

Application of Artificial Intelligence to Planning Sustainable Construction: an Ecological Quality Assessment of Soil

Kangsuk Kim¹, Dongwon Ki¹, Ilsuh Son², Kyongjoo Oh², Joonhong Park^{1,*}

¹Department of Civil and Environmental Engineering, Yonsei University, Seoul,
Republic of Korea, parkj@yonsei.ac.kr

²Department of Information and Industrial Engineering, Yonsei University, Seoul,
Republic of Korea, johanoh@yonsei.ac.kr

ABSTRACT: Accurately assessing environmental impact is important for planning sustainable mega-construction projects. Soil ecology plays a significant role in global ecosystems. However, methods for classifying and assessing soil ecological quality have not been well established. To enable such an evaluation, mapping quantitative information on soil ecological quality is required. Thus, the present study was undertaken in order to develop a predictive model for soil ecological quality using an Artificial Intelligence (AI) algorithm and to evaluate the applicability of soil ecological quality maps for planning eco-friendly road construction. Soil microbial diversity as an indicator of soil ecological quality was measured in 196 soil samples in a forest region. A Decision Tree (DT) model was optimized by correlating the field-measured microbial diversity data with existing soil property data (7 soil properties from the Agricultural Soil Information System of Korea). Based on accuracy and coherence analysis, the Ordinary Pairwise Partitioning (OPP) model classifying into 4 groups was selected to predict soil ecological quality. A soil ecological quality map for the forest area was constructed using the OPP model. In order to evaluate the applicability of the soil ecological quality map for planning eco-friendly road construction, optimal eco-road routes were simulated with varying weights assigned to soil ecological quality among various environmental and ecological factors. When the weight was greater than 14%, the soil ecological quality significantly influenced the determination of the route. These results from environmental informatics support the validity of using soil ecological quality information when planning for sustainable construction.

Keywords: sustainable development, soil ecological quality, microbial diversity, artificial intelligence, decision tree, environmental GIS

1. INTRODUCTION

Soil ecology is the foundation of the entire biosphere [9], and it plays a significant role in global ecosystems. Also, soil is an

integral component of our environment [7]. Soil quality or health can be defined as the 'capacity of soil to function as a vital living system, within ecosystem and land use boundaries, to sustain biological productivity,

to maintain or enhance water and air quality, and to promote plant, animal and human health' [6]. The wise use of soil and management of soil ecology is essential for sustainable development and land use [2]. The soil ecological system is a key factor in determining soil quality. Despite its significance, soil ecological quality has not been considered in environmental impact assessments of sustainable development. One of the main reasons for this omission is the fact that the methodology for assessing soil ecological quality has not been well established. Moreover, mapping soil ecological quality has yet to be developed for planning and designing sustainable development and land use.

Many researchers consider soil microbial diversity to be a sensitive indicator of soil ecological quality [1,10,11,19,25]. In soil ecological systems, soil microbes are predominant organisms over animals and plants, and they are involved in virtually every soil function [17]. In general, soil microbial diversity is reduced when microbial communities are exposed to environmental stimuli and stresses [15,25,29]. Thus, microbial diversity is a good indicator of soil ecological quality. In addition, the development of modern molecular tools, such as terminal restriction fragment length polymorphism (T-RFLP), facilitates rapid and quantitative measurements of microbial diversity from field samples [8,22].

To map soil ecological quality, information on microbial diversity is required for all the locations in an area of interest. However, it is not cost-effective to measure microbial diversity for every location. Instead, correlating a smaller size of field-measured diversity data (i.e., a training set of measured data) with existing soil environmental property data based on the geographical information system (GIS) can be used to predict soil microbial diversity. If this approach is viable, unmeasured microbial diversity data can be estimated based on the training set of measured diversity data. Then, geographical

information of microbial diversity can be mapped for a whole area of interest. This approach requires: (i) the availability of GIS-based soil environmental systems and (ii) a data-mining and predictive model framework based on artificial intelligence (AI). Fortunately, GIS-based soil environmental information systems are available in many countries. For example, the Agricultural Soil Information System (ASIS) of Korea provides information on soil physical and chemical properties, geological characteristics, soil erosion, drainage, and topography (<http://asis.rda.go.kr/>). However, an AI-based data-mining and predictive model framework has yet to be developed for microbial diversity mapping. In this work, the specific research aims were to (1) develop an AI-based data-mining and predictive model for microbial diversity using Decision Tree (DT) algorithms, (2) assess the accuracy and coherence of the developed model, and (3) demonstrate the validity of applying soil ecological quality maps to the planning of an eco-road route as an example of sustainable construction.

2. METHODS

2.1 *Field measurement and data collection*

Field-measured microbial diversity data were obtained from a forest area in which road construction was planned between Yongcheon and Seorak in South Korea. The locations of soil sampling were determined using a stratified random sampling method [16]. A total of 196 soil samples were taken from 15-20 cm below the top of the subsurface environment. Soil sampling was conducted during one late autumn season (November 2006) and one spring season (May 2007). Due to the dynamic nature of the soil, measurements are recommended to be taken at times of the year when the climate is stable and when there have been no recent soil disturbances [5]. Thus, late autumn and early spring seasons is generally chosen as appropriate periods for soil

sampling [21]. Soil samples were packed in sterile plastic bags and preserved at 4 °C prior to analysis.

For microbial diversity analysis, DNA extraction, polymerase chain reaction (PCR), and T-RFLP were performed. Microbial diversity was estimated using the Shannon index (H). In this estimation, the Shannon index was calculated from T-RFLP fingerprint profiles using Atlas' method (1997) as briefly described below [3].

$$H = (2.3/N) (N \log_{10} N - \sum n_i \log_{10} n_i)$$

where N = sum of peak areas in a given T-RFLP; n_i = area of a T-RF i ; and i = number of T-RFs of each T-RFLP pattern.

For soil environmental properties, data for nine variables were obtained from the ASIS database. In order to select primary influencing variables, we eliminated weakly interrelated variables through the Chi-Square test. The chosen primary influencing variables are: (1) gravel content of topsoil, (2) soil texture, (3) erosion class, (4) slope, (5) available depth, (6) drainage class, and (7) topography. These were used as input variables for the AI-based data mining and predictive model developed in this work.

2.2 Data pre-processing

Microbial diversity is a continuous variable. For the DT model approach, it is necessary to convert this continuity into a discrete classification. The measured microbial diversity values were classified into two, three, or four classes based upon the corresponding percentiles in the distribution of measured diversity values. For example, for three classes, the 0-33 percentile data belong to class 1, the 33-66 percentile data belong to class 2, and the 66-100 percentile data belong to class 3.

In the development of DT models, it is common practice to split the available data into two sub-sets: a training set and a test set [23]. The whole data set was randomly partitioned into the training data (70% of the data for each class) and the test data

(remaining 30% of the data) because the training and test sets have to be representative of the whole data set. In a similar fashion, to ensure the validity of the model, 10 pairs of sub-sets were generated for 10-fold cross validation. The training data were used for the training DT model. The test data were then used to measure the predictive accuracy of the trained model.

2.3 AI algorithms and operation

DT models were developed using SAS Enterprise Miner version 9.1 (SAS Institute Inc., 2002-2003, USA). The main procedure for developing a DT model is to split the data of the target variable based on its response to input variables (Figure 1). The Gini index, which is a standard created to measure inequality of wealth, was used to split tree nodes.

A recently developed ordinal pairwise partitioning (OPP) approach, proposed by Kwon et al. (1997), was used to modify the conventional DT model [13]. In a previous data pre-processing procedure, we partitioned the training set and test set into pairs. We used a partitioning method with the format of (N) & (remaining higher classes) which was also proposed by Kwon et al. (1997). In the OPP approach, using the forward method, the data set was partitioned into the pair (1, 2&3&4), the pair (2, 3&4), and the pair (3, 4) in the case of 4-class discretion. The case of 3-class discretion was conducted in the same way.

2.4 Accuracy and coherence analysis

Predictive accuracy was measured by hit rates between field-measured and model-predicted microbial diversity class values. For the modelling, training data were used. Then, the model predictions were carried out with the test data in order to evaluate the accuracy of the results. Each model was trained 10 times with the 10 prepared sub-sets of training data and run with the 10 sub-sets of test data. The hit rates from the 10 test runs were averaged and the average hit rate was reported as model accuracy.

For each model, coherence analysis was also carried out in order to assess the effect that randomly selecting training data has on the consistency of model outcomes. In doing so, for the whole area of the tested field, model predictions were independently run 10 times by models developed with different, randomly selected sub-sets of training data. For each polygon (a minimal area unit), the standard deviation was calculated and normalized by the range of discrete classes (i.e., maximum class – minimum class). The normalized standard deviation values were plotted in histograms (Figure 2).

2.5 Application of the soil ecological quality map: Eco-road route determination

With an optimal model, predicting the classes of soil ecological quality of unmeasured sites (scoring) was conducted using only input data of the sites.

Spatial predictions for soil ecological quality were implemented by coupling the output of the optimal model with GIS using a map overlay function. ArcGIS version 9.2 (ESRI Inc., 2006, USA) was used to construct the soil ecological quality map.

To determine eco-friendly routes for road construction, overall environmental quality information for the site, including soil ecological quality, is needed. In this study, overall environmental quality was subdivided into different components, as shown in Table 1. We weighted each component, except for soil ecological quality, as proposed by Choi (2007) [4].

To evaluate the importance of soil ecology in the natural environment, a simulation was conducted in which the weighting factor applied to soil ecological quality was varied [12] (Table 1). An overall environmental quality map was constructed that includes six environmental and ecological factors. With the overall environmental quality map, optimal eco-road routes were simulated using an ArcGIS-based Least Cost Path method.

Table 1: Overall environmental quality-related factors and their weights used in the GIS-based simulation.

Class	Sub-class	Weight (%) in sub-class													
Natural Environment (weight=0.64)	Soil	0	10	13	14	16	17	18	20	40	60	80	100		
	Ecological Quality														
	Animal & Plant	70	63	60.9	60.2	58.8	58.1	57.4	56	42	28	14	0		
	Topography & Geography	30	27	26.1	25.8	25.2	24.9	24.6	24	18	12	6	0		
Living Environment (weight=0.36)	Water	25	25	25	25	25	25	25	25	25	25	25	25		
	Quality														
	Air	19.4	19.4	19.4	19.4	19.4	19.4	19.4	19.4	19.4	19.4	19.4	19.4		
	Quality														
	Noise & Vibration	55.6	55.6	55.6	55.6	55.6	55.6	55.6	55.6	55.6	55.6	55.6	55.6		

3. RESULTS AND DISCUSSION

3.1 Model development

Figure 1 is a schematic of the flow of our AI-based data-mining and predictive model framework that was developed to estimate and map microbial diversity information. The new model framework consists of (i) data-mining training and (ii) prediction modelling. In the training stage, a multivariable correlation between the measured microbial diversity class data and the seven input soil environmental property variables was optimized using AI algorithms. For this optimization, discretely classified values of microbial diversity were used. In the prediction modelling (scoring) stage, the optimized AI algorithm was used to estimate the values of microbial diversity at unmeasured sites from ASIS-based soil environmental variables.

In this work, conventional DT (C-DT) was chosen as an AI-based algorithm for data-mining and prediction of soil ecological quality because it is frequently used in environmental and ecological data-mining problems, and it is relatively stable and accurate [14,18,20,24,27,28]. To improve the model performance of the conventional DT, an OPP approach was used as well (OPP-DT model). The respective performances of the C-DT and OPP-DT models were compared as described below. In addition, since it was unclear which degree of discrete classification for microbial diversity data allows better model performance, three

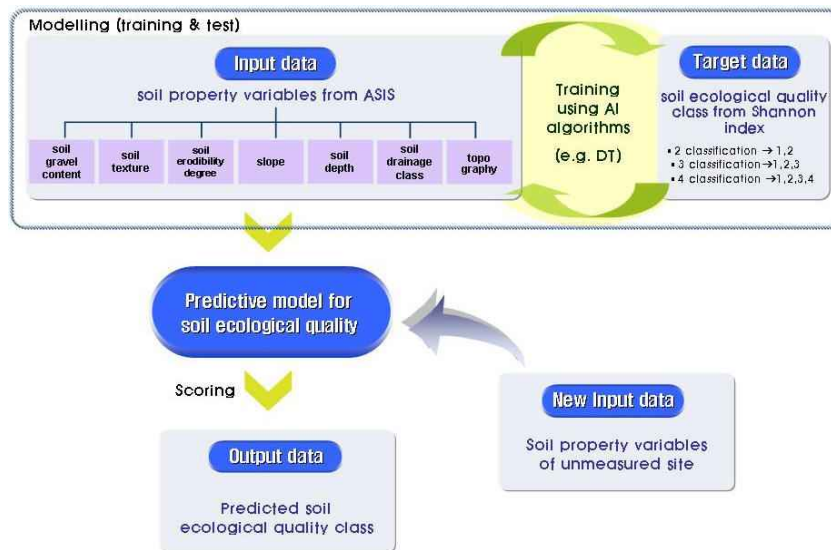


Figure 1: Schematic flow of AI-based data-mining and predictive model framework developed in this work.

different degrees of classification (2-class, 3-class, and 4-class discretion) were tested during model development, and the effects on model performance were explored. For these reasons, our model development produced five DT variant models: a conventional DT model with two discrete classes of microbial diversity (C-DT-2c), a conventional DT model with three discrete classes (C-DT-3c), a conventional DT model with four discrete classes (C-DT-4c), an OPP-DT model with three discrete classes (OPP-DT-3c), and an OPP-DT model with four discrete classes (OPP-DT-4c).

3.2 Model accuracy and coherence analysis

To compare the performance of the five model variants, accuracy and coherence were evaluated. Model accuracy was evaluated based on hit rate measurements between measured and model-predicted microbial diversity values from the test set. The average hit rates for each model are summarized in Table 2.

The OPP approach enhanced the model accuracy by 33% and 78% compared to the C-DT-3c and C-DT-4c models, respectively. These enhancements are greater than the results (approximately 4-10% improvement) presented by Kwon et al. (1997) [13]. With

respect to random conditions, C-DT models gave a slight improvement (14-27%). In contrast, OPP-DT models provided significant improvement (70-103%). In terms of overall accuracy, the OPP-DT-4c model showed greater improvement than OPP-DT-3c, while the OPP-DT-3c model has the benefit of providing relatively even hit rates among the different classes.

Table 2: Average hit rates for the tested AI-model variants.

Model	Class	Hit rate (%)	
		Conventional	OPP
DT-2c	1	62.19	
	2	57.50	
	Overall	60.63	
DT-3c	1	47.78	61.18
	2	34.86	53.51
	3	52.78	55.00
	Overall	42.47	56.56
DT-4c	1	52.73	65.43
	2	30.87	51.00
	3	14.38	39.29
	4	35.45	47.14
	Overall	28.50	50.72

Model coherence was measured to evaluate the consistency of the model predictions. Our coherence analysis was

designed to assess the effect of randomly selecting the training data on the consistency of model outcomes. This approach is similar to Bootstrap analysis, which is typically used in phylogenetic analysis [26]. According to the results in Figure 2, as the degree of classification (class number) increases, the degree of model coherence increases. For the 4-class models, the normalized standard deviations are smaller than 0.45. This suggests that a higher degree of classification may improve model coherence. If the degree of classification increases, however, the degrees of freedom will increase, causing the complexity and instability of modelling to grow significantly.

Based on the accuracy and coherence analyses, the OPP-DT-4c model was selected for predicting soil ecological quality of unmeasured sites. However, there is no exact answer for which optimal model has to be selected. It is generally recommended that model users consider model performance in response to different scenarios and their own preferences.

3.3 Application of the soil ecological quality map: Eco-road route determination

As an example of using the AI-based soil ecological quality map described here for sustainable construction, a soil ecological quality map was generated for the tested field sites using the OPP-DT-4c model. Using the generated map, planning eco-road route was performed for the same tested field. The optimal eco-road route was selected in the regions that have a minimum sum of overall environmental factors, including soil ecological quality. Since scientific experts show only weak consensus on the significance of soil ecology in road construction planning, we evaluated optimal eco-road route outcomes in response to varying the weights of soil ecological quality among various environmental and ecological factors (Table 1). When the weight of soil ecological quality was greater than 14% among other ecological factors, the soil ecological quality significantly influenced the route determination in comparison to the route calculated when soil ecological quality was not considered at all (data not shown).

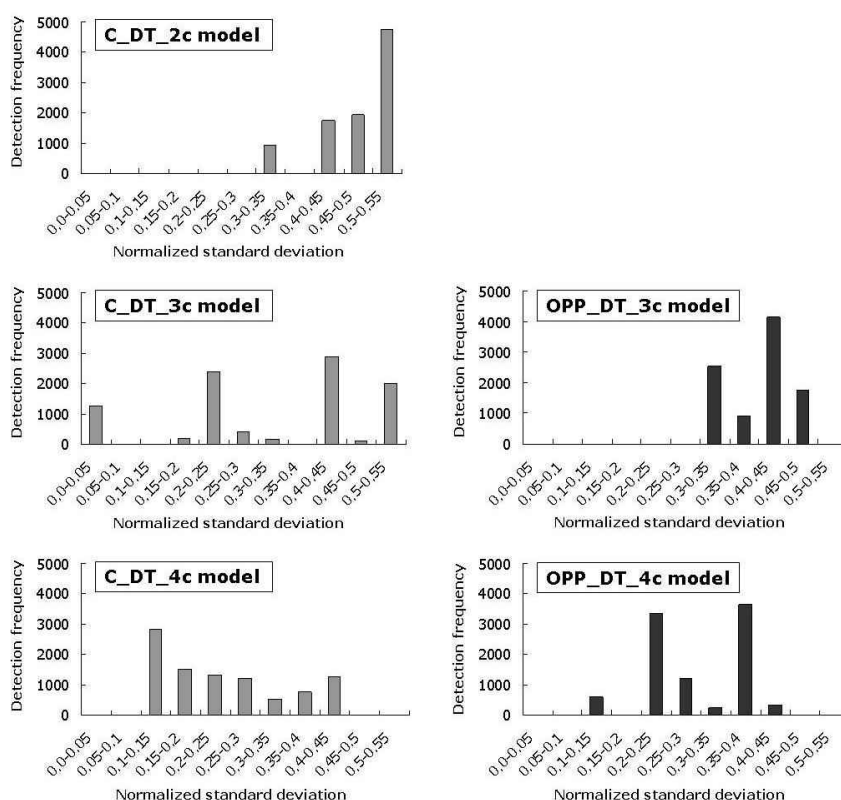


Figure 2: Results of model coherence analysis: Histogram of normalized standard deviations.

These results strongly suggest that planning the eco-road route is sensitive to soil ecological quality. This conclusion is consistent with our previous study [12]. However, in the previous study, the accuracy and coherence of the developed model were not evaluated in the mapping. Thus, the current work with those additional metrics provides a more conclusive outcome.

4. CONCLUSION

This study developed a novel approach for using AI-based algorithms to predict soil ecological quality from existing GIS-based soil environmental information systems. Therefore, this approach contributes to ecological information technology by providing a novel, efficient, and cost-effective approach to mapping soil ecological quality assessments. This will help to reduce the amount of fieldwork required for soil ecological quality mapping. In addition, the model simulation results support the use of our novel ecological information technology in planning sustainable construction projects. Finally, the AI-based information technology developed here can be applied to estimating unmeasured environmental information from existing GIS-based information systems. In this way, the proposed technology can be deployed in a wide range of ecological and environmental applications.

ACKNOWLEDGEMENT

This research was supported by Korea Ministry of Environment as the "GAIA project".

REFERENCES

- [1] Anderson, T. 2003. Microbial eco-physiological indicators to assess soil quality. *Agriculture, Ecosystems & Environment*. Vol 98. No 1-3. pp285-293.
- [2] Arshad, M & Martin, S. 2002. Identifying critical limits for soil quality indicators in agro-ecosystems. *Agriculture, Ecosystems & Environment*. Vol 88. No2. pp153-160.
- [3] Atlas R & Bartha, R. 1997. *Microbial Ecology: Fundamentals and Applications*. Forth Edition. Benjamin/Cummings Publishing Company. USA.
- [4] Choi, Y. 2007. A study on the GIS based environmental impact evaluation for route alignment. M.S. dissertation. Ajou University. Korea.
- [5] Dick, R, Thomas, D & Halvorson, J. 1996. Standardized methods, sampling, and sample pretreatment. In: *Methods for Assessing Soil Quality*. SSSA special publication. Vol 49. pp107-121.
- [6] Doran, J. 2002. Soil health and global sustainability: translating science into practice. *Agriculture, Ecosystems & Environment*. Vol 88 No 2. pp119-127.
- [7] Doran, J & Zeiss, M. 2000. Soil health and sustainability: managing the biotic component of soil quality. *Applied Soil Ecology*. Vol 15. No 1. pp3-11.
- [8] Dunbar, J, Ticknor, L & Kuske, C. 2001. Phylogenetic Specificity and Reproducibility and New Method for Analysis of Terminal Restriction Fragment Profiles of 16S rRNA Genes from Bacterial Communities. *Appl. Environ. Microbiol.* Vol 67. No 1. pp190-197
- [9] Filip, Z. 2002. International approach to assessing soil quality by ecologically-related biological parameters. *Agriculture, Ecosystems & Environment*. Vol 88. No 2. pp169-174.
- [10] Hill, G, Mitkowski, N, Aldrich-Wolfe, L, Emele, L, Jurkonie, D, Ficke, A, Maldonado-Ramirez, S, Lynch, S & Nelson, E. 2000. Methods for assessing the composition and diversity of soil microbial communities. *Applied Soil Ecology*. Vol 15. No 1. pp25-36.
- [11] Kennedy, A & Smith, K. 1995. Soil microbial diversity and the sustainability of agricultural soils. *Plant and Soil*. Vol 170. No 1. pp75-86.

- [12]Ki, D, Kang, H, Lee, S, Heo, J & Park, J. 2008. Sensitivity analysis of the effect of soil ecological quality information in selecting eco-friendly road route. The Journal of Korean Society of Soil and Groundwater Environmnet. In Press.
- [13]Kwon, Y, Han, I & Lee, K. 1997. Ordinal Pairwise Partitioning OPP) Approach to Neural Networks Training in Bond rating. International Journal of Intelligent Systems in Accounting, Finance & Management. Vol 6. No 1. pp23-40.
- [14]Moret, S, Langford, W & Margineantu, D. 2006. Learning to predict channel stability using biogeomorphic features. Ecological Modelling. Vol 19. No 1. pp47-57.
- [15]Pankhurst, C, Hawke, B, McDonald, H, Kirkby, C, Buckerfield, J, Michelsen, P, O'Brien, K, Gupta, V & Doube, B. 1995. Evaluation of soil biological properties as potential bioindicators of soil health. Australian Journal of Experimental Agriculture. Vol 35. No7. pp1015-1028.
- [16]Paterson, E. 1998. Development of a framework for soil monitoring for the state of the natural heritage reporting. Scottish Natural Heritage Commisioned Report F98AC112B. Scottish Natural Heritage. UK.
- [17]Rombke, J & Breure, A. 2005. Status and outlook of ecological soil classification and assessment concepts. Ecotoxicology and Environmental Safety. Vol 62. No 2. pp300-308.
- [18]Scheffer, J. 2002. Data mining in the survey setting: Why do children go off the rails? Res. Lett. Inform. Math. Sci. Vol 3. pp161-189.
- [19]Schloter, M, Dilly, O & Munch, J. 2003. Indicators for evaluating soil quality. Agriculture, Ecosystems & Environment. Vol 98. No 1-3. pp255-262.
- [20]Sesnie, S, Gessler, P, Finegan, B & Thessler, S. 2008. Integrating Landsat TM and SRTM-DEM derived variables with decision trees for habitat classification and change detection in complex neotropical environments. Remote Sensing of Environment. Vol 112. No 5. pp2145-2159.
- [21]Stenberg, B. 1999. Monitoring soil quality of arable land: Microbiological indicators. Acta agric. Scand., Section B, Soil and Plant Sci. Vol 49. No 1. pp1-24.
- [22]Tiedje, J, Asuming-Brempong, S, Nusslein, K, Marsh, T & Flynn, S. 1999. Opening the black box of soil microbial diversity. Applied Soil Ecology. Vol 13. No 2. pp109-122.
- [23]Tan, P, Steinbach, M & Kumar, V. 2005. Introduction to Data Mining. Addison Wesley. USA.
- [24]Vayssieres, M. 2000. Classification trees: An alternative non-parametric approach for predicting species distributions. Journal of Vegetation Science Vol 11. No 5. pp679-694.
- [25]Winding, A, Hund-Rinke, K, & Rutgers, M. 2005. The use of microorganisms in ecological soil classification and assessment concepts. Ecotoxicology and Environmental Safety. Vol 62. No 2. pp230-248.
- [26]Wong, K, Suchard, M & Huelsenbeck, J. 2008. Alignment Uncertainty and Genomic Analysis. Science. Vol. 319, No. 5862. pp473-476.
- [27]Yang, C, Prasher, S, Enright, P, Madramootoo, C, Burgess, M, Goel, P & Callum, I. 2003. Application of decision tree technology for image classification using remote sensing data. Agricultural Systems. Vol 76. No 3. pp1101-1117.
- [28]Zhang, B, Valentine, I & Kemp, P. 2005. A decision tree approach modelling functional group abundance in a pasture ecosystem. Agriculture, Ecosystems & Environment. Vol 110. No 3-4. pp279-288.
- [29]Zhang, W. 2005. Responses of soil microbial community structure and diversity to agricultural deintensification. Pedosphere. Vol 15. No 4. pp440-447.