Journal of Zhejiang University SCIENCE
ISSN 1009-3095
http://www.zju.edu.cn/jzus
E-mail: jzus@zju.edu.cn

Automated soil resources mapping based on decision tree and Bayesian predictive modeling*

ZHOU Bin (周 斌)†, ZHANG Xin-gang (张新刚), WANG Ren-chao (王人潮)

(*Institute of Agricultural Remote Sensing and Information Technology Application, Zhejiang University, Hangzhou 310029, China*)

†E-mail: zhoubin@zju.edu.cn

Received Oct. 9, 2003; revision accepted Mar. 22, 2004

Abstract: This article presents two approaches for automated building of knowledge bases of soil resources mapping. These methods used decision tree and Bayesian predictive modeling, respectively to generate knowledge from training data. With these methods, building a knowledge base for automated soil mapping is easier than using the conventional knowledge acquisition approach. The knowledge bases built by these two methods were used by the knowledge classifier for soil type classification of the Longyou area, Zhejiang Province, China using TM bi-temporal imageries and GIS data. To evaluate the performance of the resultant knowledge bases, the classification results were compared to existing soil map based on field survey. The accuracy assessment and analysis of the resultant soil maps suggested that the knowledge bases built by these two methods were of good quality for mapping distribution model of soil classes over the study area.

Key words: Soil mapping, Decision tree, Bayesian predictive modeling, Knowledge-based classification, Rule extracting
Document code: A **CLC number:** S159-3; P283.8; P283.7

INTRODUCTION

Land resources management and ecological environmental decision-making requires knowledge about the spatial distribution, and quantity and quality of soil resources. Soil maps have traditionally been made by interpretation of remotely sensed imagery supported by ground surveys. Thus, soil mapping becomes expensive, labor-intensive, and time-consuming exercises. Moreover, it also is subjective, and may result in inconsistencies in the assignment of soil type boundaries or names between different interpreters, and over time with individual interpreters (Skidmore, 1989).

A number of studies used expert systems (also called knowledge-based systems) to do automatic soil mapping by reasoning like an expert (Huang and Jensen, 1997). Most expert systems are organized on three levels: data, knowledge base, and inference engine. The heart of the expert system approach is its knowledge base that contains a great deal of domain-specific knowledge (Luger and Stubblefield, 1993).

The usual method of acquiring knowledge in a computer-usable format to build a knowledge base involving human domain experts and knowledge engineers (Huang and Jensen, 1997). The domain expert explicitly expresses his or her knowledge about a subject in a language that can be understood by the knowledge engineer, who translates the domain knowledge into computer-usable format and stores it in the knowledge base.

This process presents a well-known problem

*Project supported by the National Natural Science Foundation of China (No. 40101014), and by the Science and technology Committee of Zhejiang Province (No. 001110445), China

when creating expert system that is often referred to as the "knowledge acquisition bottleneck". To solve this problem, much effort has been exerted in the artificial intelligence community to automate knowledge acquisition to obtain low-cast and high-quality knowledge bases. Studies on automated knowledge acquisition belong to the sub-field of artificial intelligence known as machine learning (Huang and Jensen, 1997).

Machine learning techniques for classification can be based on either statistical pattern recognition or data mining techniques for induction of decision trees or production rules. Thus, the objective of this study was to employ two techniques, namely decision tree and Bayesian predictive modeling methods, for automated soil mapping. This method eliminates or reduces the difficulty caused by the "knowledge acquisition bottleneck", and should allow expert system techniques to be adopted more easily by soil resources mapping scientists.

THEORETICAL BASIS FOR AUTOMATED SOIL MAPPING

Theoretical basis for soil inference

The theoretical basis for soil inference is based on the classic concept of Jenny (1941; 1980) that soil is a product of interaction among climatic factors, landform, parent material, organism, and hydrological factors over time. Therefore, we may infer the soil type at a given location if we have data on local environmental conditions. This can be expressed in qualitative terms by

$$S = f(Cl, Og, Pm, Tp, t) \quad (1)$$

where Cl represents climate conditions, Og is for organism, Pm is parent material, Tp stands for topography, and t is time.

Eq.(1) illustrates the general relationship between the soil and its environmental factors. However, the details of the relationship are different at different places. It is very difficult at this stage to derive a mathematical formula for the relationship because of the complexity and limited understanding of both soil forming processes and the paleo-environment.

Over decades of study of soil-environment relationships, a great deal of empirical knowledge has been accumulated. Particularly, local soil scientists who study and map soils in their respective regions have accumulated detailed knowledge on soil-environment relationships. It is our belief that this empirical knowledge can be used to approximate relationship in Eq.(1) for soil resources category inference.

Methodology of machine learning

Machine learning is the science of computer modeling of learning processes. It enables a computer to acquire knowledge from existing data or theories using certain inference strategies such as induction or deduction (Huang and Jensen, 1997). In this study, we focus on inductive learning and its application in building knowledge bases for automated soil mapping.

A motivation for the use of this approach to build a knowledge base is that it requires only a few good examples to function as training data. This is often much easier than explicitly extracting complete general theories from the domain expert. In machine learning from examples usually a set of examples is linked to a given output, the system then derives rules or statistical measures to link both parts. Both decision tree and Bayesian predictive modeling methods employed in this study need a few good training examples to derive the required knowledge.

Generally, ground sampling is the best way for obtaining such training data. However, it is difficult to collect enough samples especially in a non-agricultural area under financial and personnel constraints. So it would be profitable to extend the usefulness of existing soil surveys. During soil survey, empirical soil-environment models are developed to make predictive statements about the spatial distribution of soil classes. These models are developed by inductive reasoning from field observations and are used to delineate soil classes. Models for predicting soil-environment distribution relate soils/soil classes to topographic position in

certain landforms, geology, vegetation communities, and remotely sensed data. Thus, it should be theoretically possible to integrate these data types within a once mapped area to re-build predictive rules for soil resources mapping.

The association between an existing soil map for the Longyou area in Zhejiang Province, China, and other environmental and satellite sensing spatial data was analyzed by using two rule induction methods(decision tree method and Bayesian predictive modeling) whose resulting maps were then compared with the original. The underlying assumption was that these rule induction methods can reveal relationships between environmental variables that may help to predict the distribution of soil types and thus, in some way, mimic the mental model of the former surveyors.

STUDY AREA AND MATERIALS

Description of the study area

The study area (28°44′ to 29°17′N, 119°02′ to 119°20′E) locates in Longyou County, Zhejiang Province, China. According to the availability of the 1:50000 scale geological map, an area smaller than the administration region of Longyou County was selected as the study area of 777.8 km^2.

The climate is monsoon subtropical, with mean annual precipitation being about 1672 mm and the mean annual accumulation temperature reaching 5503 ºC. The area lies in the transitional zone between the North-Zhejiang Mountain range and the South-Zhejiang Mountain range with the alluvial plain formed by the Qujiang river running through the central part.

The elevation in the study area ranges from 33 m to 1439 m above sea level. The southern part is mountainous; the parent material is mainly gneiss, granite and tuff. The northern part is hilly region, mainly of red sandstone, limestone, and purple sandy shale parent material. The central part is the flat Jinqu Basin with variable fluvial deposits.

Based on the former soil survey (The Second Chinese National Soil Inventory), there are five soil groups in the study area, namely Red soil, Yellow soil, Lithomorph soil, Fluvio-aquic soil and Rice paddy soil, with red soil being the main zonal soil of the region.

Environmental variables used

The predictive variables considered important for determining the distribution of soils over the study area included: parent material, landuse, elevation and its derived terrain attributes, namely slope, aspect, profile curvature, plan curvature, upslope contributing area. In addition, the first 4 principal components of bitemporal TM imageries of study area are employed to characterize the soil-forming environment. The existing soil map used as a ground-truth map.

The 1:50000-scale soil map, geology and landuse map of the study area were digitized manually. Information on topographic data such as elevation, slope, aspect, profile curvature, plan curvature and upslope contributing area were obtained from a digital elevation model (DEM) of the study area. The contour lines, 3D points, streamlines and ridgelines were digitized manually from the Chinese State Bureau of Surveying and Mapping 1:50000-scale series map with a 10-m contour interval. These above topographic objects were input to Arc/Info to build a TIN, and then were transferred into a raster-based DEM. The profile curvature, plan curvature and upslope contributing area were derived by Arc/Info Grid module, and then rescaled into 0~255 to make these variables easily analyzable.

Two Landsat TM multispectral imageries of Longyou County obtained on 4th May, 2000 and 5th June, 1997 respectively were used in this study. After layer stacking these two TM imageries into a single imagery with 12 bands (excluding two thermal bands of TM), principal components analysis was performed on the bitemporal TM imagery, in order to reduce the number of features and thereby improve computational efficiency. Only the first 4 principal components accounting for 99.4% of the total scenes variance were adopted. These principal components were rescaled to range in brightness from 0 to 255 to improve the computational efficiency.

Layerstacking the dataset and sampling the training data

The raster data model was chosen to represent data layers and results in our method because the raster model is more suitable for representing continuous spatial variation of soil. All the data were geometrically corrected to a Gauss Kruger projection using the same map base (1:50000 scale contour map) and resampled to a regular 30 m grid. These datasets were inputted, and stored as a stacked layer in the ERDAS Imagine.

The objective of training is to provide examples of the concepts (in this study, it means soil-environment mental model) to be learned. When building a knowledge base for soil classification, the examples should be a set of training objects, each of which is represented by a class-attribute value class vector such as [*SoilClass_i*, *EnvironVariable*_1, …, *EnvironVariable_n*]. The learning algorithm attempts to deduce from this training data set some generalized concepts, i.e., rules that can be used to classify the remaining data (Huang and Jensen, 1997).

A total of 90266 pixels (approximately 10% of the total pixels) were selected using stratified random sampling. The training pixels were then subdivided randomly into two datasets; one consisting of 80786 pixels that were only used to develop the knowledge base and the other with 9480 pixels that were only used for accuracy assessment.

KNOWLEDGE ACQUISITION PROCESS

Decision tree modeling

(1) A brief introduction of decision tree modeling

Decision tree is a non-parametric method for analyzing hierarchical relationships. Trees can identify and express nonlinear and non-additive relationships in a simple form. The idea behind decision tree is to recursively subdivide the training set of examples into homogeneous groups, using discriminating variables. The variable selection criterion is based on the entropy measure from information theory. Thus a variable is chosen, which results in the best discrimination of the dataset into the given classes. The procedure is repeated for each new subset of the dataset, until all data in a subset belong to as "pure" subsets as possible.

There are a number of decision tree algorithms, such as Quinlan's (1986; 1993) ID3 and C4.5, and so on. The C4.5 algorithm was selected in this research because of its following advantages (Quinlan, 1993): 1) C4.5 is flexible. Unlike many statistical approaches, it does not depend on assumptions about the distribution of variables values or the independence of the variables themselves; 2) C4.5 is based on a decision-tree learning algorithm that is one of the most efficient forms of inductive learning.

(2) decision tree generation

A recursive "divide and conquer" strategy is used by C4.5 to generate a decision tree from the above training data. The input training data was a text file with each line representing a training object. In this study, tree construction proceeds until the number of cases reaching each leaf is small (by default, n<5) or the leaf is homogeneous enough (by default, its deviance is <1% of the deviance at the root node). The resultant tree structure is shown in Fig.1.

The decision tree obtained using the continuous quantitative variables elevation, slope, aspect,

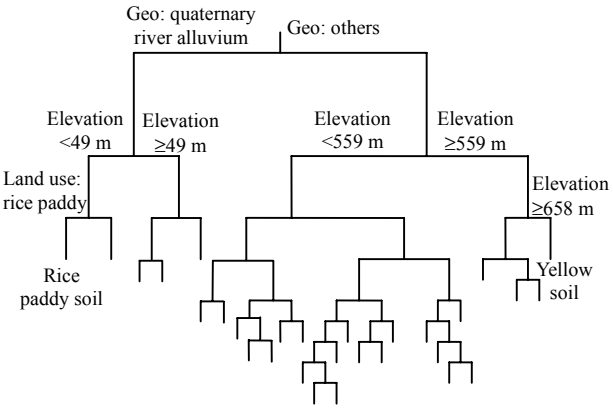

Fig.1 The structure of the classification tree with 29 terminal nodes

Some simplified splitting criterion is given for the marginal nodes with the largest reduction in deviance. All of the splitting criterion (i.e. production rules) are listed in Appendix I

upslope contribution area, PC1−PC4 and the categorical factor geology, landuse to predict Longyou soil classes had 89 nodes. Table 1 compares the original and pruned tree statistics. The misclassification error rate increases slightly from 0.1891 to 0.1894 for the pruned tree with 29 terminal nodes. Yet more severe pruning further reduced the number of predicted classes: pruning to 20 or 11 terminal nodes yielded 4 classes (Table1).

It was clear that geology, landuse and elevation, which were used to split the trees at the highest nodes (Fig.1) were the most important variables for predicting Longyou soil classes. Of the DEM-derived terrain attributes, only slope and upslope contributing area were used to build the tree when it was pruned to 29 terminal nodes.

(3) From decision tree to production rules

Although decision tree is an important form of knowledge representation, it is rarely used directly in knowledge bases. Decision trees are often too complex to be understood, especially when they are large. A decision tree is also difficult to maintain and update. Therefore, it is often desirable to transform a decision tree to another form of knowledge representation, such as production rules. In fact, each path from the root to a leaf in a decision tree can be translated into a production rule. For example, the path from the root to the most left leaf in the decision tree in Fig.1 can be represented by a production rule, i.e.:

(geology="Quaternary river alluvium") and (elevation<49m) and (landuse="rice paddy field")→ (soil=Rice paddy soil)

The performance of each rule was summarized by the statistics (n/m). n is the number of training cases covered by the rule and m shows how many of them do not belong to the class predicted by the rule. The rule's accuracy was estimated by the Laplace ratio ($n-m+1$)/(n+2). The meaning of rule parts was shown in the following simple illustration:

```
RULE 2:
GEO=Quaternary river alluvium
   // if Geology type is "Quaternary river alluvium",
ELE ≤ 49 m
   //and if elevation is less than and equal to 49 m,
LDU = rice paddy field
   //and if landuse type is "rice paddy field",
→ SOIL = Rice paddy soil  [0.897]
   //then soil class is "Rice paddy soil".
   //The accuracy of this rule is 0.897.
```

Totally, 29 predicting rules were transformed from the decision tree (see Appendix I).

Because the rules are easy to understand, they can also be examined by human experts. With caution, they may be edited directly. Based on the result of the decision tree, the production rules were built as a file with the aid of ERDAS IMAGINE 8.4/ knowledge Engineer program. This file became the knowledge base and was the core part of the knowledge classifier. The subsequent knowledge classification was preformed by using the ERDAS IMAGINE 8.4/knowledge classifier program, and the resultant soil map is shown in Fig.2.

Bayesian predictive modeling

(1) A brief introduction to Bayesian predictive modeling

Bayesian statistics constitute an alternative method for building predictive relationships between

Table 1 Summary of original and pruned tree results

Variables employed by decision tree (ordered by appearance in tree)	No. of terminal nodes	Misclassication error rate	No. of soil classes predicted
GEO+ELE+LDU+PC3+PC4+PC1+ASP+ USCA + SLP + PC2	89	0.1891	5
GEO+ELE+LDU+PC3+PC +SLP+USCA+PC2	33	0.1892	5
GEO+ELE+LDU+PC3+SLP+USCA+PC2+PC4	29	0.1894	5
GEO+ELE+LDU+PC3+SLP+USCA+PC2	20	0.1917	4
GEO+ELE+LDU+PC3+SLP+PC2	11	0.2001	4

GEO: geology; ELE: elevation; LDU: landuse; ASP: aspect; USCA: upslope contribution area; SLP: slope; PC1, PC2, PC3 and PC4: 1st, 2nd, 3rd and 4th principal components of bitemperal TM data, respectively

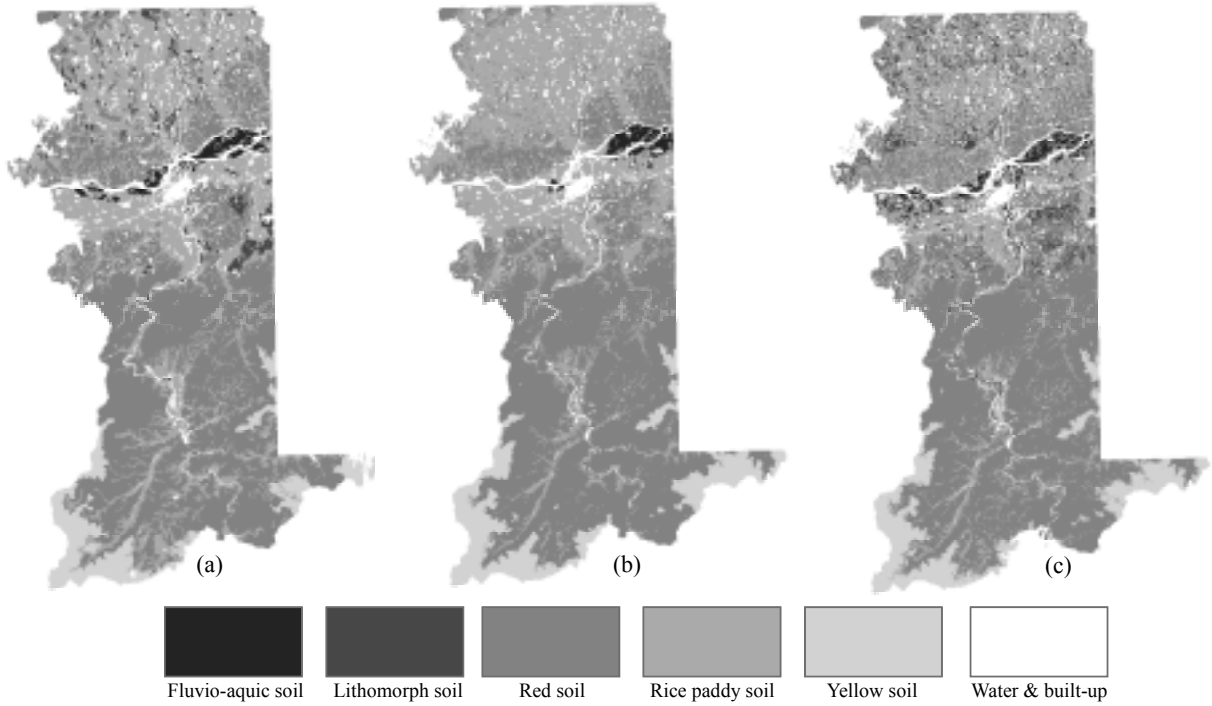

Fig.2 The soil maps of the study area
(a) Reference soil map; (b) Inferred soil map based on classification tree;
(c) harden soil map resulting from Bayesian predictive modeling

soil types and their environment. Mapping soil resources under this probability model is a process of deriving probability values of the given soil resource at each location (pixel) to a set of prescribed soil resource classes.

Bayes' theorem uses a prior and conditional probabilities to calculate the probability of an uncertain event occurring. In this instance, let S_a be the soil class (for a=1, …, n classes) occurring at location X_{ij}, that is at the ith row and jth column of the GIS raster database. Let E_b be an item of predictor variable (for b=1, …, k items of evidence) known at location X_{ij}. Set up a hypothesis (H_a) that class S_a occurs at location X_{ij},

A rule may be defined as

$$E_b \Rightarrow H_a \qquad (2)$$

That is, given a piece of evidence, then infer H_a.

The probability of a given soil class at a given cell is then inferred using Bayes' theorem to update the probability of the rule that the hypothesis (H_a) occurs at location (i, j) given a piece of evidence (E_b), i.e.

$$P(H_a \mid E_b) = [P(E_b \mid H_a)P(H_a)] / P(E_b) \qquad (3)$$

where $P(E_b \mid H_a)$ is the a priori conditional probability that there is a piece of evidence E_b given a hypothesis H_a that class S_a occurs at location (i, j); $P(H_a)$ is the probability for the hypothesis (H_a) that class S_a occurs at location (i, j) and is estimated by the experienced soil scientist or calculated on the field samples. $P(E_b)$ is the probability of the evidence alone, or, the probability that any cell has an item of evidence $\{E_b\}$.

Given the complexity of soil landscape, we would not expect a high degree of precision from one or two rules alone. The surveyor would conventionally consider a range of evidence simultaneously and a similar function is performed here by combing rules in a single estimate of probability.

Thus parallel inference method was employed in the inference process to calculate the joint probability of the soil classes' occurrence given *n* soil environment factors.

Assuming conditional independence of the evidence, the joint probability from *n* contributing rules can be calculated as follows (Cook *et al.*, 1996):

$$p(H_a \mid E_1,...,E_b) = \frac{o(H_a \mid E_1,...,E_b)}{o(H_a \mid E_1,...,E_b)+1} \tag{4}$$

$$o(H_a \mid E_1,...,E_b) = o(H_a)\prod_{b=1}^{k} L_b \tag{5}$$

$$L_b = \frac{o(H_a \mid E_b)}{o(H_a)} \tag{6}$$

$$o(H_a \mid E_b) = \frac{p(H_a \mid E_b)}{1-p(H_a \mid E_b)} \tag{7}$$

$$o(H_a) = \frac{p(H_a)}{1-p(H_a)} \tag{8}$$

where $p(H_a \mid E_1,...,E_b)$ is the joint probability that a hypothesis H_a that class S_a occurs at location (i, j) given n evidence E_b; L_b is the weighting value of the bth environmental variable for the soil S_a; $o(H_a \mid E_1,...,E_b)$, $o(H_a \mid E_b)$ and $o(H_a)$ represent the conditional odds format of $p(H_a|E_1, \ldots, E_b)$, $p(H_a \mid E_b)$ and $P(H_a)$, respectively.

(2) Bayesian probability modeling

In a soil mapping exercise the joint probabilities are generally developed using sample sites throughout the area to be mapped. At these sites the soil class to be predicted are identified, as are the on-site values of the evidence variables. In this study, the joint probabilities were calculated using the training data mentioned above.

Based on the above result of classification tree, the evidence variables that might be helpful in identifying the soil type of the study area were geology, landuse, elevation, slope, upslope contribution area, and PC1~PC4. Thus only these 9 variables were adopted in the Bayesian predictive modeling. For each soil type, a point-by-point comparison with each predictor variable was used to derive a probability lookup table.

The inference process was carried out using a raster data model with which probability values were computed for each grid cell. For a given soil type the inference system took a set of environmental conditions of a pixel from the GIS database. It then used each of the probability lookup tables to calculate the joint probability value from each of the environmental variables. The combining inference was used on these probability values to obtain the joint probability value for the pixel. The process continued onto the next pixel until all pixels in the area were visited. A map (P^a) of overall probability values for the given soil type *a* over the mapping area was then produced. The process continued onto the next soil type until all soil types were considered.

Thus, the soil in a given location (i, j) can be expressed by an n-element vector

$$\boldsymbol{P}_{ij} = (p_{ij}^1,...p_{ij}^a,...p_{ij}^n) \tag{9}$$

where p_{ij}^a is a probability measure of the soil at point (i, j) to the prescribed soil categorical unit a, and n is the number of soil categorical units in the area.

We call vector $\boldsymbol{P}$ the soil probability vector (SPV). Thus, SPV at point (i, j) will be represented as $\boldsymbol{SPV}_{ij}$. The assemblage of these probability maps (SPV) forms the probability representation of the soil resources over the area.

In order to compare the inferred results with the existing soil maps at the point level, we hardened the SPV to produce a crisp representation of soil information for each pixel by assigning to it the soil type which has the highest posterior probability $\{\max\ p(H_a \mid E_1,...,E_b) \mid a=1,\ldots,n\}$ at that location. The resultant soil maps obtained by using Bayesian predictive modeling are shown in Fig.3.

RESULTS AND DISCUSSIONS

Maps results from each of the methods detailed

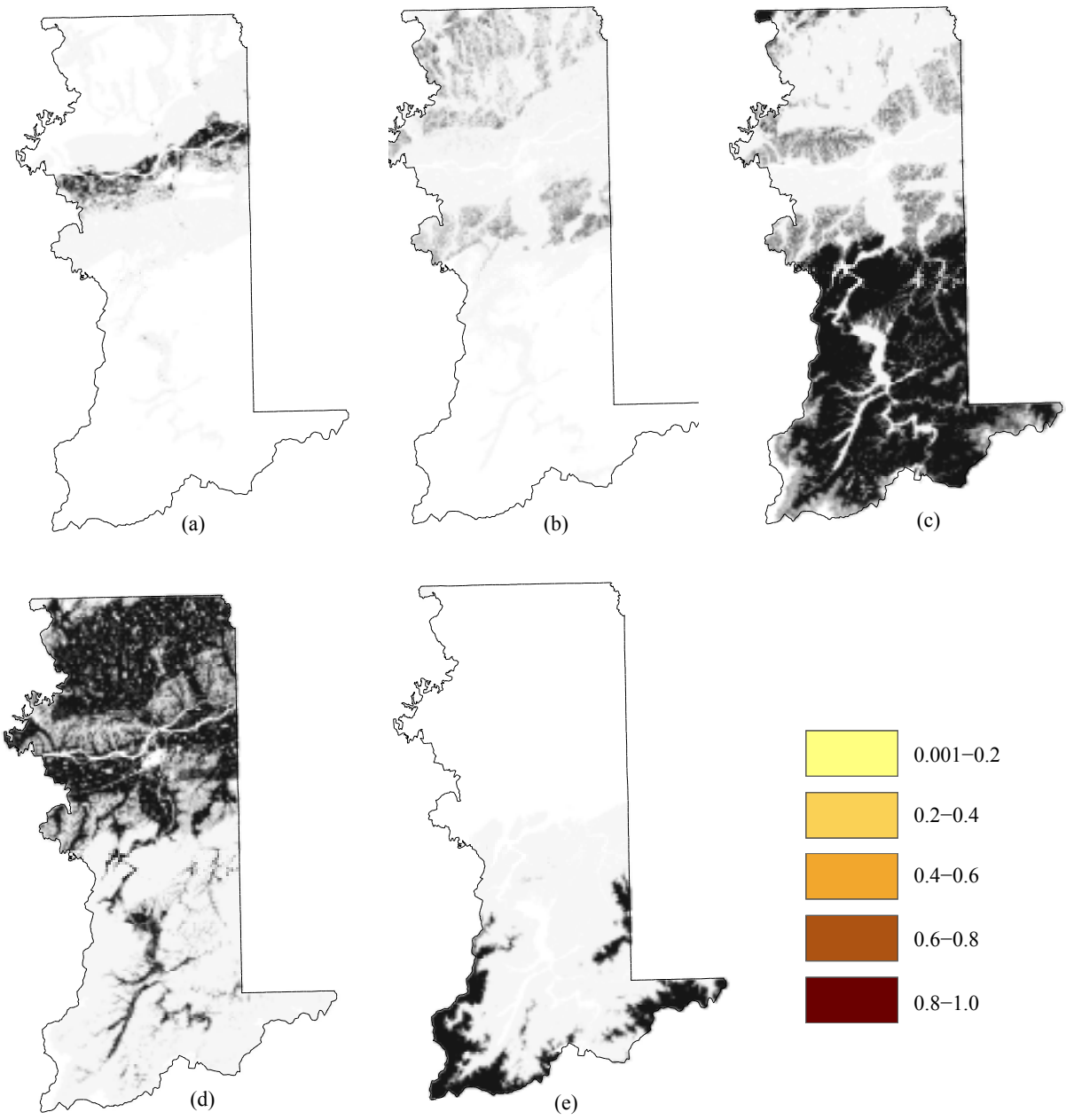

Fig.3 The probability maps of the soil in study area
(a) Red soil; (b) Yellow soil; (c) Lithomorph soil; (d) Rice paddy soil; (e) Fluvio-aquic soil

above are shown in Fig.2 and Fig.3. The resulting maps were overlaid with the reference and the correspondences between them were tested by cross-validation (Table 2 and Table 3). User's accuracy, producer's accuracy and Kappa statistics were then applied to measure the overall agreement between the predicted and the reference data. These misclassification rates suggested that there were some differences between the inferred soil maps by two methods and the existing soil map.

For both methods, Yellow soil and Fluvio-aquic soil were predicted more successfully than Rice paddy soil and Red soil according to Kappa statistics. Neither method can successfully predict

Table 2 Cross-tabulation of reference vs classes predicted with the decision tree

		Reference soil map						User's accuracy	Kappa
		Red soil	Yellow soil	Lithomorph soil	Fluvio-aquic soil	Rice paddy soil	Total		
Predicted soil map	Red soil	4088	50	90	1	795	5024	0.814	0.6307
	Yellow soil	172	581	0	0	30	783	0.742	0.7236
	Lithomorph soil	4	0	7	0	0	11	0.636	0.6273
	Fluvio-aquic soil	0	0	0	104	12	116	0.897	0.8948
	Rice paddy soil	433	0	134	56	2923	3546	0.824	0.7088
	Total	4697	631	231	161	3760	9480	—	—
Producer's accuracy		0.870	0.921	0.030	0.646	0.777	Overall accuracy: 0.813		

Table 3 Cross-tabulation of reference vs classes predicted by using Bayesian predictive modeling

		Reference soil map						User's accuracy	Kappa
		Red soil	Yellow soil	Lithomorph soil	Fluvio-aquic soil	Rice paddy soil	Total		
Predicted soil map	Red soil	3828	125	57	2	601	4613	0.830	0.6627
	Yellow soil	85	453	0	0	5	543	0.834	0.8224
	Lithomorph soil	125	0	67	0	77	269	0.249	0.2303
	Fluvio-aquic soil	4	0	0	136	42	182	0.747	0.7429
	Rice paddy soil	655	53	107	23	3035	3873	0.784	0.6414
	Total	4697	631	231	161	3760	9480	—	—
Producer's accuracy		0.815	0.718	0.290	0.845	0.807	Overall accuracy: 0.793		

Lithomorph soil, although the Kappa value for it in Table 2 reached to 0.6273, the producer's accuracy for this class was much worse (0.030) which suggested almost 97% of this type was classified as other types.

Both of the two methods predicted Red soil with similar accuracy, Kappa values for decision tree and Bayesian predicting modeling were 0.6307 and 0.6627, respectively. User's accuracy and producer's accuracy were also close to each other.

Rice paddy soil was in similar situation with red soil. With some slight difference, decision tree and Bayesian predicting modeling displayed the comparable effect for soil prediction.

However, great differences existed in the prediction of Fluvio-aquic soil and Yellow soil by the two methods. For Fluvio-aquic soil, producer's accuracy of Bayesian predicting modeling was better than that of decision tree by 20%, whereas user's accuracy of the former was worse than the latter by 15%. For Yellow soil, Bayesian predicting modeling had better user's accuracy but worse producer's accuracy than decision tree modeling with differences of accuracy values reaching to 10% and 20%, respectively.

Overall accuracy provided another index for evaluating the prediction result. About 81.3% and 79.3% of the total validating data were correctly predicted by decision tree modeling and Bayesian predicting modeling. This result suggested that, on general basis, these two methods had equivalent effect on soil class prediction over the study area although they displayed different effect for some of the specific soil type predictions.

The differences between the inferred soil maps and the reference soil map might be caused by inaccuracies in the GIS data layers and the ambiguity or incompleteness of the knowledge bases. Errors in GIS data layers will reduce the mapping accuracy of the resultant soil maps derived by the machine learning techniques (Skidmore *et al*., 1991). The extent to which errors were accumulated from the

GIS data layers may be uncertain, but are analyzed theoretically and empirically in the following part.

In this study, the existing soil map produced from soil surveys was digitized to be used as a "true" map and provide the soil classes information required for the machine learning. However, soil maps often contain a great deal of uncertainty as much of the quantitative and qualitative knowledge of the soil scientists regarding the occurrence of given soil units is not maintained in the map. There are some major problems regarding the use of current soil maps in geographic analysis (Burrough, 1986) including limited coverage at a fixed scale, locational errors, attributes errors, and insufficient information in the mapping units due to the crisp logic and cartographic techniques with which soil maps are produced.

Therefore, the knowledge of soil scientists about soil variation cannot be fully represented by soil maps so that some of the information mapped cannot be captured by the machine learning methods. The knowledge bases acquired by machine learning algorithm can be consequently ambiguous or incomplete.

In addition, locational errors are introduced into soil maps by improper positioning of boundaries between soil bodies. The introduction of locational errors is not due purely to the mistakes made by soil mapping experts but also due to the nature of soil boundaries. Soil varies gradually and the boundaries between different types of soil are often diffused than sharp (Mark and Csillag, 1990). However, soils have to be delineated into homogeneous polygons on soil maps. Therefore, it is difficult for any soil-mapping expert to draw a boundary between two soils without introducing locational errors.

The idea that the low predictive success of the two methods stems from positional inaccuracies in the polygon boundaries was tested by performing new map correlations after masking out 1 pixel-width on either side of map polygon boundaries of the original reference soil map. This decreased the number of validation samples from 9480 to only 7186. The overall accuracy of the predicted map by using the decision tree increased from 81.3% to 89.0% and Kappa statistics increased from 0.6307 to 0.8029 for Red soil, from 0.7088 to 0.8371 for Rice paddy soil. The overall accuracy of the predicted map by using the Bayesian probability method increased from 79.3% to 86.5% and Kappa statistics increased from 0.6627 to 0.8114 for Red soil, from 0.6414 to 0.7701 for Rice paddy soil.

Table 4 lists the details of the decrease number for each soil classes. The results showed that the validating data for Lithomorph soil had the highest decrease rate, more than half of the total points located in the above masking-out region which implied Lithomorph soil dramatically tend to be spatially intermittent.

Based on the results, we can find that both methods had difficulty in singling out the soil-environmental relationships for soil units that tend to be spatially intermittent, especially in the hilly region of the study area. It is true that at high elevations the environment is less heterogeneous than at low elevations in the study area. Thus, the soil class (Yellow soil) at high elevations tended to be spatially contiguous always had the higher classification accuracy. Understanding the relationships between spatially contiguous soil units and their environments would be easier than under-

Table 4 Details of the decrease number for each soil classes

	No. of original validation data	No. of validation data after masking-out	No. of decrease	Rate of decrease
Red soil	4697	3673	1024	21.8%
Yellow soil	631	536	95	15.1%
Lithomorph soil	231	112	119	51.5%
Fluvio-aquic soil	161	123	38	23.6%
Rice paddy soil	3760	2742	1018	27.1%
Total	9480	7186	2294	24.2%

standing the relationships between intermittently distributed soil units and their environments.

The reference soil map is not completely accurate, neither is the other categorical maps (geology and landuse) nor the DEM, and further errors are propagated by using the DEM to derive terrain attributes. Lithomorph soil often occurs around small rock outcrops or along small spurious ridges and divides which could not be described well with the environmental variables, especially the topographic and parent materials variables employed. Therefore, it was difficult to establish the relationships between Lithomorph soil and these environmental variables. In detail, Lithomorph soil and Red soil essentially fall within similar elevation range, and have similar spectral reflectance due to their being overlaid by similar landcover types (mainly rangeland) in the northern part of the study area, where only the parent materials can be used to separate them. Thus prediction of these two soil classes is highly sensitive to any errors in the geology map. The low consistency measure between our resultant map and the existing soil map related to the Lithomorph soil implied that there might be some errors or omissions existing in the geology map. A larger scale (for example, 1:10000 scale) geology map would be helpful for improving the accuracy.

Yet another possibility for the inconsistency rate of the machine learning methods is that the current environmental conditions are not those under which the soils developed. For example, we can easily be aware that the landuse type of some areas that underlaid the rice paddy soil in the reference soil map has changed from rice paddy field to other landuse type (rangeland or dry land) by overlaying landuse map or remotely sensed imageries. Thus, the soil type must have changed to some extent due to absence of the conditions of Rice paddy soil development.

Therefore, the valid test of accuracy for the results is field ground-truthing of the predicted maps but not only by comparing with the existing soil map. Under financial and personnel constraints, however, it is generally difficult to collect enough field samples to assess the error in a statistically significant manner.

The trees confirm that the spatial correlations evident from visual comparisons of overlays of the geology, landuse, spectral features and elevation with the Longyou soil map. In fact, Wang *et al*. (1986) considered landform as the highest category level of their soil visual interpretation. The soils on uniform landform were further subdivided on the basis of parent material and spectral reflectance. These researches including this study suggested that landform, geology, and multispectral imagery played important roles in the mental model used by soil experts to map the soil distribution in Zhejiang Province.

As the terrain attributes derived from DEM, it is possible that profile curvature and plan curvature are not useful for predicting soil-environment relationships when the soil mapping units are large groups rather than genus of species and there is a mismatch between the scale of generalization of soil units and the terrain attributes. Lagacherie and Holmes (1997) concluded that the curvature variables were of limited usefulness in his study area in France.

Bayesian probability methods achieved moderate success in predicting soil map classes over the study area using geology, bitemporal TM imageries, landuse, DEM and its derived terrain attributes. Moreover, this representation of soil information is different from the conventional crisp representation. The existence of the soil at a location in a soil unit is expressed in terms of a probability value between 0.0 and 1.0, and not a yes or no. Fig.3b and Fig.2a show the distribution of probability values and existing soil map, respectively, for Yellow soil. Compared with the existing soil map, the probability maps reveal more details at the spatial level. Therefore, the probability predictive modeling method is capable of eliminating the minimum mapping size problem in conventional soil mapping and by allowing more detailed spatial patterns of soil information to be represented.

CONCLUSION

This paper presents a process based on deci-

sion tree technique and Bayes' Theorem for acquiring knowledge about soil-environment relationships and also presents a case study that demonstrated the use and potential of this process for extracting useful information from the existing maps. The case study showed that the process was moderately successful in extracting knowledge from the existing soil maps and other GIS data related to soil development. The implication is that this process would be useful in acquiring knowledge about resource-environment relationships for mapping natural resources by using the decision tree and probability model.

The decision trees suggested that there exist spatial correlations between some local environmental variables and the Longyou soil types. These variables were confirmed to be geology, landuse, spectral features, elevation and some of its derived terrain attributes, with geology, landuse, and elevation playing the most important roles in the mental model used by soil experts to map the soil distribution in local area.

The probability images produced from the Bayesian predictive modeling have potential advantages over standard soil survey maps in terms of revealing spatial patterns of soil information and in terms of production cost. Rigorous field-testing is required to quantify the potential advantages of this technique in the derivation and representation of the spatial pattern of soil types.

References

Burrough, P.A., 1986. Principles of Geographical Information Systems for Land Resources Assessment. Clarendon Press, Oxford, p.193.

Cook, S.E., Corner, R.J., Grealish, G.J., Gessler, P.E., Chartres, C.J., 1996. A rule-based system to map soil properties. *Soil Science Society America Journal*, **60**:1893-1900.

Huang, X.Q., Jensen, J.R., 1997. A machine-learning approach to automated knowledge-base building for remote sensing image analysis with GIS data. *Photogrammetric Engineering & Remote Sensing*, **63**(10): 1185-1194.

Jenny, H., 1941. Factors of Soil Formation: A System of Quantitative Pedology. McGraw-Hill, New York, p.281.

Jenny, H., 1980. The Soil Resource: Origin and Behaviour. Springer-Verlag, New York, p.377.

Lagacherie, P., Holmes, S., 1997. Addressing geographical data errors in a classification tree for soil unit predictions. *Int. J. Geographical Information Science*, (11):183-198.

Luger, G.F., Stubblefield, W.A., 1993. Artificial Intelligence: Structures and Strategies for Complex Problem Solving. Second Edition, The Benjamin/Cummings Publishing Company, Inc., Redwood City, California, p.740.

Mark, D.M., Csillag, F., 1990. The nature of boundaries on 'area-class' maps. *Cartographica*, (27):65-78.

Quinlan, J.R., 1986. Induction of decision tree. *Machine Learning*, **1**(1):81-106.

Quinlan, J.R., 1993. C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, San Mateo, California.

Skidmore, A.K., 1989. An expert system classifies eucalypt forest types using thematic mapper data and a digital terrain model. *Photogrammetric Engineering and Remote Sensing*, **55**(10):1449-1464.

Skidmore, A.K., Ryan, P.J., Dawes, W., Short, D., O'Loughlin, E., 1991. Use of an expert system to map forest soils from a geographical information system. *Int. J. Geographical Information Systems*, **5**(4):431-445.

Wang, R.C., Wang, S.F., Su, H.P., 1986. The research on soil visual interpretation and mapping technique by using MSS imagery. *Journal of Zhejiang Agricultural University*, **12(**2):103-111 (in Chinese).

Appendix I 29 predicting rules transformed from the decision tree*

Rule 1:
ELE < 49 m
GEO = 6,9,17,26[1)]
LDU = 4,7,10,16,17,18[2)]
→ Soil = Fluvio-aquic soil [0.878]

Rule 2:
ELE < 49 m
GEO = 6,9,17,26
LDU = 1,2,3,5,8,9,15,20
→ Soil = Rice paddy soil [0.897]

Rule 3:
ELE >= 49 m
GEO = 6,17
LDU = 1,2,4,7, 9,11,15,16
→ Soil = Rice paddy soil [0.780]

Rule 4:
ELE >= 49 m
GEO = 6,17
LDU = 3,6,8,10,13,18,19,20
→ Soil = Red soil [0.599]

Rule 5:
ELE >= 49 m
GEO = 9,26
→ Soil = Rice paddy soil [0.915]

Rule 6:
ELE < 559 m
GEO !=[3] 6,9,17,26
LDU = 1,2,6
PC3 < 98
SLP < 18
→ Soil = Red soil [0.670]

Rule 7:
ELE < 559 m
GEO != 6,9,17,26
LDU = 1,2,6
PC3 < 98
SLP >= 18
→ Soil = Red soil [0.716]

Rule 8:
ELE < 559 m
GEO != 6,7,9,12,16,17,23,26,27
LDU = 1,2,6
PC3 >= 98
USCA < 3
→ Soil = Red soil [0.649]

Rule 9:
ELE < 559 m
GEO != 6,7,9,12,16,17,23,26,27
LDU = 1,2,6
PC3 >= 98
USCA >= 3
PC2 < 126
→ Soil = Rice paddy soil [0.545]

Rule 10:
ELE < 559 m
GEO != 6,7,9,12,16,17,23,26,27
LDU = 1,2,6
PC3 >= 98
USCA >= 3
PC2 >= 126
→ Soil = Red soil [0.664]

Rule 11:
ELE < 559 m
GEO = 12,16,23,27
LDU = 1,2,6
PC3 >= 98 AND < 115
→ Soil = Rice paddy soil [0.538]

Rule 12:
ELE < 559 m
GEO = 12,16,23,27
LDU = 1,2,6
PC3 >= 115
→ SOIL = Red Soil [0.610]

Rule 13:
ELE < 73 m
GEO = 12,13,15,16,18,19,22,23,27
LDU != 1,2,6
PC3 < 111
PC4 < 116
→ Soil = Red soil [0.702]

Rule 14:
ELE < 73 m
GEO = 12,13,15,16,18,19,22,23,27
LDU != 1,2,6
PC3 < 111
PC4 < 95
→ Soil = Rice paddy soil [0.544]

Rule 15:
ELE < 73 m
GEO = 12,13,15,16,18,19,22,23,27
LDU != 1,2,6
PC3 < 111
PC4 >= 95 AND <116
→ Soil = Rice paddy soil [0.597]

Rule 16:
ELE < 73 m
GEO = 12,13,15,16,18,19,22,23,27
LDU != 1,2,6
PC3 < 111
PC4 >= 116
→ Soil = Red soil [0.746]

Rule 17:
ELE < 73 m
GEO = 12,13,15,16,18,19,22,23,27
LDU != 1,2,6
PC3 >= 111
→ SOIL = Red soil [0.824]

Rule 18:
ELE >= 73 m and < 559 m
GEO = 12,13,15,16,18,19,22,23,27
LDU = 13
→ Soil = Lithomorph soil [0.638]

Rule 19:
ELE >= 73 m and < 559 m
GEO = 12,13,15,16,18,19,22,23,27
LDU = 4,7,8,9,10,14,15,16,18,20
→ Soil = Red soil [0.713]

Rule 20:
ELE >= 73 m and < 559 m
GEO = 12,13,15,16,18,19,22,23,27
LDU != 1,2,6
→ Soil = Red soil [0.706]

Rule 21:
ELE < 559 m
GEO = 4,8,10,14,20,21,24,25
LDU != 1,2,6
PC2 < 98
→ Soil = Red soil [0.825]

Rule 22:
ELE < 559 m
GEO = 1,2,3,4,5,11
LDU != 1,2,6
PC2 < 23
→ Soil = Red soil [0.492]

Rule 23:
ELE < 559 m
GEO = 1,2,3,4,5,11
LDU != 1,2,6
PC2 >= 23 and <83
→ Soil = Rice paddy soil [0.683]

Rule 24:
ELE < 559 m
GEO = 1,2,3,4,5,11
LDU != 1,2,6
PC2 >= 83 and <98
→ Soil = Red soil [0.524]

Rule 25:
ELE < 559 m
GEO = 1,2,3,4,5,11
LDU != 1,2,6
PC2 >= 98
→ Soil = Red soil [0.900]

Rule 26:
ELE >= 559 m and < 658 m
GEO = 6,9,17,26
PC2 < 104
→ Soil = Rice paddy soil [0.542]

Rule 27:
ELE >= 559 m and < 602 m
GEO != 6,9,17,26
PC2 >= 104
→ Soil = Red soil [0.683]

Rule 28:
ELE >= 602 m and < 658 m
GEO = 6,9,17,26
PC2 >= 104
→ Soil = Yellow soil [0.565]

Rule 29:
ELE >= 658 m
GEO != 6,9,17,26
→ Soil = Yellow soil [0.834]

* GEO: geology; ELE: elevation; LDU: landuse; ASP: aspect; USCA: upslope contribution area; SLP: slope; PC1, PC2, PC3 and PC4: 1st, 2nd, 3rd and 4th principal components of bitemperal TM data, respectively
1) Geology codes see Appendix II; 2) Landuse codes see Appendix III; 3) != means "not equal to"

Appendix II Geology codes derived from the 1:50000 scale geological map of the study area

Code	Geological type	Code	Geological type
1	Granite, J_3	15	Silty mud stone, K_1c
2	Granite, J_3Sc	16	Conglomerate, K_1z
3	Felsite, $\nu o\pi J_3$	17	Silty mud stone, K_2j
4	Rhyolite, J_3j^3	18	Gritstone, K_2q
5	Arkose quartzite, K_1	19	Silty mud stone, O_1
6	Diorite	20	Silty mud stone, Z_1x
7	Andesite	21	Tuff, J_3
8	Diabase	22	Siliceous shale, $\in_1 h$
9	Ultrabasic rock	23	Marlite, $\in_3$-O_1x
10	Peridotite, Pt	24	Quartz sandstone, $P_{1\text{-}2}l$
11	Gneiss, Pt_1	25	Quartz sandstone, T_{3W}
12	Silty mud stone, J_3l^1	26	River alluvium, Qhy^{al}
13	Conglomerate, J_3l^2	27	River alluvium, Qp_2
14	Tuff, J_3x^1		

Appendix III Landuse codes derived from the 1:50000 scale landuse map of the study area

Code	Landuse type	Code	Landuse type
1	Irrigated rice paddy field	11	Bamboo
2	Rice paddy field without irrigation	12	Shrubbery
3	Irrigable land	13	Sparse woodland
4	Dryland	14	Afforestation land
5	Vegetable land	15	Built-up
6	Other gardens	16	Water
7	Mulberry field	17	Overbank flood plain
8	Tea garden	18	Grassland
9	Citrus garden	19	Bare rock
10	Woodland	20	Other unused land