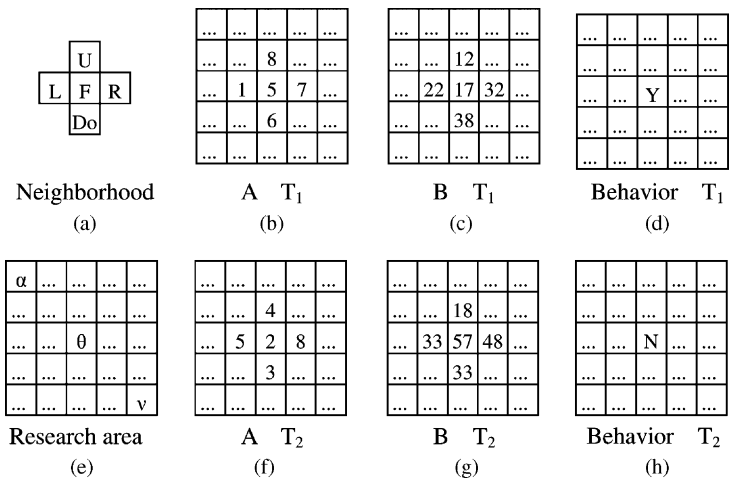

Fig. 4. The sea surface expressed as a raster grid.

Table 1
Basic decision table

ID	T	A_F	B_F	D
...	...	...	...	...
θ	T_1	5	17	Y
θ	T_2	2	57	N
...	...	...	...	...

the attributes of the neighbourhood and the decision attribute at the focus. In this way, the attribute-encoded information is catalogued in the decision table ready for analysis.

Factors A and B together with the time are jointly referred to as the condition attribute C, which has the associated behavior attribute D at the focus cell.

Once the table is populated, a number of analysis procedures can be performed—such as a standard regression analysis. In STAMM, the basis of the analyses are a set of rules which can be mined. These rules are, for example from Table 1, of the form: (ID, θ) ∧ (T, T_1) ∧ (A, 5) ∧ (B, 17)→(D, Y). The detailed description of the mining method and algorithm will be left for Sections 2.4 and 2.5. The above rule means that for the row representing the condition and decision identified by θ, when factors A and B are 5 and 17, respectively, at time T_1, then the decision attribute, or activity of organism at the focus, is Y. Such a simple relationship, however, does not consider the spatial relationships which may be operating in the neighbourhood. Thus, results derived from such rules do not reflect the influence of surrounding factors on the activity at the focus.

Assuming that the value of D is affected by the value of factors A and B in the focus cell and that in its neighbourhood, we should construct the table as Table 2 to extract rules such as:

$$(\mathrm{ID}, \theta) \wedge (T, t_1) \wedge (A_U, 8) \wedge (A_F, 5) \wedge (B_U, 12) \wedge (B_F, 17) \rightarrow (D, \mathrm{Y})$$

In addition to the values of neighbour cells, relative values between neighbouring cells, reflecting for instance gradients, or potential energy between two points or areas, can also be used to construct more sophisticated rules such as those in Table 3. An example rule set constructed from Table 3 may be of the following form:

$$(\mathrm{ID}, \theta) \wedge (T, t_1) \wedge (A_U, 8) \wedge (A_F, 5) \wedge (B_U, 12) \wedge (B_F, 17) \wedge (B_R - B_F, 15) \rightarrow (D, \mathrm{Y})$$

Rules such as these which integrate the attributes of space, time and spatial relationship can be extracted from Table 3.

2.4. Describing the ecological association rule (EAR)

In order to systematically study the mining of ecological association rules, we define some preliminary concepts to describe the rules.

Table 2
Attributes of events with spatial neighbourhood

ID	T	A_U	A_L	A_F	A_R	A_{Do}	B_U	B_L	B_F	B_R	B_{Do}	D
...	...	...	...	...	...	...	...	...	...	...	...	...
θ	T_1	8	1	5	7	6	12	22	17	32	38	Y
θ	T_2	4	5	2	8	3	18	33	57	48	33	N
...	...	...	...	...	...	...	...	...	...	...	...	...

Table 3
Attributes of events with spatial relative relationships

ID	T	A_U	...	A_{Do}	B_U	...	B_{Do}	A_F–A_U	A_R–A_F	...	B_F–B_U	B_R–B_F	D
...	...	...	...	...	...	...	...	...	...	...	...	...	...
F	T_1	8	...	6	12	...	38	−3	2	...	5	15	Y
F	T_2	4	...	3	18	...	33	−2	6	...	39	−9	N
...	...	...	...	...	...	...	...	...	...	...	...	...	...

Definition. An ecological association rule relates a decision attribute D with a condition C such that:

$$C \to D \quad C \subset I, D \subset I \text{ and } C \cap D = \Phi \tag{1}$$

where $I = \{i_1, i_2, \ldots, i_m\}$ denotes the collection of attributes associated with the ecological event (EE), and the item i_m may denote a temporal, spatial, non-spatial, or a relative value attribute. Once the values of I are set, we call it an ecological event (EE) and use EEID to identify it. The list of ecological events (EE) are arranged as a decision table and identified as an ecological decision table (EDT). Each recorded row in the table is an ecological event. The formula in (1) implies that the condition attribute C can be used to infer or resolve the status of the decision attribute D.

For describing the importance of each rule, we define four indices:

Index 1: Confidence

$$C_{\mathrm{CD}} = \frac{|\{\mathrm{EE} : C \wedge D \subseteq \mathrm{EE}, \mathrm{EE} \in \mathrm{EDT}\}|}{|\{\mathrm{EE} : C \subseteq \mathrm{EE}, \mathrm{EE} \in \mathrm{EDT}\}|} = c\% \tag{2}$$

C_{XY} expresses the Confidence of the rule $C \to D$. The formula means that c% of conditions C support the decision D on the premise that C is in the EDT. That is to say, C_{CD} shows how much of the probability of D is captured in C. For example, if there are 100 ecological events within which the consistency of chlorophyll lies in the range of R in a special area A, and of those events 60 result in fish assembling at the focus. That means the Confidence of the rule:

$$(\text{consistency of chlorophyll}, R) \wedge (\text{Position}, A) \to (\text{fish, assembling}) \text{ is } 60\%.$$

Index 2: Support

$$S_{\mathrm{CD}} = \frac{|\{\mathrm{EE} : C \wedge D \subseteq \mathrm{EE}, \mathrm{EE} \in \mathrm{EDT}\}|}{|\mathrm{EDT}|} = s\% \tag{3}$$

S_{CD} is the Support of the rule $C \to D$. It is defined as s% of the events in the table which support C and D. That is to say, S_{CD} shows how much the probability of D together with C is. For example, if there are a total of 200 records in the table and there are 60 records that (chlorophyll, R) $\wedge$ (Position, A) and (fish, assembling) in the same record.

That means the overall Support of the rule:

$$(\text{chlorophyll}, R) \wedge (\text{Position}, A) \to (\text{fish, assembling}) \text{ is } 30\%.$$

Expected Confidence

$$\mathrm{EC}_{\mathrm{CD}} = \frac{|\{\mathrm{EE} : D \subseteq \mathrm{EE}, \mathrm{EE} \in \mathrm{EDT}\}|}{|\mathrm{EDT}|} = e\% \tag{4}$$

$\mathrm{EC}_{\mathrm{CD}}$ is the Expected Confidence of the rule $C \to D$. That means e% events in the table support D. That is to say, $\mathrm{EC}_{\mathrm{CD}}$ shows how much the probability of D is in the table. For example, there are 200 records in table and there are 80 records that (fish, assembling). That means the Expected Confidence of the rule:

$$(\text{chlorophyll}, R) \wedge (\text{Position}, A) \to (\text{fish, assembling}) \text{ is } 40\%.$$

Lift

$$L_{\mathrm{CD}} = \frac{C(C \to D)}{\mathrm{EC}(C \to D)} \tag{5}$$

L_{CD} is the Lift of the rule $C \to D$. It describes, to D, how important C is. If there are 200 records in the table, and of those 100 are of records (chlorophyll, R) $\wedge$ (Position, A), 80 records (fish, assembling), 60 records (chlorophyll, R) $\wedge$ (Position, A) $\wedge$ (fish, assembling). That means the Lift of the rule:

$$(\text{chlorophyll}, R) \wedge (\text{Position}, A) \to (\text{fish, assembling}) \text{ is } 1.5.$$

The indices also can be expressed by probability formulas which are summarised in Table 4.

The Confidence evaluates the suitability of the rule when it is used to infer the ecological event. The Support describes how important the rule is in the ecological event table. The Expected Confidence describes the support for D when C is not taken into account. The Lift describes the influence on D from C. The higher the Lift is, the more the influence on D from C.

Table 4
Probability formula of the index of the association rule

Index	Description	Formula
Confidence	Probability of D at the premise of C	$P(D/C)$
Support	Probability of intersection of D and C	$P(C \cap D)$
Expected Confidence	Probability of D	$P(D)$
Lift	Ratio of confidence to expected confidence	$P(D/C)/P(D)$

In general, the value of the Lift of a useful rule should be larger than 1 because if the value of Confidence is larger than that of the Expected Confidence, it implies that C does affect D or there is some relationship between them. If the value of the Lift is smaller than 1, the rule is not of value.

The Support or the Confidence can reflect the appropriateness of the rule mined directly. The definitions of EAR and its indices show there are always association rules between any two attributes although the values of the indices of the rules are different. So, an infinite number of rules will be found from the table if we do not impose the minimum value of Confidence or Support. In fact, the interesting rules are those whose Support or Confidence is larger than a special value. So the threshold for the Support or Confidence should be given when mining the rule. The threshold of Support is called the minSupport and the threshold of Confidence is named the minConfidence. Too many valueless rules will be mined and the computing time will be too long if the threshold is too small. And, valuable rules will be lost if the threshold is too large.

2.5. Mining algorithm

After the ecological events are expressed in Table 3, the mining algorithm should be applied. Our algorithm is designed as a recursive algorithm based on the Frequent Item Set (Agrawal et al., 1993) paradigm. It can be divided into two phases:

(1) Searching attributes whose Confidence or Support is larger than the minConfidence or minSupport level. These attributes are named as the Frequent Item Set.
(2) Selecting the rules from these Frequent Item Sets. In general, if $A_1A_2A_3A_4$ and A_1A_2 are Frequent Item Sets and the Confidence or the ratio of Supports of $A_1A_2A_3A_4$ and A_1A_2 is larger than minConfidence, that rule $A_1A_2 \rightarrow A_3A_4$ is selected.

The detail process list of the algorithm is as follows:

(1) Frequent set $L_1 = \{\text{large } 1 - \text{item sets}\}$
(2) for ($k = 2$; $L_{k-1} \neq \Phi$; k++) do begin
(3) $C_k = -$extension (L_{k-1}); //new candidate set, extensions from L_{k-1}
(4) for all ecological events EE $\in$ EDT do begin
(5) $C_t =$ subset (C_k t); //candidate set contained in EDT
(6) for all candidates $c \in C_t$ do
(7) c. count++
(8) end
(9) $L_k = \{c \in C_k / \text{c. count} \geq \text{minsupp}\}$
(10) end
(11) Answer $= \cup_k L_k$

The algorithm requires firstly getting the frequent 1-item set L_1, then getting the next frequent 2-item set L_2, this iteration is repeated till the point at which L_k is none, and the process is stopped. For each attribute in the aggregate of candidate K-item set, C_k is an attribute aggregate which is added to the frequent set L_{k-1}.

In general, if the attribute is a numeric value and its domain of values is large, it should firstly be partitioned into different intervals, after which it can be mapped for each (attribute, interval) pair to a boolean attribute which is more suited to applying the algorithm. For example, if the domain of value of A is [l, r], it can be divided into ranges [l_k, r_k] and the (A, value) can then be mapped to (A, K).

3. Experiment

In this section, we use the model to study why and when the fishing ground will come into being in the Dasha area in the north of Yellow Sea. The fishing ground is assumed to be established where the catch per net is larger than 500 boxes.

Our analysis grid, surrounding the fishing ground is shown in Fig. 5 with the decision focus cell denoted by a box surrounded by a neighbourhood comprising cells denoted $T_1, T_2, \ldots, T_9, T_a, \ldots, T_g$. We extract

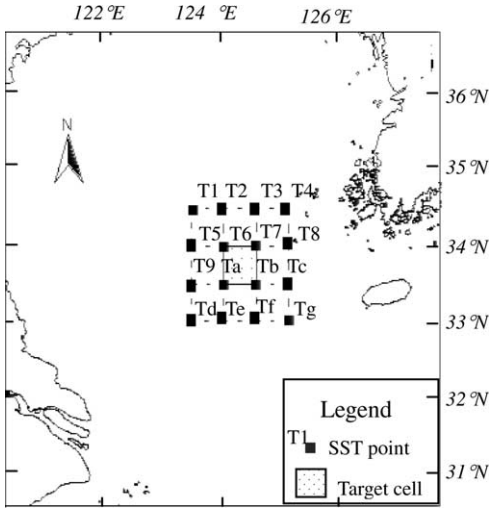

Fig. 5. The neighbourhood definition.

the temperature of the neighbourhood and the fish density at the target as in Table 5, where ID denotes the identification number for the target cell, Week denotes the week number. For example, 8801 denotes the first week in the year of 1988. T_1, T_2, ..., T_f, T_g denotes the temperature at their respective positions. T_g–T_1, T_g–T_2, ..., T_d–T_3, T_d–T_4 are the relative values of SST between the respective cells. Density is the catch per net in the target cell.

The SST is derived from the NOAA satellite image and the density is calculated from fishing catch data. We assume that temperature represents the condition attributes and density as the decision attribute.

3.1. A correlation analysis

We used the nonparametric Spearman rank correlation to test the relationship between the condition attributes and the decision attribute in Table 5. We found that the significance level is relatively large that the correlation is not significant. However, despite this we can not say that there is no relationship because a correlation, or rank correlation, analysis may not be suitable in this application.

3.2. Spatiotemporal assignment extracted by STAMM

In order to apply the STAMM algorithm, we simplified the attribute information by firstly partitioning the values into intervals as follows and constructing the attribute (Table 6):

Intervals for temperature: a: ∼12.9; b: 13.0–15.0; c: 15.1

Intervals for the density: Yes: density > 500 boxes per net; No: density ≤ 500 boxes per net

Intervals for the relative temperature:

(1) Y: $\Delta T \geq 1$; N: $\Delta T < 1$. When the distance between two positions is one cell.
(2) Y: $\Delta T \geq 1.5$; N: $\Delta T < 1.5$. When the distance between two positions is two cells.
(3) Y: $\Delta T \geq 2$; N: $\Delta T < 2$. When the distance between two positions is three cells.

Many rules can be extracted from the table and the choice as to which are useful or not depends on the analysis, or interpretation, which is possible and

Table 5
Fishing ground condition attributes for different weeks, and the density attribute at the focus

ID	Week	T_1	T_6	T_7	T_a	T_b	T_g	...	T_g–T_1	T_a–T_7	Density
888	8801	11.5	14.2	14.1	15.1	17	18	...	6.5	1	0
888	8901	11.7	11.2	12.1	11.3	12.3	13.8	...	2.1	−0.8	4170
888	9001	13.3	14	12	14.7	15	16	...	2.7	2.7	1275
888	9101	13	13.1	13	14.8	13.8	15	...	2.0	1.8	2650
888	9201	14.2	13.7	13.7	14.6	14.3	15.2	...	1	0.9	0
888	9301	13	13.9	12.9	14	14	17	...	4	1.1	1112
888	9401	12.3	14.8	15	13.4	14	17.1	...	4.8	−1.6	740
888	9501	12.5	12.7	12.3	14	13	16.4	...	3.9	1.7	0
888	9601	12.7	12.6	11.6	13.1	13.5	14.6	...	1.9	1.5	0
888	9701	11.5	13	13.9	11.7	13.9	16.5	...	5	−2.2	0
888	8802	11.4	14.5	14.0	15.3	16.7	17.8	...	6.4	1.3	0
...	...	...	...	...	...	...	...	...	...	...	...

Table 6
Coded interval attributes for the STAMM decision table analysis

ID	Week	T_1	T_6	T_7	T_a	T_b	T_g	...	T_g–T_1	T_a–T_7	Density
888	8801	a	b	b	c	c	c	...	Y	Y	No
888	8901	a	a	a	a	a	b	...	Y	N	Yes
888	9001	b	b	a	b	b	c	...	Y	Y	Yes
888	9101	b	b	b	b	b	b	...	Y	Y	Yes
888	9201	b	b	b	b	b	c	...	N	N	No
888	9301	b	b	a	b	b	c	...	Y	Y	Yes
888	9401	a	b	b	b	b	c	...	Y	Y	Yes
888	9501	a	a	a	b	b	c	...	N	Y	No
888	9601	a	a	a	b	b	b	...	Y	Y	No
888	9701	a	b	b	a	b	c	...	Y	Y	No
888	8802	a	b	b	c	c	c	...	Y	Y	No
...	...	...	...	...	...	...	...	...	...	...	...

guided by our knowledge of the fishery. For example, we can choose one rule with the highest Confidence to analyse:

$$(\text{ID}, 888)\wedge(\text{WEEK}, 1)\wedge(T_6,b)\wedge(T_a,b)\wedge(T_g\text{–}T_1, \text{Y}) \rightarrow (\text{Density}, \text{Yes})\ 100\%$$

The rule shows, for the target cell in the first week, there will be a fishing ground if the temperature of its left-upper position point lies in the range of 13–15 °C and that of its left-lower is in the range of 13–15 °C and there is a special difference of temperature in the direction of southeast-northwest. The difference of temperature in the direction of southeast-northwest makes the fish concentrate in a relative small area hence the density of fish is high.

4. Discussion

The example experiment shows that in situations where simple linear or nonparametric correlation approaches fail, rule-based data-mining approaches such as the STAMM algorithm developed in this paper can successfully model the information.

The STAMM model can incorporate fuzzy knowledge as prior experience in formulating the definition of the neighbourhood and the appropriate condition attributes. Through the model the fuzzy knowledge can be refined in the selection of those rules with good performance criteria (confidence, expected confidence, support and lift). As the experiment shows, prior fuzzy experience suggested an expected influence on fishing assembling, SST, and SST gradients, around the focus area. Analysis by the STAMM model yielded a robust rule which incorporated this prior knowledge as shown in Section 3.2.

The model based on a raster grid makes it convenient to get information on spatial relationships between cells. And the raster data structure is suitable for analysing temporal change. This is also the format for data derived from remote sensing information, for example, temperature, chlorophyll, or sand, and so on. However our STAMM modelling approach can handle other forms of spatial neighbourhood arrangements including points, cells, polygons and indeed 3-dimensional or multi-dimensional information.

There are two points which we draw attention to:

The first is how to define the neighbourhood. If the neighbourhood is defined too small (in relation to the influence of the surrounding spatial area on the focus) or some relationships are missing, this will result in the rule set not being complete, or that it cannot be successfully mined. But if the neighbourhood is too large or there are too many potential relationships (compared to the available information), long computations will be required and the resulting rule set may not be robust.

Similarly, in the definition of neighbourhood conditions, the attributes that should be selected into the table are a problem. If too many attributes are selected, that will make the computing time long. But if too few attributes are selected, that will result in the rule mined being incomplete, or the rule cannot be mined.

If the neighbourhood defined includes m cells and n attributes that are to be taken into account, then there are $mn + nC_m^2$ condition attributes in the decision table and it should search the table $\sum_{k=1}^{mn+nC_m^2} C_{mn+nC_m^2}^k$ times. If one cell is added into the neighbourhood and the number of attributes is n, the search time will be $\sum_{k=1}^{(m+1)n+nC_{m+1}^2} C_{(m+1)n+nC_{m+1}^2}^k$. If one more attribute is added into the table and the cell number in the neighbourhood is m, the search time will be $\sum_{k=1}^{m(n+1)+(n+1)C_m^2} C_{m(n+1)+(n+1)C_m^2}^k$. That is to say, if the cell number is 9 and the attribute number is 1, it should search the table 69×10^9 times. If the cell number is 10 and the attribute number is 1, it should search the table 36×10^{15} times. If the cell number is 9 and the attribute number is 2, it should search the table 1.2×10^{27} times. So caution must be exercised when we select the neighbourhood and the attributes. If we do not take the difference of values between two cells, the number of searches will decrease markedly.

Another problem is how to partition the values into intervals. If the interval is not partitioned properly such that the range of each interval is too short, too long, or the points divided are not the appropriate points, the rule mined will not be compete, or the rule cannot be mined.

There are two ways to resolve these two problems:

The best way is to obtain prior expert knowledge, even if it is qualitative or fuzzy, to determine the form of relationships that might be usefully examined. In complex problems, this knowledge is needed to guide the selection of an appropriate neighbourhood range and the necessary attributes, with their form of relationship with the focus. These condition attributes need to be encoded, or partitioned into suitable intervals, in order to simplify and guide the resulting computations. Incorporation of prior knowledge is thus a key element in the design and interpretation of the analyses.

If there is no prior background on the research problem, the solution has to depend on brute-force computing. So in defining the neighbourhood, or selecting the attributes, we need to define the neighbourhood large enough (to encompass conceivable interactions) and select enough attributes. Using our four previously defined performance criteria, the algorithm can choose those cells, and the attributes which are significantly related with the decision attribute in the focus cell. When partitioning the intervals, a few sets of intervals may need to be trialled. The mining alrogithm will need to be applied to each set of intervals, in order to search the rule sets with reference to the performance criteria and to select the best set of condition attributes and rules.

While the computer can mine the rules, it cannot provide understanding about these rules. So a key task after mining the rules, is to analyse the rationality of the selected condition attributes and rule sets and to iteratively select those which are meaningful.

5. Final remarks

With the development of technology, the difficulty to obtain data is decreasing and the difficulty to analyse the huge volume of data is increasing. That is to say, it is technically more difficult to extract knowledge from the huge volume of data which are currently being stored in databases. We have presented in this paper a data-mining approach that uses the theory of raster GIS to mine the relationships, or rules, between environmental attributes and the behavior of living organisms. Spatial neighbourhood is used to get the attributes of spatial structure and then a decision table is constructed from which the temporal attribute is analysed as a condition attribute. After recursion, ecological association rules are mined and the spatiotemporal assignments are extracted using a set of performance criteria developed here. Our example experiment showed that the STAMM model can successfully analyse the nonlinear rule-based relationship even when nonparametric correlations fail.

The characteristics of STAMM may be summarised as following:

(1) The continuos space and time of ecological process is discretised by the model.
(2) The uncertain or fuzzy prior-experience or knowledge can be reflected in the model.
(3) The prior knowledge will be refined as quantitative rules when data are mined by the model. This refined rule can be used in expert systems to predict the behavior, or spatio-temporal occupancy of living organisms.

(4) The model can deal with the nonlinear relationships between the environment pattern and the living organism.

The biggest merit of STAMM is that it can extract the spatiotemporal assignment which affects the behaviour or spatial occupancy of living organisms.

Acknowledgements

We wish to thank an anonymous referee and Professor S.E. Jorgensen for their comments on the manuscript. This research was partially funded by the projects 2003AA637030, 2002AA639400 and 2001AA633010, which are supported by the National Hi-technology Program of China. The authors also wish to thank Prof. Quanqin Shao and Professor Baoyin Liu (FIO, SOA, China) for their assistance during the research. We are grateful to Dr. Stephen K. Brown, Dr. Sebastien Brosse, Dr. We Ji and Dr. Steven Mackinson who provided some relevant references.

References

Agrawal, R., Imielinski, T., Swami, A., 1993. Mining association rules between sets of items in large databases. In: Proc. 1993 ACM-SIGMOD Int. Conf. Manage. Data, Washington, DC, pp. 207–216.

Bian, L., 2003. The representation of the environment in the context of individual-based modeling. Ecol. Model. 159, 279–296.

Binns, N.A., Eiserman, J.P., 1979. Quantification of fluvial trout habitat in Wyoming. Trans. Am. Fish. Soc. 198, 215–228.

Brosse, S., Guegan, J.F., Tourenq, J.N., Lek, S., 1999. The use of artificial neural networks to assess fish abundance and spatial occupancy in the littoral zone of a mesotrophic lake. Ecol. Model. 120, 299–311.

Brown, S.K., Buja, K.R., Jury, S.H., 2000. Habitat suitability index models for eight fish and invertebrate species in Casco and Sheepscot Bays, Maine. North Am. J. Fish. Manage. 20, 408–435.

Du, Y., Zhou, C., Shao, Q., 2001. Sea surface temperature and purse net productivity in East China Sea. High Technol. Lett. 11 (2), 56–60 (in Chinese).

FAO-CRODT-ORSTOM, 1995. A simulated GIS exercise to demonstrate its usefulness in the management of Senegalese demersal fisheries. Report of the training course on the application of GIS to fisheries. FAO project GCP/RAF/288/FRA, Rabat, Morocco.

Fayyad, U.M., Piatetsky-Shapiro, G., Smyth, P., 1996. From data mining to knowledge discovery: an overview. In: U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, R.U. Uthurusamy (Eds.), Advances in Knowledge Discovery and Data Mining. AAAI/MIT Press, pp. 83–115.

Gertseva, V.V., Gertsev, V.I., 2002. A model of fish population distribution in the space of inhabitation. Ecol. Model. 147, 161–170.

Giske, J., Huse, G., Fiksen, O., 1998. Modelling spatial dynamics of fish. Rev. Fish. Biol. Fish. 8, 57–91.

Hays, G.C., Dray, M., Quaife, T., 2001. Movements of migrating green turtles in relation to QVHRR derived sea surface temperature. Int. J. Remote Sens. 22 (8), 1403–1411.

Hu, J., 1995. Theory of Fishing Ground. China Agriculture Press, Beijing, pp. 156–158 (in Chinese).

James, F.C., McCulloch, C.E., 1990. Multivariate analysis in ecology and systematics: Panacea or Pandora's box? Ann. Rev. Ecol. Syst. 21, 129–166.

Ji, W., Jeske, C., 2000. Spatial modeling of the geographic distribution of wild life populations: a case study in the lower Mississippi River region. Ecol. Model. 132, 95–104.

Jorgensen, S.E., 1997. Ecological modelling by "Ecological Modelling". Ecol. Model. 100, 5–10.

Koperski, K., Han, J., 1995. Discovery of spatial association rules in geographic information databases. In: Advances in Spatial Databases, Proceedings of 4th Symposium, SSD'95. Springer-Verlag, Berlin, pp. 47–66.

Kracker, L.M., 1999. The geography of fish: the use of remote sensing and spatial analysis tools in fisheries research. Professional Geographer 51 (3), 440–450.

Letcher, B.H., Rice, J.A., 1997. Prey patchiness and laval fish growth and survival: inferences from an individual-based model. Ecol. Model. 95, 29–43.

Mackinson, S., 2000. An adaptive fuzzy expert system for prediction structure, dynamics and distribution of herring shoals. Ecol. Model. 126, 155–178.

Meaden, G.J., 2000. Applications of GIS to fisheries management. In: Wright, D., Bartlett, D. (Eds.), Marine and Coastal Geographic Information Systems. Taylor & Francis, London, UK, pp. 205–206.

Ricker, W.E., 1975. Computation and interpretation of biological statistics of fish populations. Bull. Fish. Res. Board Can. 191, 1–382.

Solanki, H.U., Dwivedi, R.M., Nayak, S.R., 2001. Synergistic analysis of SeaWiFS chlorophyll concentration and NOAA-AVHRR SST features for exploring marine living resources. Int. J. Remote Sens. 22, 3877–3882.

Somers, I.F., Long, B.G., 1994. Note on the sediments and hydrology of the Gulf of Carpentaria. Aust. J. Mar. Freshwater Res. 45, 283–291.

Su, F., Zhou, C., Shao, Q., 2000. Analysis of spatio-temporal fluctuations of East China Sea fishery resources using GIS. In: Rodriguez, G.R., Brebbia, C.A. (Eds.), Environmental Coastal Regions III. WIT Press, Southampton, pp. 249–257.

Su, F., Zhou, C., Shao, Q., 2001. GIS spatio-temporal analysis of fishery resources in East China Sea. High Technol. Lett. 11 (5), 60–63 (in Chinese).

Wei, C., 1988. Methods to predict the fishing situation for Spanish mackerel in the south of Yellow Sea. Acta Oceanol. Sinica 10 (2), 216–221 (in Chinese).