# K-Nearest Neighbor Method and Neural Network for Classification of Forest Encroachment by Using multi Remote Sensing Data

Ahmed1 A. Mehdawi and Baharin Bin Ahmad
Faculty of Geoinformation and Real Estate, Institute of Geospatial Science and Technology (INSTeG), Universiti Teknologi, Malaysia

**Abstract:** This study gives sophisticated result in the use of K-Nearest Neighbor (KNN) and Neural Network (NN) techniques as classification tools for deforestation. The major focus is on the data and technique that can be used to identify the changes in forest features. This study will concentrate on identifying forest encroachment in tropical forests such as the forests of Cameron Highland Malaysia. This technique study will establish a strong mechanism that can be used by different sectors such as forestry, local administration, surveying and agriculture. The main contribution of this study is that it utilizes of K-Nearest Neighbor (KNN) Method and (NN) with multi-remote sensing data to detect any change may it happened on forest. Hopefully, this study will serve as a reference for any future research on utilizes of KNN and NN techniques as classification tools to identify of tropical forest encroachment.

[Ahmed1 A. Mehdawi and Baharin Bin Ahmad. **K-Nearest Neighbor Method and Neural Network for Classification of Forest Encroachment by Using multi Remote Sensing Data.** *Life Sci J* 2013; 10(4): 1129-1134]. (ISSN: 1097-8135). http://www.lifesciencesite.com 148

**Keywords:** K-Nearest Neighbor Method, Neural Network, Remote Sensing, Malaysia and tropical forests encroachment

## Introduction

Preservation farm land utilizes arranging within humid tropical lowland forests and appropriate remote sensing approaches to differentiate amongst the many floristically different forest types.Satellite images and airborne photographs would definitely be a most important provider of greenery data in remote tropical zones in which various other data sources, for instance maps of vegetation, soil along with topography are often inaccessible (Ferrier, 2002 and Lawal et al., 2011). Floristically various tropical rain forest zones repeatedly lack wide-ranging field data as a result of logistical difficulties as well as challenges with outlining incredibly diverse or perhaps even negatively known flora (Ruokolainen et al., 1997). Whenever associated with species makeup or structural circumstances on the ground, satellite images produce a low-cost source most typically associated with data for the purpose of determining and mapping rain forest kinds but for some data only, a necessary with respect to preservation planning and maintainable forest management (Favrichon, 1998; Ferrier andGuisan, 2006; MargulesandPressey, 2000).

Tropical rain forest kinds are likely to be categorized through extended physiognomic and design characteristics, mainly because they normally seem structurally comparatively homogeneous within terrain dimensions. On top of that, the field details are in most cases as well coarse just for identifying floristic variance as a result of selecting mistake coming from variations which are usually as a consequence of certain geographical elements. Accordingly, just a couple of classes are actually employed to differentiate rain forest kinds making use of remote sensing methods (Achard et al., 2001; Kleinn et al., 2002; Stibig et al., 2003). Lately, numerous researches have shown the fact that rain forests have a relatively wide range of soil-related fine-grained floristic and architectural variation, that isn't viewable through the existing greenery maps.

The quality of natural environment classification accuracy and reliability originating from a satellite data is determined by the classification algorithm (Kleinn et al., 2002) as well as the image resolution (pixel window or segment size) utilized for the process. Additionally, distinctly identified forest classes are required to assess classifiers to get thematic accuracy. A commonly utilized supervised classification strategy is discriminant analysis (Thenkabail et al., 2004).

Discriminant analysis mission to find the linear mix off variables (e.g., spectral features) that most effective discriminates within classes. A non-parametric replacement for discriminant analysis is the K nearest neighbours (K-NN) classifier. The K-NN technique has usually been employed to calculate forest inventory variables from satellite imagery, such as total volume and basal area for temperate and boreal zones (Gjertsen et al., 2000 and McRoberts et al., 2007), on the other hand, it happens to be significantly less common with regard to image relying classification of forest kinds. Within the tropics it has been examined once, with promising outcomes, to get estimating frequent floristic dissimilarities. A particular advantage of the non-parametric K-NN classifier is that it will not make any distributional assumptions in regards to the variables used. In K-NN, the pixel whose class is unfamiliar is a member of a class as outlined by its spectrally nearest neighbours (e. g. spectrally most equivalent pixels) whose class identities are recognized.

The K-nearest neighbor (KNN) multisource inventory has turned out to be timely, cost-efficient, as well as precise while in the tropical forest and Malaysian trials. This strategy designed for improving field point inventories will be perfectly appropriate for the evaluation and observation requires of Authorities agencies, such as the Forest Service, that carry out natural and additionally agricultural resource inventories. It includes wall-towall maps of forest features, continues the natural data variety perfectly found on the field inventory and presents accurate and localized estimates in accordance metrics throughout significant areas as well as ownerships.

Within a pixel-level classification, the KNN algorithm assigns every unidentified (target) pixel the field features of essentially the most equivalent reference pixels for which field data occurs(Franco-Lopez et al., 2001). Likeness is defined in conditions with the feature space, traditionally measured as Euclidean or Mahalanobis as the equation 1 distance between spectral featuresin addition Euclidean distance function is used for measuring the distance between the holdout point (star) and the training point (blue/white) as shown in figure 1.Associated with principle purpose should be to explore whether discriminant analysis and k-NN classifiers will be able to efficiently classify study area (Cameron Highland as rain forest types to identify the most suitable classification approach via accuracy to detect forest change Cases that are near each other are seemed to be “neighbours.” When a new case (holdout) is presented, its distance from each of the cases in the model is computed. The classifications of the most similar cases – the nearest neighbours – are tallied and the new case is placed into the category that contains the greatest number of nearest neighbors.

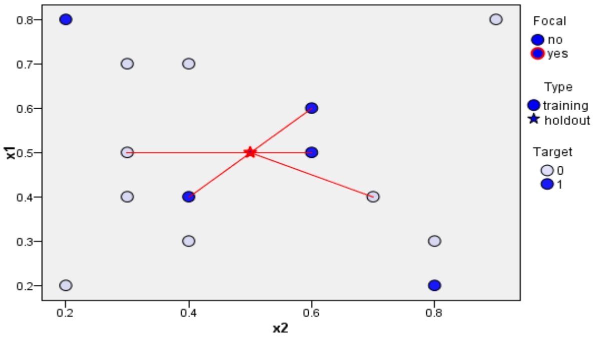


Figure 1shown distance between the holdout point (star) and the training point (blue/white)

$$d_E(x,y)= \sum_{i=1}^{N}\sqrt{x_i^2 - y_i^2} \qquad (1)$$

In kNN classification of Forest Inventory and Analysis FIA-defined forest kinds. Particular importance is determined relating to increasing mapping productivity by lessening classification feature space, reducing the number of distance estimations in the most adjacent neighbor search, along with eliminating redundancy in redundant nearest neighbor searches by constructing a database of feature patterns connected with different forest kind classes.

An artificial neural network (NN) is a biologically motivated computational version having a running elements (known as neurons) as well as associations in between them by using coefficients (weights) likely to the relationships which usually constitute the neuronal design. The connection weights are the ”memory” of the method. Single neurons which carry out data processing operations are usually linked together towards networks, weights are modified through data training. The dialog dependent neural network NN application we built is running with a better performance as comber with KNN. The main options consist of determining the number of input/output files, ANN layers and nodes, training noise, maximum training time, acceptable training errors, and quantity of unclassified classes. Simply because demonstrated in result of using back propagation neural net work performed by ENVI software in figure 2 we can discover the robust indicator of the ability of using NN instead of KNN however still the type of data perform the most important function to evaluate the techniques efficiency.

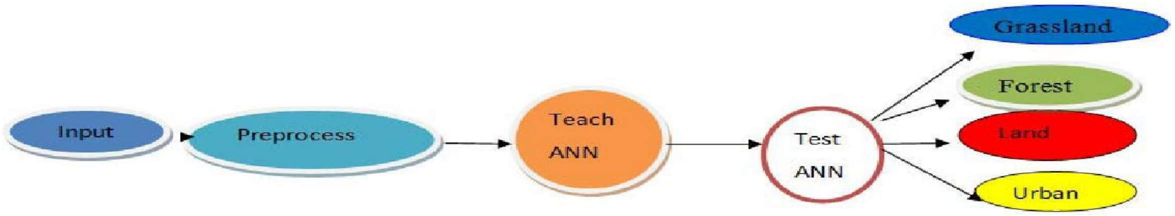


Figure 2 show the overview of the trained bands

**Methods and Dataset**

General Information forthe Study Area

The district of Cameron Highlands (4 ° 28' N) (101 ° 23' E), Pahang, Malaysia, is located on the main range of Peninsular Malaysia as shown in Figure 3 and 4. It covers a total area of 71 000 715 $km^2$ (Fortuin, 2006).Generally, the terrain is mountainous and strongly dissected with 10–35 ° slopes. More than 66 per cent of the land has a gradient of more than 20 °. The Cameron Highlands are about 715 $km^2$ in area settled between roughly 900 and 1800 m and surrounded by forested peaks rising to 2032 m. Malaysian lowlands are heavily disturbed, so upland forests like those of the Cameron Highlands are an important refuge for biodiversityas shown in figure 3 by QuickBird satellite. The Cameron Highlands are significantly cooler than Malaysia’s lowlands, with a mean daily minimum of 14.8℃, a mean daily maximum of 21.1℃, which suits temperate crops. The rainfall averages 2660 mm yr-1, humidity is high and there is no marked dry season.

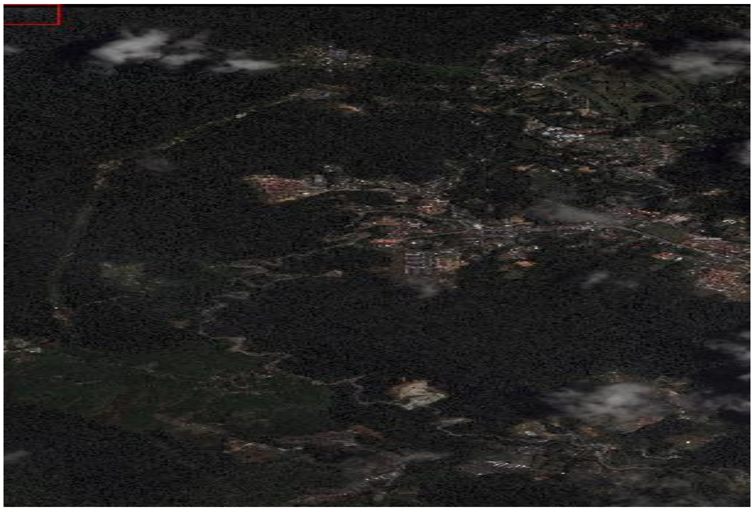


Figure 3: showed QuickBird Image for study area in 2007

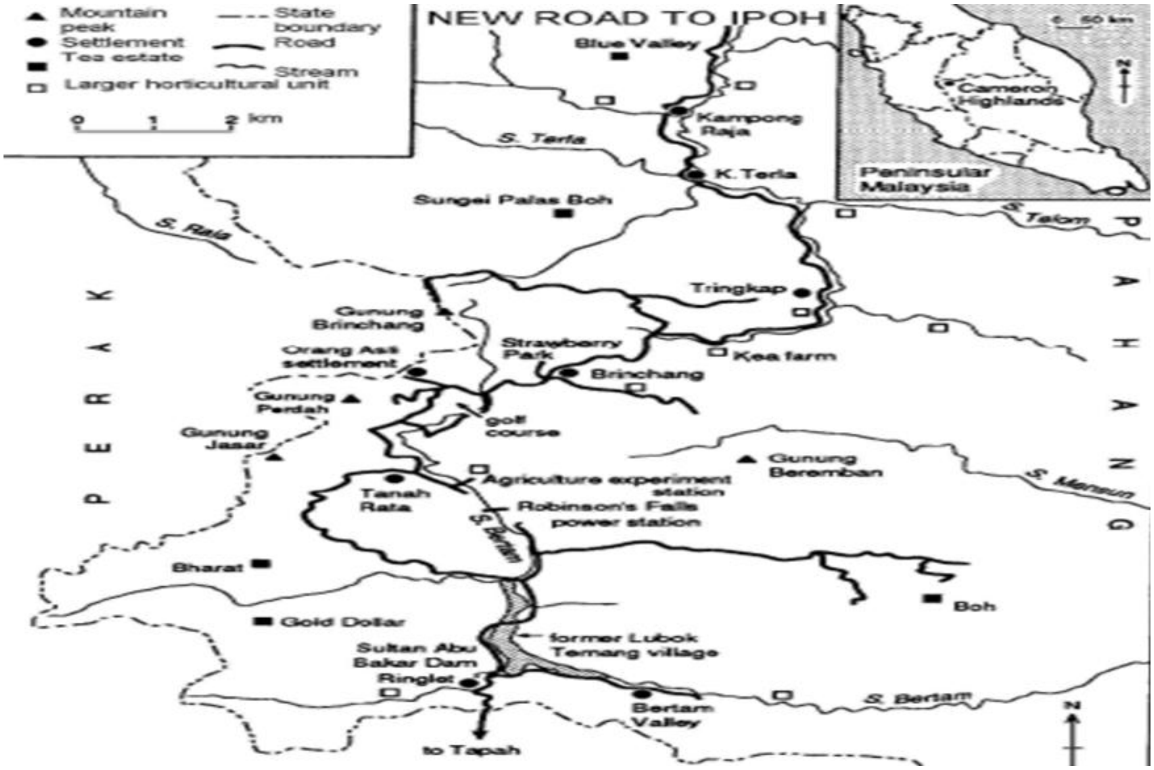


**Figure 4**: The location of Cameron Highland located between Perak and Pahang (Malaysian)

**Dataset**

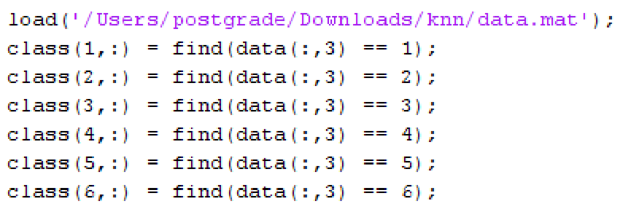
This dataset is categorized as multiclass dataset. The dataset consists of one feature (average of reflectance) and 6 classes. The classes are named as follows, (1) Bush, (2) Cover, (3) Forest, (4) Tea, (5) Vege, and (6) soil. Each class consists of 2151 samples of total 12906 samples processed on Mata lab software and Quickbird image which processed in ENVI software as originality conceder as optical data( Luo et al 2012). The summary of the dataset is tabulated in Table 1 as it used for matlab as input function.

Table 1 Summary of the dataset

| Class | Representation of Class | Feature |
|---|---|---|
| Bush | 1 | Reflectance |
| Cover | 2 | Reflectance |
| Forest | 3 | Reflectance |
| Tea | 4 | Reflectance |
| Vege | 5 | Reflectance |
| Soil | 6 | Reflectance |

**Methods**

Raining and testing data is divided using the method of 3-fold cross validation. For experimenting using MATLAB, user requires to firstly choose the directory of m file as shown below in figure 5, beside that the flow chat of the method summarized the study steps as showed in figure 6.



```
load('/Users/postgrade/Downloads/knn/data.mat');
class(1,:) = find(data(:,3) == 1);
class(2,:) = find(data(:,3) == 2);
class(3,:) = find(data(:,3) == 3);
class(4,:) = find(data(:,3) == 4);
class(5,:) = find(data(:,3) == 5);
class(6,:) = find(data(:,3) == 6);
```

Figure 5 shows the bath at Matlab software to the user to manage the large number of data.

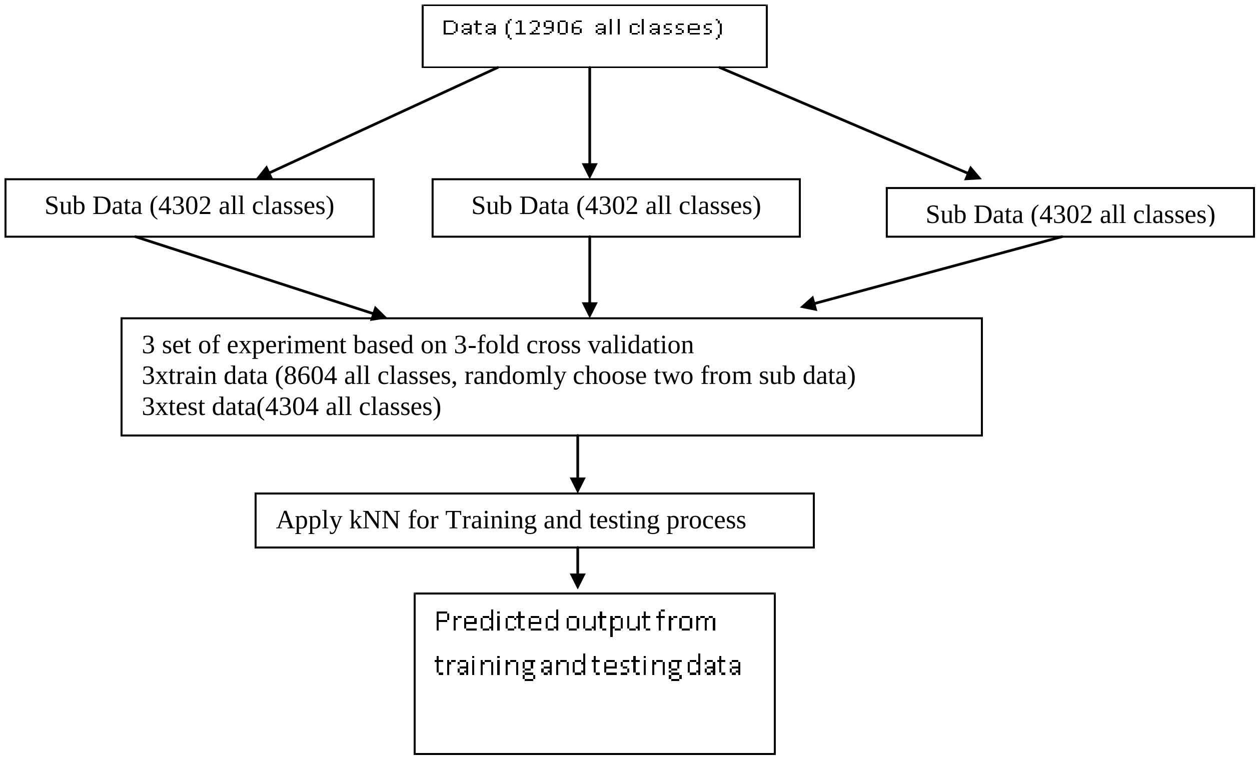

Data (12906 all classes)
Sub Data (4302 all classes)
Sub Data (4302 all classes)
Sub Data (4302 all classes)
3 set of experiment based on 3-fold cross validation
3xtrain data (8604 all classes, randomly choose two from sub data)
3xtest data(4304 all classes)
Apply kNN for Training and testing process
Predicted output from training and testing data

Figure 6 explain flowchart of the methods.

With regard to calculating by using Euclidean distances, take into account the spectral distance$d_{pi,p}$, that is definitely calculated in the feature space through the focus on pixel p to every single reference pixel $p_i$by which the forest kind class is recognized. For every single pixel $p$, sort the k-nearest field plot pixels (from the feature space) just by$d_{pi,p} \leq \ldots \leq d_{pk,p}$. The imputed value with the pixel $p$ will likely be expressed as an objective of the nearest units, every single such unit calculated as outlined by this distance breaking down function:

$$w_{pi,p} = \frac{1}{d^t_{pi,p}} \Big/ \sum_{i=1}^{k} \frac{1}{d^t_{pj,p}} \qquad (2)$$

In which $t$ can be described as distance decomposition factor set equivalent to 1 for all trials.

To help impute class variables which include forest type and MATLAB variables as in table 2, the distance decomposition function calculates a weighted mode value.

Table 2 Description of MATLAB's Variables

| No | MATLAB variable | Descriptions |
|---|---|---|
| 1 | trainingdata | Training data |
| 2 | trainingClass | Class of training data |
| 3 | sampledata | Testing data |
| 4 | sampleClass | Class of testing data |
| 5 | predictedClass | Predicted Class of testing data |
| 6 | AccuracyPercentage | Accuracy from testing data |
| 7 | test1-6 | Testing data with their original output class |
| 8 | testdata1-6 | Testing data with their predicted output class |

To get a class variable, the error rate (*Err*) suggests the difference of opinion between an estimated value $\hat{y}$ and the real response$y$ in a dichotomous circumstance these types of which $y$does indeed or doesn't necessarily participate in class $i$ (Efron and Tibshirani 1993). Therefore, by implementing the overall accuracy (OA) (Stehman 1997,) described as follows:

$$OA = 1 - Err, \qquad (3)$$

$$Err = \sum_{i=1}^{n}(y_i - \hat{y}) / \mathrm{n} \qquad (4)$$

Here is the exclusive circumstance with the mean square error for a great signal variable. A lot of these estimators have been recommended over the typical Kappa estimator regarding causes offered by Franco-Lopez et al. (2001).

Errors have been calculated by means of leave-one-out cross-validation. This method omits training sample models individually together with mimics the application of independent data.Per omission; we applied the kNN prediction principle towards the outstanding small sample. Eventually, the errors out there forecasts have been made clear. Overall, we employed the prediction rule *n* times and predicted the results to get *n* units.This kind of rates involving prediction error usually are practically unbiased (Efron and Tibshirani 1993).

The dialog based neural network application we built is running under with a better performance (Ahmed and Baharin, 2012). The fundamental options include determining the number of input/output files, ANN layers and nodes, training noise, maximum training time, acceptable training errors, and quantity of unclassified classes as shown in figure 7.To enhance the truth of classification, we managed to save the acceptable classes after which built a new sub-training looking for second tier classification. Several unique features from the new training set were made such as the removal of unclear road and crop/grass sample pixels and adding more representative pixels.

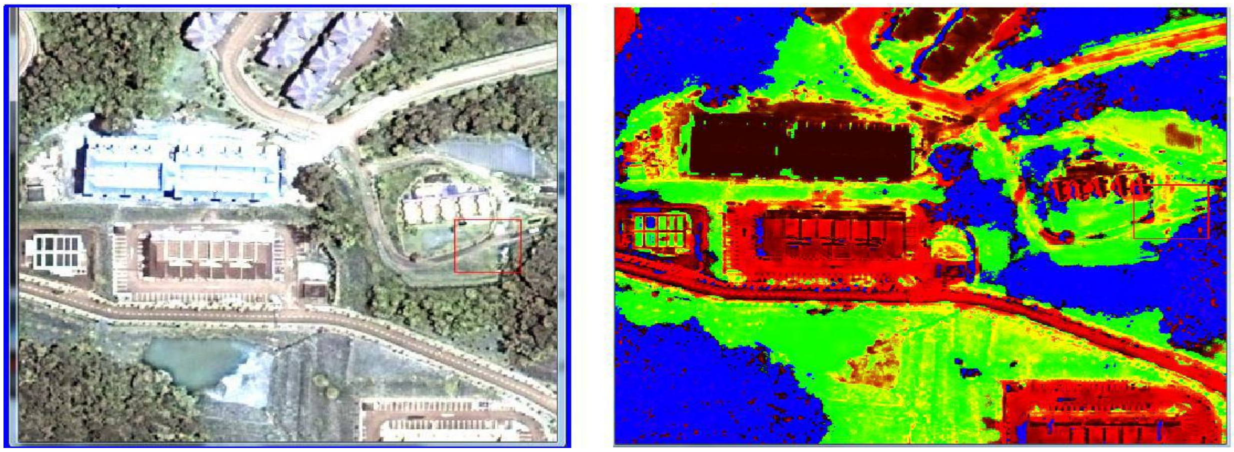


Figure 7 show the original area before and after ANN classification

## Result and Discussion

Simply because taken into account within the Methods portion, while k increases, a better probability is present which a reference observation is going to cross typically the inequality in addition to require the whole Euclidean distance contrast using the target. Our studies verified this unique relationship among increasing *k* as well as range of Euclidean distance measurements.

The majority of significantly, our studies show that using the *KNN* algorithm could substantially enhance mapping productivity by reducing the amount of measurements if necessary.

While using new training set and just training the not-effective classes (zoom-in) we've got a much better output as proven by using NN techniques,. The result was enhanced through getting an precision of 96.55%, although some classes almost didn't change because of the street class in the multispectral data such as QuickBird.

Implementing our database-assisted mapping may possibly appreciate experience success whenever classes were much more spectrally distinct as proven by using KNN with Spectroradiometer Data.

**References**

[1] Achard, F., Eva, H., and Mayaux, P. (2001). Tropical forest mapping from coarse spatial resolution satellite data: production and accuracy assessment issues. International Journal of Remote Sensing, 22, 2741−2762.

[2] Dano Umar Lawal, Abdul-Nasir Matori, Ahmad Mustafa Hashim, Imtiaz Ahmed Chandio, Soheil Sabri, Abdul-Lateef Balogun and Haruna Ahmed Abba, 2011. "Geographic Information System and Remote Sensing Applications in Flood Hazards Management: A Review". Research Journal of Applied Sciences, Engineering and Technology, 3 (9). pp. 933-947. ISSN 2040-7467 [Thomson ISI]

[3] Efron, B.; Tibshirani, R.J. 1993. An introduction to the bootstrap. New York: Chapman and Hall. 436 p.

[4] Favrichon, V. (1998). Modeling the dynamics and species composition of a tropical mixed-species uneven-aged natural forest: effects of alternative cutting regimes.Forest Science, 44, 113−124.

[5] Ferrier, S. (2002). Mapping spatial pattern in biodiversity for regional conservation planning: where to from here? Systematic Biology, 51, 331−363.

[6] Ferrier, S., and Guisan, A. (2006). Spatial modelling of biodiversity at the community level.Journal of Applied Ecology, 43, 393−404.

[7] Franco-Lopez, H., Ek, A. R., and Bauer,M. E. (2001). Estimation and mapping of forest stand density, volume, and cover type using the k-nearest neighbors method. Remote Sensing of Environment, 77, 251−274.

[8] Fortuin, R. (2006). Soil Erosion in Cameron Highlands, an Erosion Rate Study of a Highland Area. Saxion University Deventer.fragmentation and its impact on species diversity: an analysis using remote sensing and GIS. *Biology and Conservation,* 14, 1681-1698.

[9] Gjertsen, A. K., Tomter, S., &Tomppo, E. (2000). Combined use of NFI sample plots Ad Landsat TM data to provide forest information on municipality level. Proceedings of conference on remote sensing and forest monitoring, June 1–3, 1999, Rogow, Poland (pp. 167−174).

[10] Kleinn, C., Corrales, L., and Morales, D. (2002). Forest area in Costa Rica: a comparative study of tropical forest cover estimates over time. Environmental Monitoring and Assessment, 73, 17−40.

[11] Kleinn, C., Corrales, L., and Morales, D. (2002). Forest area in Costa Rica: a comparative study of tropical forest cover estimates over time. Environmental Monitoring and Assessment, 73, 17−40.

[12] Luo, W., Li, H and Liu, G. (2012) Joint Change Detection and Image Registration for Optical Remote Sensing Images. Research Journal of Applied Sciences, Engineering and Technology 4(6): 634-640, 2012 ISSN: 2040-746

[13] Margules, C. R., and Pressey, R. L. (2000). Systematic conservation planning. Nature, 405, 243−253.

[14] McRoberts, R. E., Tomppo, E. O., Finley, A. O., and Heikkinen, J. (2007). Estimating areal means and variances of forest attributes using the k-nearest neighbours technique and satellite imagery. Remote Sensing of Environment, 111, 466−480.

[15] Mehdawi., A and Ahmad., B. (2012) Classification of Forest Change by Integration of Remote Sensing Data with Neural Network Techniques. International Conference on System Engineering and Technology. September 11-12, 2012, Bandung, Indonesia.

[16] Prance, G. T. (1989). American tropical forests. In H. Lieth, and M. J. A. Werger (Eds.), Tropical rain forest ecosystems: Biogeographical and ecological studies (pp. 99−132). New York: Elsevier.

[17] Ruokolainen, K., Linna, A., and Tuomisto, H. (1997). Use of Melastomataceae and pteridophytes for revealing phytogeographical patterns in Amazonian rain forests.Journal of Tropical Ecology, 13, 243−256.

[18] Stehman, S.V. 1997. Selecting and interpreting measures of thematic classification accuracy. Remote Sensing of Environment. 62: 77–89.

[19] Stibig, H. J., Beuchle, R., &Achard, F. (2003).Mapping of the tropical forest cover of insular Southeast Asia from SPOT4-Vegetation images.International Journal of Remote Sensing, 24, 3651−3662.

[20] Thenkabail, P. S., Enclona, E. A., Ashton, M. S., Legg, C., and De Dieu, M. J. (2004). Hyperion, IKONOS, ALI, and ETM plus sensors in the study of African rainforests.Remote Sensing of Environment, 90, 23−43.

10/22/2013