2013 International Conference on Advances in ICT for Emerging Regions (ICTer): 222 - 227

Spatial data mining technique to evaluate forest extent changes using GIS and Remote Sensing

P.K.S.C.Jayasinghe[1], Masao Yoshida[2],

[1] *Department of Information Technology, Faculty of Computing, Sri Lanka Institute of Information Technology, Malambe, Sri Lanka*

[2] *Faculty of Agriculture, Ibaraki University, Chou 2-21-19, Ami Machi, Ibaraki ken, Japan*

[1]subash.j@sliit.lk, [2]myoshida@mx.ibaraki.ac.jp

Abstract*— Development of new computational, visual analytical and statistical methods to process, analyse, and understand complex and massive geospatial and temporal data is of vital importance at present in the world. Therefore, Spatial Data Mining (SPD) technique is very useful tool to access environmental phenomena. Spatial data mining is the process of discovering interesting and previously unknown, but potentially useful patterns from large spatial datasets. The main purpose of the present study was to identify forest extent changes during two decades using SPD techniques. For this study, multi-temporal satellite images (Land sat 5 TM 1992 and ASTER 2006) were used. Nuwaraeliya was selected as the study area for this research. Two thematic maps were derived from following tow approaches. The first and second approaches were consisted of unsupervised and supervised classification, respectively; derived thematic maps (unsupervised and supervised) were combined with Geographical Information System (GIS) overlay technique to generate a new map. These three maps were reclassified and converted to American Standard Code for Information Interchange (ASCII) format which is suitable formatting interface for SDM modelling. In order to carry out spatial data mining, Back-propagation algorithm was used. Overall accuracy of ASTER was 96.2 whereas that of land sat TM was 94. Results revealed that the extent of forest cover was lost by 5.28% in the present study area within in the period from 1992 to 2006. The results of this study are expected to be useful for researchers, managers and policy makers for updating existing forest maps, detecting forest changes and planning.

Keywords: - **Remote sensing, forest, GIS, Spatial data mining.**

I. INTRODUCTION

Spatial Data Mining (SDM) is a special kind of data mining. The main difference between data mining and SDM is that in SDM tasks it uses not only non-spatial attributes but also spatial attributes. Therefore, the SDM is an application of data mining technique to spatial data where the algorithms of SDM are capable to deal with noisy, fuzzy and incomplete data. Artificial Neural Network (ANN) can be used as the SDM technique [1] [2]. The SDM is a growing research field that is still very early stage. Therefore, it is well anticipated that more and more new uses of spatial data and novel spatial data mining approaches will be developed in the coming years.

The explosive growth of spatial data and widespread use of spatial database emphasize the need for the automated discovery of spatial knowledge. The complexity of spatial data and intrinsic spatial relationships limits the usefulness of conventional data mining techniques for extracting spatial patterns. Efficient tools for extracting information from geo-spatial data are crucial to organizations which make decisions based on large spatial datasets and many application domains including geology and environmental management, public safety, transportation, Earth science, epidemiology, and climatology [3]. Thus, developing of new computational, visual analytical, and statistical methods to process and understand complex and massive geospatial and temporal data is of vital importance to access environment phenomenon.

Forests are considered to be one of the world's most important and valuable natural resources. It plays an important and significant role in keeping the balance of the environmental stability [4]. However, at present, this resource is being degrading continuously because of the careless human interventions, bad political decisions, inappropriate policies, poor resource administration and supervision etc. A forest is an ecosystem, and therefore, the deforestation means not only the loss of trees but also the loss of the ecosystem and the environment [5]. Rapid decreases in the extent of forest resources can be mainly attributed to the increasing demand by the growing population, urban development, expansion of agricultural areas and industrial development [6]. Deflection of the forest cover is critical to the living creatures and has many ecological, social, and economical adverse effects including climatic changes, breakdown of nutrient and water cycle, soil degradation, floods and desertification.

Remote sensed images are used for determining land use of a given point at a given time. In order to be useful, the data must be classified appropriately into categories representing a set of identified characteristics. But computer assisted image classifications are still in a poor stage to produce land use/cover maps and statistics with high enough accuracy [7]. Several image classification techniques from automated to manual digitization can be found in the literature. However, most of these applications in image processing still rely on concepts developed in the early 70s and it is argued that they do not make use of spatial concepts [8] [9]. Meanwhile many studies managed to derive broad land use types, however with some difficulties that encountered when trying to characterize the complex land use/cover patterns accurately and precisely [10] [11]. If the full potential of the new image data sets for land use mapping is to be realized, a new inferential remote-sensing analysis technique has to be applied. In such cases, SDM technique is considered to be the on of the best methodologies used for the land use mapping.

The primary objective of this study is to derive current forest extent map using SDM technology. This paper describes our initial efforts, achievements and challenges in addressing some of the above areas. In particular, we present two time series change detection techniques for forest monitoring, illustrative examples of large scale vegetation disturbances, event identification, characterization and relationship mining.

978-1-4799-1276-6/13/$31.00 ©2013 IEEE

The other secondary objectives are;

1. To test the suitability and the accuracy of the temporal satellite images. Sri Lanka was selected to test this spatial data mining approach.
2. To monitor the changes of the forest cover. The result is expected to be helpful for forest resource management and future planning for the development of the areas.

II RELATED WORK

There are many related researches can be found in the literature. Fen Wu et al [12] used SDM assessment to identify the vegetation disaggregation classification in the farming-pastoral ecotone of North China. The method was used to classify the vegetation classes such as the closed forest, shrubland, and grassland with the exclusive spectral feature parameters. U.Kumar et al [13] were used multi-temporal remote sensing data with efficient SDM algorithm to monitor rapid urbanization which is important for natural resource management and sustainable planning activities. K.R. Manjula et al [14] expressed the widespread use of spatial database and spatial data mining technique to understand inter–relational nature of spatial data and the role of different driving factors for deforestation and the relationship among these factors. Kalli et al [15] and Seng et al [16] utilized the spatial data mining tasks including association rules mining, classification and prediction on forest fire.

III. METHODOLOGY

A. Remote sensing data

Two temporal satellite images were used for this study, namely the ASTER image acquired in 2006 and Land sat 5 TM image acquired in 1992. ASTER data is consisted with three types of spatial resolutions images. For the present study, Visible Near Infrared (VNIR) data of Terra/ASTER level 1B was used (Table 1).

Table 1: Basic information on remote sensing images

Sensor	Acquired date	Spatial Resolution
ASTER	05-08-2006	15m
Land sat 5 TM	13-03-1992	28.5m

B. Satellite image proccessing

As a part of the pre-processing stage, all images were first imported to ERDAS Imagine 9.2 (an image processing application) followed by geo-referencing and re-projection into UTM, WGS 84 datum, zone 44 north using a first-order polynomial and nearest-neighbour transformation. Then, a subset of the study area was created from the original satellite images. The ASTER and Land sat TM images were re-sampled. Finally, the post processing algorithms for all images such as radiometric, geometric and topographic corrections were implemented.

C. Unsupervised Classification

The unsupervised classification approach is an automated classification method that creates a thematic raster layer from a remotely sensed image. The two most frequently used algorithms are the K-mean and the ISODATA clustering algorithm [17]. ISODATA technique was used for unsupervised classification for both images in ERDAS 9.2 image-processing application. After the classification is completed, the analysis employs a posteriori knowledge to label the spectral classes into information classes. Initially, twenty five spectral clusters were formed to separate the image information into a more readable form. These twenty five clusters were carefully judged using expert knowledge and ground reference data. Spectrally similar classes of identical land cover types were merged. These merged clusters were evaluated according to the land use/cover classes (five classes) listed in Table 2. Finally, a labeling function was applied to generate a thematic forest cover and other land use/cover map.

D. Supervised Classification

Supervised approach requires pre analysed input from an analyst in order to automate the classification algorithm to associate pixel values with the correct land cover category. Supervised classification is a data-driven modelling tool in that the process derives statistical relationships between the inputs variables and the ground-truth habitats. The 'signature' is in the form of a statistical probability distribution in as many dimensions as there are input images. The probability distribution is calculated using the Maximum Likelihood Estimator (MLE).

In the process of supervised classification, we identified homogeneous sample pixels as training pixels in the image that can be used as representative samples for each land use category to train the algorithm to locate similar pixels in the image. For each land use/cover type, five areas of interest were prepared as the signatures of training samples. The training areas were created in order to discriminate the individual classes.

Table 2: Land use/cover classes

Class No.	Class name	Definition
1	Forest	Forest cover
2	Tea	Tea plantation
3	Residential	Houses, building s and roads
4	Farm lands	Vegetable, paddy and irrigated lands
5	Water	Water bodies

The land use map (study area) was developed by Survey Department of Sri Lanka (2002) and experiences of field visits were used to prepare the training signatures. After obtaining satisfactory discrimination between the classes during spectral signature evaluation, supervised classification was done using the parallelepiped non-parametric rule provided by ERDAS 9.2. Finally, thematic land use/cover map was generated with five classes (Table 2).

E. GIS overly processing technique

The combination of several classifications mapping provides best results than single classification. In a GIS overlay processing technique combines and mapping with previous supervised and unsupervised classification to produce an improved forest cover map. Firstly, the maps derived from supervised and unsupervised classification were converted to ESRI Grid format. Then, both of derived thematic maps were combined with GIS overlay function in ArcGIS application.

F. Spatial Data Mining

The above created three thematic maps (supervised, unsupervised and GIS overlay processing) are now in raster

format (grid). Then, reclassified each layer as forest cover area was assigned to 1 and other land use/cover was assigned to 2 (Table 3). Then, these raster layers were exported to ASCII format using the raster to ASCII tool in ArcGIS 9.2. The header information and no_data values written to the ASCII file was removed. A simple Visual Basic program (Microsoft Visual Studio. Net) was used to convert exported ASCII data file into the data format suitable for SDM modeling (Fig. 1).

The available data set was divided into two sets, training and testing. The ANN model was trained using randomly selected data, while the remaining data were utilized for testing of the network performance. The PC version of the Alyuda NeuroIntelligence 2.2 application was used to implement the three-layered network (Input layer – hidden layer – output layer). Output values of the network were set to 1 and 0 (Table 3). The back propagation algorithm was then applied to calculate the weights between the input layers and the hidden layers and between the hidden layers and the output layer, by modifying the number of hidden layers and the learning rate (iteration time) (Fig. 2).

The size of the hidden layer can be a crucial question in network design. For each network, the number of hidden units is varied and the network with the best performance in terms of the final network error is used for the classification. The hidden layer changed from 10 to 30, and the training iteration time changed from 100 to 2000. The initial learning rate was set to 0.5, and the momentum term was set to 0.2. Modifications of these parameters were then made by examining the dynamics of the error as suggested by [18]. The input layer consisted of three inputs which were corresponding to supervised, unsupervised and GIS post processing thematic maps, whereas output layer consisted of two neurons which were

Table 3: Output pattern of SDM modelling

Forest	Other land use/cover
1	0
0	1

Figure1: Data preparation methodology for ANN

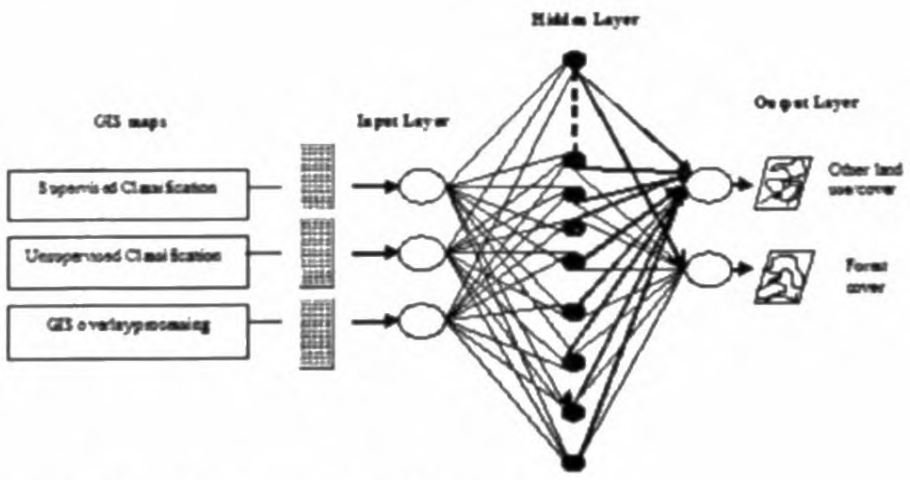

Figure2: Typical architecture for SDM modelling

corresponding to the forest cover and other land use/cover. In this study, the Root Mean Square Error (RMSE) and the Mean Error (ME) were used to assess the model accuracy. The model prediction with the lowest RMSE and ME was considered to be the most successful model [19].

After making and evaluating a number of network structures, the 3-18-2 network structure showed the lowest RMSE and ME values among the network structures (Fig. 3 and 4). The training results indicated that the best ANN model for predicting a suitable area for sugarcane was the 3-18-2 network structure.

When the number of hidden layers was less than 18, the network scale was too small, and the model prediction accuracy was low. When the number of hidden layers was greater than 18, the model could be over fitted. When the model was over fitted, the training accuracy was high, but the prediction accuracy might be decreased [20].

Identifying the relationship between training time and training accuracy is the next step of the SDM modeling. Prediction accuracies of the 3-18-2 network using the back propagation method with numbers of iteration are shown in Fig. 5. The peak in training accuracy was reached at a learning rate after 1400 iterations. If the iteration time exceeded 1400, the model could be over trained, which is another kind of over fitted situation. Based on this analysis, the network structure, training method and training times were considered to greatly affect the identification of a suitable SDM model. This experiment showed that the 3-18-2 network structure trained by the back propagation algorithm at 1400 training times was the best SDM model for the prediction of forest land cover according to the given input criteria.

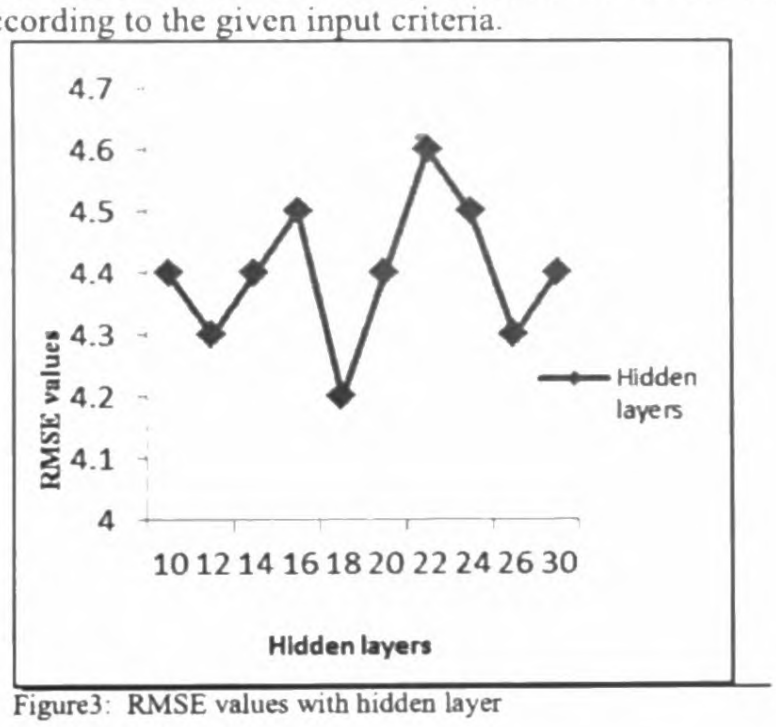

Figure3: RMSE values with hidden layer

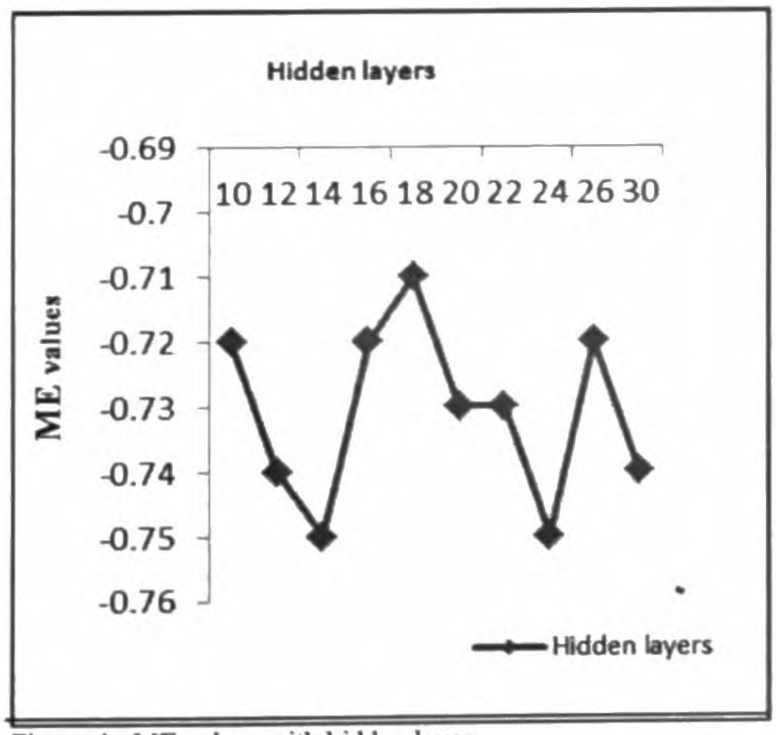

Figure 4: ME values with hidden layer

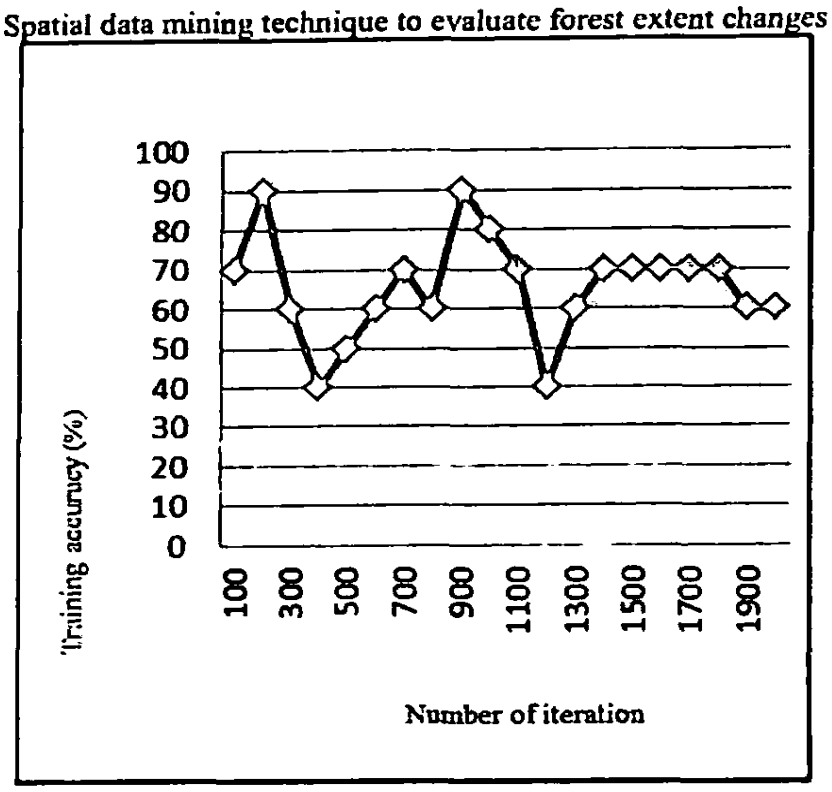

Figure5: Number of iteration with training accuracy

After the best ANN architecture was identified from the back propagation algorithm, these setting were used for the remaining experiments and classification process. Similar procedures were followed for the land sat 5 TM satellite image to produce forest cover map for 1992.

Based on above the setting of SDM modeling, the entire study area was evaluated using PC version of Alyuda Neurointelligence application. The result was then saved in spread-sheet format and converted back to GIS format for the output process. This information was evaluated to create a final thematic map based on the evaluation result from the SDM modeling. Finally, image was converted back to ERDAS *.img* image format for accuracy assessment.

IV. ANALYSIS & RESULTS

A. Classification Accuracy

The thematic maps derived from the image classification analyses are often compared with field observation, recent published air photos and survey results for the accuracy. Accuracy assessment is a general term for comparing predicted results with geographical reference data that are assumed to be the actual [21].

By incorporating the neighbourhood information, significant increases in overall classification accuracy was achieved. The improvement was identified across the classes, partly due to the homogeneity of the test sites chosen. Forests were identified more accurately because the confusion with arable land and orchards was decreased.

To evaluate the accuracy of the classified image, "Accuracy Assessment" tool in ERDAS 9.2 was used based on the random sampling method in which 130 points were automatically selected from referenced topographic map. The referenced values were recorded on the "Accuracy Assessment Table" based on the previous land use map of the study area. A non-parametric Kappa test was also used to measure the classification accuracy [22]. It provides a more rigorous assessment of the classification accuracy. The overall accuracy of the ASTER 2006 was 96.2 whereas that of the land sat TM 1992 was 94.6.

Overall Kappa statistics for the land sat TM 1992 image and the ASTER 2006 were 0.88 and 0.92, respectively. ASTER 2006 satellite image provided comparatively higher accuracy. Finally, these maps were re-sampled to 15x15m resolution and post classification smooth was applied for ease of analysis. One thematic map has only two classes namely forest and other land use/cover. The data was analyzed by MS excel application. In this approach, two thematic forest cover maps were produced for the year 2006 and year 1992. The methodology described above was applied for both satellite images. The forest extent maps of 1992 and 2006 are shown in Fig. 6. Table 4 presents the results of the forest cover areas and forest land changes between 1992 and 2006.

V. DISCUSSION

According to the results, the extent of forest cover was lost by 5.28% (2973.49 ha) in the present study area within the period from 1992 and 2006. The results of the final classified maps showed that the forest cover areas were rapidly decreased in the southern part of the study area. Moreover, the forest cover in the north-east area of the map was also seemed to be lost. In the southern part of the present study area, the residential areas are being growing very rapidly during last two decades. That might be the main reason for the depletion of the forest covers. The other reasons might be the increase of farmlands for cultivation and the natural hazards.

Table 4: Results of SDM modeling

Year	Forest area (ha)	%
1992	19973.97	36.51
2006	17000.48	31.23
Forest lost	2973.49	05.28

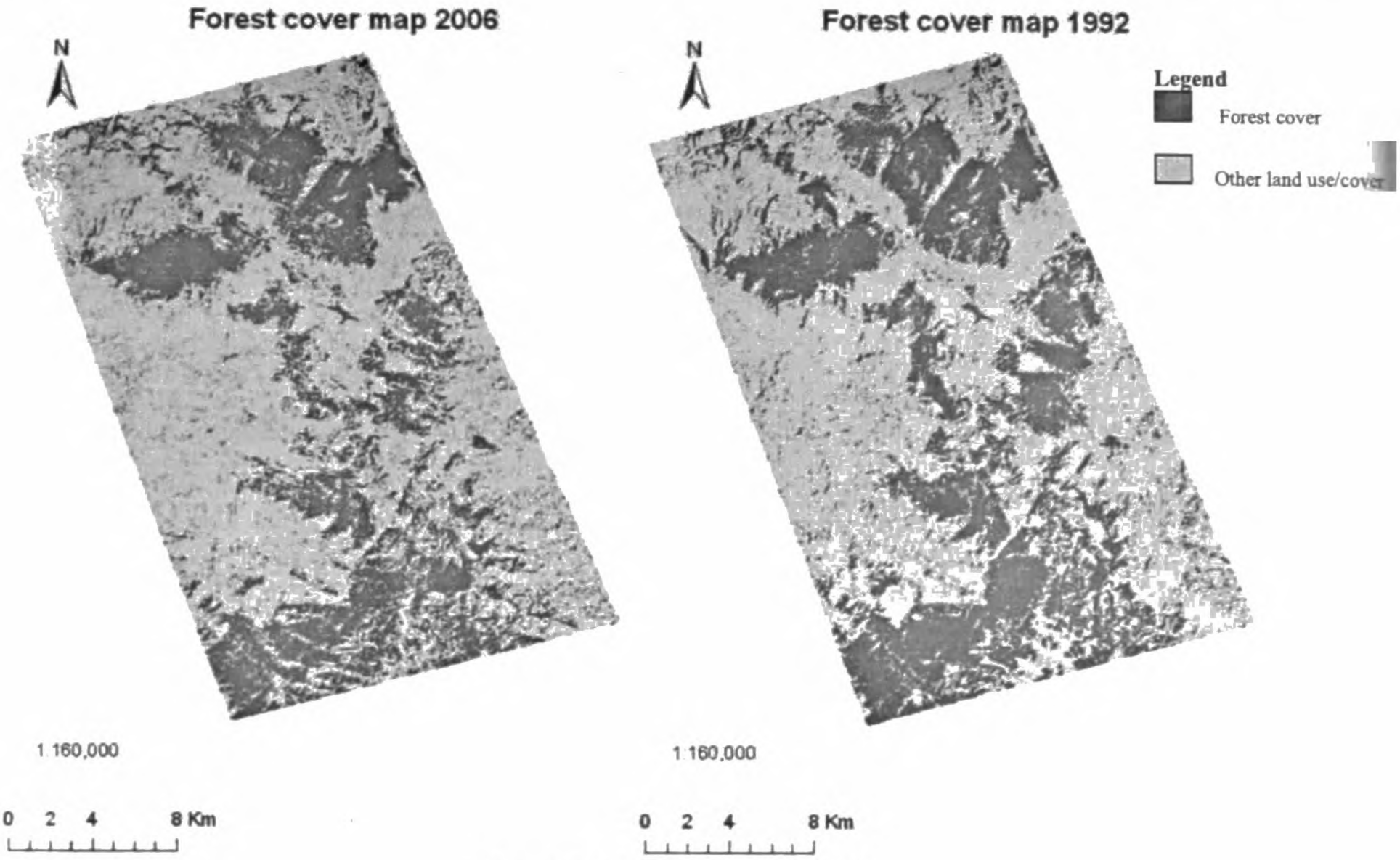

Figure6: Forest extent map in 1992 and 2006

Remotely sensed images remain difficult to classify for various reasons. However, SDM present a promising mode to improve the classification of remotely sensed images. Many authors reported higher accuracy when classifying spectral images with an SDM approach rather than with statistical methods such as maximum likelihood [23]. However, a more important contribution of the SDM is their ability to incorporate additional data into the classification process.

VI. CONCLUSION

In this paper, we attempted to develop high accuracy forest cover map by the combination of multi-classification approaches and using modern computer SDM technique. In the present study, a new methodology (three different classification approach combined with SDM) for land classification was developed. Multi years' satellite images used for identifying forest extent changes during the period between 1992 and 2006 in Nuwaraeliya of Sri Lanka. The use of multi-year satellite data in conjunction with GIS and neural network provided an opportunity for forest cover monitoring, surveying and change detection, which can be helpful in monitoring deforestation would be required for future national policy planning such as forest conservation measures.

Present study exposed the areas where the forests are altered. This provides resource managers with a basis for making a practical land use decisions. A strong relationship was observed between forest degradation and increase of land use change. Remote sensing is one of the most valuable tools in mapping and monitoring due to its advantages over traditional procedure in terms cost of effectiveness and timeliness with the availability of information over lager areas. Moreover, an integration of multi-sensor and multi-temporal data effectively improves the temporal attribute and the reliability of multi-data. Although satellite remote sensing data is increasingly available, spatial detection of deforestation for biological conservation is not widespread in the developing world. Although analysis and initiative have been carried out to some extent and updated information is still insufficient.

This study shows that the forest covers of the southern areas are vulnerable and could be further deteriorated if proper forest management strategies and protection measures would not be applied immediately. The results of this study might be useful for researchers, managers and policy makers for updating existing forest maps, detecting forest changes and planning for biodiversity management.

There is an urgent need for effective and efficient methods to extract unknown and unexpected information from spatial data sets of unprecedentedly large size, high dimensionality, and complexity. To address these challenges, spatial data mining and geographic knowledge discovery has emerged as an active research field, focusing on the development of theory, methodology, and practice for the extraction of useful information and knowledge from massive and complex spatial databases. Therefore, this study attempted to address above mentioned challenge using SDM technology.

REFERENCES

[1] S. Wang, D. Liu, X. Wang, Spatial Reasoning Based Spatial Data Mining for Precision Agriculture, AP Web Workshops 2006, LNCS 3842, pp. 506-510, 2006..

[2] M.D. Richard, R.P.Lippmann, Neural network classifiers estimate bayesian a posteriori probabilities. Neural Computation 3, 461–483,1991.

[3] J.A. Richards, Remote sensing digital image analysis: an introduction: Springer-Verlag, New York, 304 p, 1996.

[4] FAO FRA, On Definitions of Forest and Forest Change, Food and Agriculture Organization of the United Nations, Rome,2000.

[5] M. Richards, towards valuation of forest conservation benefits in developing countries. Environ. Conserv. 21: 308–319,1994.

[6] T. Rudel, J. Roper, The paths to rain forest destruction: cross national patterns of tropical deforestation, 1975–1990. World Dev. 25:53–65,1997.

[7] B. Prenzel, P. Treitz, Comparison of function- and structure-based schemes for classification of remotely sensed data. Int J Remote Sens 26: 543–561,2005.

[8] T. Blaschke, S. Lang, E. Lorup, J. Strobl, Object-oriented image processing in an integrated GIS/remote sensing environment and perspectives for environmental applications. Environmental Information for Planning, Politics and the Public. Metropolis Verlag, Marburg, 2: 555–570, 2000.

[9] G.M. Foody, Land covers classification by an artificial neural network with ancillary information. Intl J Geo Info Sys 9:527-542, 1995.

[10] T. Fung, K. Chan, Spatial composition of spectral classes: a structural approach for image analysis of heterogeneous land use and land cover types. Int Photogramm Eng Remote Sens 60: 173-180,1994.

[11] S.E. Franklin, M.A. Wulder, Remote sensing methods in medium spatial resolution satellite data land cover classification of large areas. Progr Phy Geo 26:173-205,2002.

[12] F. Wu, J. Zhan, H. Yan, C. Shi, and J. Huang, Land Cover Mapping Based on Multisource Spatial Data Mining Approach for Climate Simulation: A Case Study in the Farming-Pastoral Ecotone of North China. Advances in Meteorology, Vol. 2013. 1-12, 2013

[13] U.Kumar. and T.V. Ramachandra Spatial Data Mining and Modeling for Visualization of Rapid Urbanization. ISIT journal, Volume 9 and 22-43, 2009.

[14] K.R. Manjula, S.Jyothi, S. Ananda Kumar Varma, Analyzing the Factors of Deforestation using Association Rule Mining, geospatialworld.net, 2012.

[15] S.N.P Kalli, S. Ramakrishna, An autonomous forest fire detection system based on spatial data mining and fuzzy logic, IJCSNS International Journal of Computer Science and Network Security, Vol.8 No.12, 49-55, 2008.

[16] C.T Seng , K. Wynne, H.L Kim, C. Y. Lee, Spatial data mining: Clustering of hot spots and pattern recognition, In Geoscience and Remote Sensing Symposium, 21-25 July 2003 (IEEE International), 3685-3687, 2003.

[17] K. Johnsson, Segment-based land-use classification from SPOT satellite data. Int Photogramm Eng Remote Sens 60: 47–53,1994.

[18] P.D. Heermann, N. Khazenie, Classification of multispectral remote sensing data using a back-propagation neural network, LEEE Transactions. Geosen Remote Sens 30:81-88,1992.

[19] K. Hornik, Approximation capabilities of multilayer feed forward network. J. Neural Network. 4: 251-257,1991.

[20] Z.S.H. Chan, H.W. Ngan, A.B. Rad, Short-term ANN load forecasting from limited data using generalization learning strategies. Neurocomputing 70: 409-419,2006.

[21] T.M. Lillesand, R.W. Kiefer, J.W. Chipman remote sensing and image interpretation. New York: John Wiley & Sons, Inc,2008.

[22] G.H. Rosenfield, K. A. Fitzpatirck-Lins, coefficient of agreement as a measure of thematic classification accuracy. Photogramm Eng Remote Sens 52: 223–227,1986.

[23] J.D. Paola, R.A. Schowengerdt, A review and analysis of back-propagation neural networks for classification of remotely-sensed multi-spectral imagery. Int J.Remote Sens 16:3033–3058, 1995.