ISSN NO. 2320-5407

Journal homepage:http://www.journalijar.com

INTERNATIONAL JOURNAL OF ADVANCED RESEARCH

RESEARCH ARTICLE

Spatial Image Data Mining Using K-Means Analysis: A Case Study of Uyo Capital City, Nigeria

[1]Christopher Ndehedehe, [2]Ogunlade Simeon,[1]Akwaowo Ekpa

1. Department of Geoinformatics & Surveying, Faculty of Environmental Studies, University of Uyo, Nigeria.
2.Department of Surveying and Geoinformatics, School of Environmental Technology, Federal University of Technology Akure, Nigeria.

Manuscript Info

Manuscript History:

Received:14 August 2013
Final Accepted: 22 August 2013
Published Online: September 2013

Key words:
Data mining, K-means,
Clustering,
Uyo capital city,
Landsat 7, KDD, GIS

Abstract

Data mining is the application of specific algorithms for extracting patterns from data. Different Data mining techniques have been used on large volumes of data to discover hidden patterns and relationships helpful in decision making.This work investigates the reliability of K-means, a popular and simplest unsupervised learning algorithm in Land Use Land Cover mapping of Uyo Capital City. The spatial subset of the classified imagery and the ground truth data sampled for this work was a 500m x 500m window. K-means classification was done using different iterations for the five clusters identified in the study area. The confusion matrix, overall accuracy and kappa coefficient results were good. The overall accuracies were 95.835% and 97.588% while the kappa coefficients were0.95 and 0.97 for 50 and 80 iterations respectively. The results were also confirmed by overlaying the various cluster groups with other validated data sources like Orthophoto and digitized vector of the same location. The use of K-means clustering analysis in land use classification may provide us with significant findings and reliable classification results like the supervised and machine learning algorithms. So far the results of K-means clustering is good enough when the exact number of clusters can be determined in the image and when very large areas are not sampled.

1.0 Introduction

Data mining is a technology used in different disciplines to search for significant relationships among variables in large data sets (Erdoğan and Timor, 2005).Data mining is the application of specific algorithms for extracting patterns from data. It is the practice of automatically searching large stores of data to discover patterns and trends that go beyond simple analysis. Data mining uses sophisticated mathematical algorithms to segment the data and evaluate the probability of several user-defined outcomes. Data mining is the extraction of hidden predictive information from large databases (Thearling, 1995).Data mining can answer questions that cannot be addressed through simple query and reporting techniques (Fayyad et al, 1996).

Data Mining, also popularly known as Knowledge Discovery in Databases (KDD), refers to the nontrivial extraction of implicit, previously unknown and potentially useful information from data in databases (Fayyad et al, 1994). While data mining and knowledge discovery in databases (or KDD) are frequently treated as synonyms, data mining is actually part of the knowledge discovery process.

1.1 Data Mining Techniques

Data mining techniques are used to operate on large volumes of data to discover hidden patterns and relationships helpful in decision making (e.g. Pandey and Pal 2011 and Alaa, 2009). Alternatively it has been called exploratory analysis, data driven discovery and deductive learning. A data mining algorithm is a well-defined procedure that takes data as input and produces output in the form of models or patterns (Erdoğan and Timor, 2005).

Data mining can be applied to a number of different applications such as data summarization, learning classification rules, finding associations, analyzing

changes and detecting anomalies (Han et al, 2006). Generally, the process of extracting patterns from data to solve a problem involves fourkey activities namely:Clustering, Classification, Regression,and Association Rule Learning (see e.g.Elena, 2001; Pandey and Pal, 2011). A brief discussion of Clustering and Classification is given below.

1.1.1 **Clustering:**Cluster analysis is a technique used in data mining. Cluster analysis involves the process of grouping objects with similar characteristics (Han et al 2001), and each group is referred to as a cluster.Data clustering is considered a data exploration technique that allows objects with similar characteristics to be grouped together in order to facilitate their further processing(Pham et al, 2004). Cluster analysis is used in various fields, such as marketing, image processing, geographical information systems, biology, and genetics.

1.1.2 **Classification:**Classification is a predictive data mining technique, makes predication about values of data using known results found from different data (see e.g.Pandey and Pal 2011, Margret, 2006). Predictive models have the specific aim of allowing us to predict the unknown value of a variable of interest given known values of other variables. Predictive modeling can be thought of as learning a mapping from an input set of vector measurements x to a scalar output y (e.g.Erdoğan, 2005; Margret, 2006). Classification maps data into predefined groups known as classes. It is often referred to as supervised learning because the classes are determined before examining the data. They often describe these classes by looking at the characteristic of data already known to belong to the classes (Margret, 2006).

2.0 Concept of Clustering

Clustering is the process of partitioning or grouping a given set of patterns into disjoint clusters (Alsabti et al, 1997). Cluster analysis is a multivariate analysis technique where individuals with similar characteristics are determined and classified (grouped) accordingly (Erdoğan, 2005). It is viewed as an unsupervised method for data analysis and classification (see e.g.Kiri et al, 2001, Yedla et al, 2010).Cluster analysis seeks to partition a given data set into groups based on specified features so that the data points within a group are more similar to each other than the points in different groups(see Deelers et al 2010, Margret, 2006).

Clustering therefore involves dividing a set of data points into non-overlapping groups, or clusters, of points, where points in a cluster are "more similar" to one another than to points in other clusters. The term "more similar," when applied to clustered points, usually means closer by some measure of proximity. When a dataset is clustered, every point is assigned to some cluster, and every cluster can be characterized by a single reference point, usually an average of the points in the cluster. Any particular division of all points in a dataset into clusters is called a partitioning.

3.0 K-means Clustering

The k-means method has been shown to be effective in producing good clustering results for many practical applications (Alsabti et al, 1997). K-means is one of the simplestunsupervised learning algorithms that solve t he well-known clustering problem. Numerous methods have been proposed to solve clustering problem. One of the most popular clustering methods is K-means clustering algorithm developed by Mac Queen in 1967. The easiness of K-means clustering algorithm made this algorithm used in several fields. The K-means clustering algorithm is a partitioning clustering method that separates data into k groups (see e.g.Fahim, et al 2006; Koheri et al 2007; Mc Queen, 1967; Yuan et al 2004 and Margret, 2006).The K-means clustering algorithm though computationally very expensive is more prominent because of its intelligence to cluster massive data rapidly and efficiently. Some authors have argued that the simplicity of the algorithm also can lead to some bad solutions (see e.g. Yedla et al, 2010, Alsabti et al, 1997), nevertheless a simple method for estimating the mean (vectors) of a set of K-groups. Various methods have been proposed to enhance the accuracy and efficiency of the k-means clustering algorithm (seeFahim,et al 2006;Kiri et al, 2001; Chen and Shixiong, 2009; Yedla et al, 2010 and Alsabti et al, 1997).

K-means algorithm has many advantages such as simplicity, low computational complexity, etc. (e.g. Ding et al, 2007) and it is widely used in remote sensing (e.g. Chehata et al, 2008; Maheshwary and Srivastav, 2008 and Zheng et al, 2008). However, K-means is sensitive to initialization (see Ding and Zhang, 2007; Yang, et al, 2010; Witten and Frank, 2005) and to the choice of the number of clusters, which usually is a critical issue and considered a drawback (Pham et al, 2004). Different random initializations of the cluster centres result in significantly different clusters at the convergence (Yang, et al, 2010). Thus, the algorithm is usually run many times with different initializations in an attempt to find a good solution (Wu, et al 2008). A fuzzy approach to data classification was suggested (see e.g. Dinesh et al, 1997; Borasca et al, 2006 and Yang, et al, 2010) as a better method in managing both the

uncertainty intrinsic in the classification problem and the relation one-to- many of a pattern with the related information classes.

The final clustering result of the K-means clustering algorithm greatly depends upon the correctness of the initial centroids, which are selected randomly (Yedla et al, 2010). It generates K points as initial centroids arbitrarily, where K is a user specified parameter. Each point is then assigned to the cluster with the closest centroid(see e.g. Chen and Shixiong, 2009, Yuan et al 2004, and Elmasri, et al, 2006). Then the centroid of each cluster is updated by taking the mean of the data points of each cluster. Some data points may move from one cluster to other cluster. Again we calculate new centroids and assign the data points to the suitable clusters. We repeat the assignment and update the centroids, until convergence criteria is met i.e., no point changes clusters, or equivalently, until the centroids remain the same. The K-means algorithm aims at minimizing an objective function know as squared error function given by:

$$J(V) = \sum_{i=1}^{c} \sum_{j=1}^{c_i} \left(\left\| x_i - v_j \right\| \right)^2$$

where

'$||x_i - v_j||$' is the Euclidean distance between x_i and v_j.

'c_i' is the number of data points in i^{th} cluster.

'c' is the number of cluster centers.

In this algorithm the Euclidean distance is used to find distance between data points and centroids (Nazeer and Sebastian, 2009).

Steps for K-means Clustering

Let $X = \{x_1, x_2, x_3, \ldots\ldots, x_n\}$ be the set of data points and $V = \{v_1, v_2, \ldots\ldots, v_c\}$ be the set of centers.

1) Randomly select *'c'* cluster centers.

2) Calculate the distance between each data point and cluster centers.

3) Assign the data point to the cluster center whose distance from the cluster center is minimum of all the cluster centers.

4) Recalculate the new cluster center using:

$$v_i = (1 / c_i) \sum_{j=1}^{c_i} x_i$$

Where, *'c_i'* represents the number of data points in i^{th} cluster.

5) Recalculate the distance between each data point and new obtained cluster centers.

6) If no data point was reassigned then stop, otherwise repeat from step 3).

In Seber, 1984 and Spath, 1985, K means uses a two-phase iterative algorithm to minimize the sum of point-to-centroid distances, summed over all k clusters: The first phase uses batch updates, where each iteration consists of reassigning points to their nearest cluster centroid, all at once, followed by recalculation of cluster centroids. This phase occasionally does not converge to solution that is a local minimum, that is, a partition of the data where moving any single point to a different cluster increases the total sum of distances. This is more likely for small data sets. The batch phase is fast, but potentially only approximates a solution as a starting point for the second phase. The second phase uses online updates, where points are individually reassigned if doing so will reduce the sum of distances, and cluster centroids are recomputed after each reassignment. Each iteration during the second phase consists of one pass through all the points. The second phase will converge to a local minimum, although there may be other local minima with lower total sum of distances. The problem of finding the global minimum can only be solved in general by an exhaustive (or clever, or lucky) choice of starting points, but using several replicates with random starting points typically results in a solution that is a global minimum.

The primary application of k-means is in clustering, or unsupervised classification (Jia Li, 2008). This study utilizes data mining in land use classification of Uyo capital city from low resolution satellite imagery (Landsat 7 imagery). K-means clustering analysis will be used as a data mining technique to cluster and classify the information classes present in the Landsat 7 imagery. The area of application is remote sensing, different from the usual data mining studies.

4.0 Study area

The area known as Uyo capital city lies within latitudes $4^0\,56^1\,30^{"}$ N and $5^0\,07^1\,40^{"}$ N, and longitudes $7^0\,49^1\,50^{"}$ E and $8^0\,01^1$ E. The present area of Uyo capital city is about 312.6 Sq. km with a population of about 3.9 million. The 1991 national population census puts Uyo population density of about 1,500 people per 1 Sq. km. Uyo capital city is originally a collection of villages, now almost seamlessly joined together to form the conurbation that it is today. A nucleated settlement pattern is exhibited in the area.

Uyo is the administrative and political capital of Akwa Ibom State, Nigeria. The area in the city can be categorized as residential, commercial and industrial based on the activities. Before now Most of the area in Uyo capital city can be classified as residential except for the commercial agglomeration in the business district. As the capital city is expanding, so new commercial zones are springing up to cater for the vast need of the growing population. It lies almost at the center of the state with roads linking all the local government areas in Akwa Ibom State, Nigeria.

4.1 Methodology

K-Means is an unsupervised classification method which calculates initial class means evenly distributed in the data space then iteratively clusters the pixels into the nearest class using a minimum distance technique. Each iteration recalculates class means and reclassifies pixels with respect to the new means. All pixels are classified to the nearest class unless a standard deviation or distance threshold is specified, in which case some pixels may be unclassified if they do not meet the selected criteria (Tou, and Gonzalez, 1974). This process continues until the number of pixels in each class changes by less than the selected pixel change threshold or the maximum number of iterations is reached.

The K-means clustering algorithm is a simple method for estimating the mean (vectors) of a set of K-groups. K-means is a nice method to quickly sort your data into clusters. All you need to know are the number of clusters you seek to find.Actually K-means clustering were used to generate class labels for the Landsat 7 imagery. In adopting k-means for land use classification, a fair idea of the available information classes is necessary.The K-means module implemented in ENVI 4.7 was used to cluster (group) the pixels in the Landsat imagery into their various land use classes. The results of the k-means classification were validated and viewed using an existing Orthophoto of the same location with other data sources in a geographic information system (GIS) environment.The performance matrix, overall accuracy and kappa coefficient was produced to evaluate the efficiency and reliability of this method.

4.2Data source

Collected maps and images (Orthophoto, Landsat 7 etc.) were sorted and classified for analysis and interpretation.Landsat 7 imagery of year 2000 and digitsed vector from an existing Orthophoto of the same year were employed in this study to produce land use/cover categories of 2000. Reflective bands 1, 2, 3, 4, 5, and 7 of each image scene were stacked and used in an image-to-image geometric projection, using the 2000 image as master.

4.3Data Preparation

In the present study a processed geo-referenced remotely sensed data was used as a base for image registration. Images were traced from Landsat 7 of year 2000. The standard image processing techniques such as, image extraction, rectification, and restoration, were applied in this work. The image obtained were made up of three bands, viz., Band 2 (visible), Band 4, and Band 7 (infrared) were used to create a False Color Composite (FCC) as shown in Figure 4.1.

This combination provides a "natural-like" rendition, while also penetrating atmospheric particles and smoke. This combination brings out urban areas in varying shades of magenta, Grasslands/agricultural areas appear as light green, forested areas are Olive-green to bright-greenhues. Sparsely vegetated areas appear as oranges and browns, cultivated areas/burnt areas appear as red etc.Pattern recognition helps in finding meaningful patterns in data. Spectral pattern recognition can be improved through Digital image processing as mentioned earlier. The red, green and blue(RGB)composite of band 742 was used for the K-means clustering in ENVI 4.7.

5.0 Analysis and Results

The K-Means unsupervised classifier uses a cluster analysis approach which requires the analyst to select the number of clusters to be located in the data, and arbitrarily locates this number of cluster centres, and then iteratively repositions them until optimal spectral separability is achieved. The spatial subset of the classified imagery and the ground truth data was a 500mx 500m window. The results of the K-means classification with 50 and 100 iterationsis shown in figure 4.2a and 4.2b respectively.

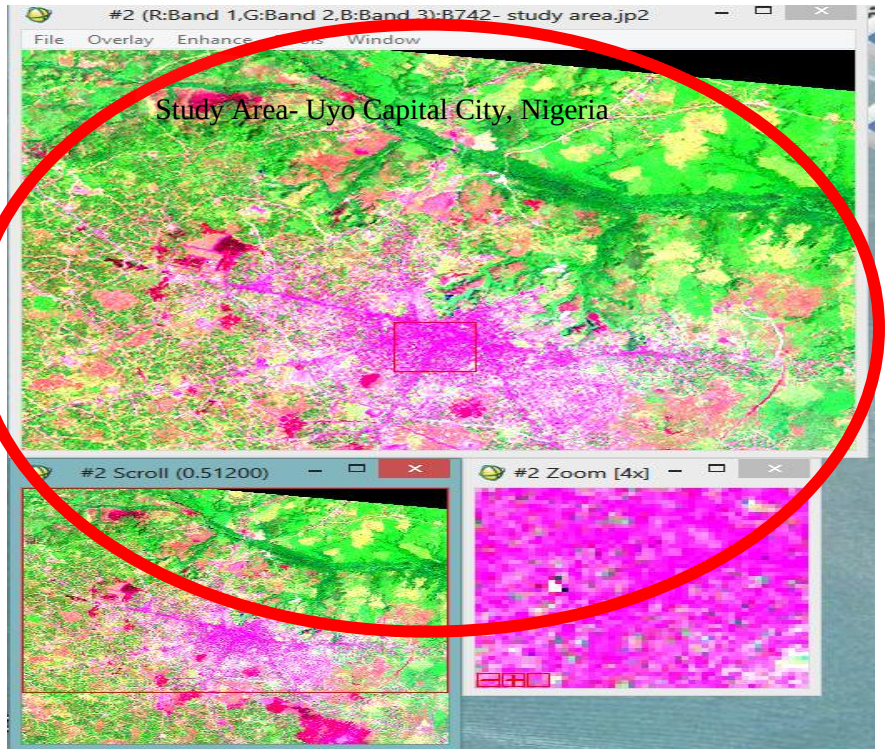
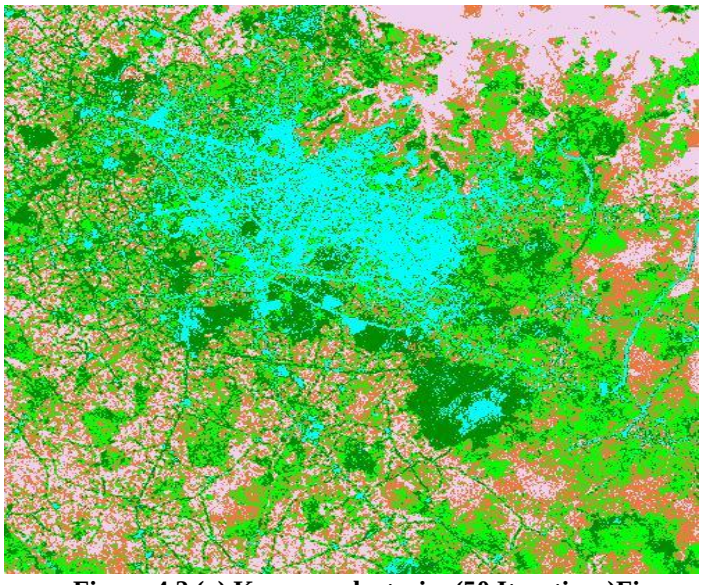


Figure 4.1 RGB composites from Landsat 7



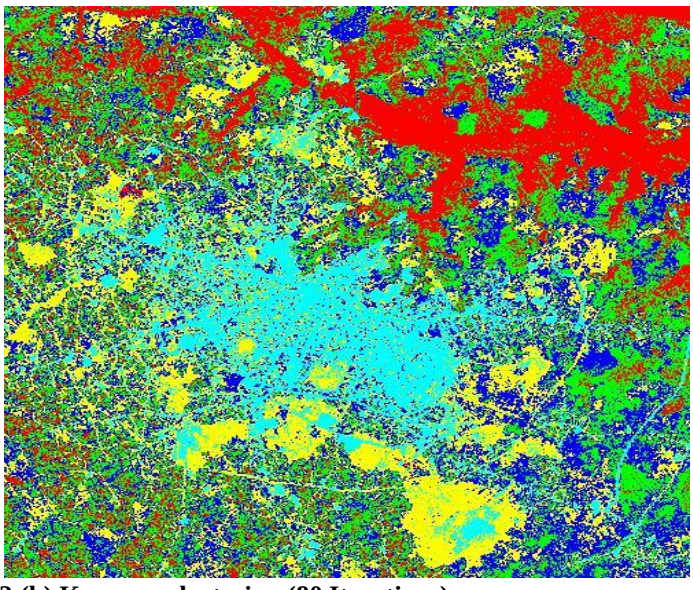
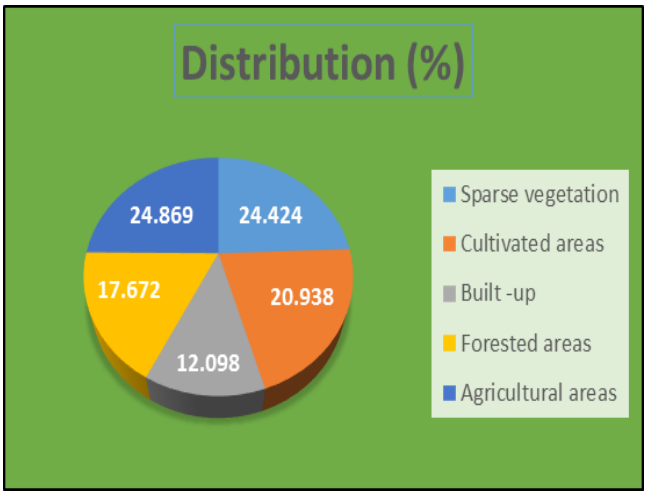


Figure 4.2 (a) K-means clustering(50 Iterations)Figure 4.2 (b) K-means clustering (80 Iterations)

Built -up
Cultivated areas
Agricultural areas
Sparse vegetation
Forested areas

Built -up
Sparse vegetation
Cultivated areas
Forested areas
Agricultural areas

Cluster Distribution of the Pixels for the Different Iterations Used



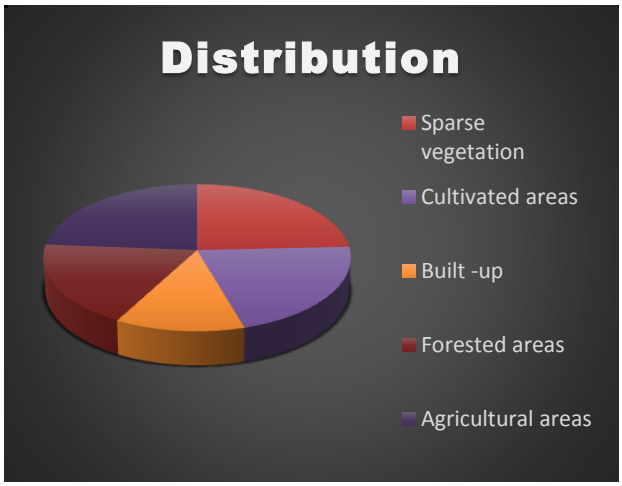


Table 5.1

Land Use map classification (50 Iterations)Land Use Map Classification (80 Iterations)

Class	*Producer Accuracy (%)*	*User Accuracy(%)*	*Class*	*Producer Accuracy (%)*	*User Accuracy (%)*
Built -up	96.61	100.00	Built -up	98.95	100.00
Cultivated areas	98.21	97.75	Cultivated areas	98.68	99.30
Agricultural areas	95.03	98.20	Agricultural areas	96.85	98.75
Sparse vegetation	91.26	94.64	Sparse vegetation	94.86	96.82
Forested areas	100.00	87.59	Forested areas	100.00	93.23

Table 5.2*Overall Accuracy and Kappa Coefficient Result*

No of Iterations	*Overall Accuracy*	*Kappa Coefficient*	*Correctly Placed Pixel*	*Misplaced Pixel*	*Total Pixels*
50	95.8355%	0.9474	193790	8421	202211
80	97.5884%	0.9695	227869	5631	233500

Table 5.3aConfusion Matrix Table for 50 Iterations

Ground Truth (Pixel)

Class	***Distribution(%)***	***Sparse vegetation***	***Cultivated areas***	***Built-up***	***Forested areas***	***Agricultural areas***	***Total***
Sparse vegetation	24.424	42635	0	0	0	2417	***45052***
Cultivated areas	20.938	0	45598	988	0	63	***46649***
Built -up	12.098	0	0	29293	0	0	***29293***
Forested areas	17.672	4083	0	0	28816	0	***32899***
Agricultural areas	24.869	0	829	41	0	47448	***48318***
Total	***100***	***46718***	***46427***	***30322***	***28816***	***49928***	***202211***

The error matrices and global performance indices obtained by applying K-means clustering is shown in table 5.3a and 5.3b. It is obvious that higher iterations can improve the overall accuracy and kappa coefficient-See table 5.2. There is confusion when discriminating agricultural field from built up regions. This is possible due to the open space within urban areas used as parks, recreational centres etc.The same situation is observed in sparse vegetation, forested areas and cultivated areas. This possibly could be due to chlorophyll content and water present in those pixels. The results were also confirmed by overlaying the various cluster groups with other validated data sources like Orthophoto, digitized vector etc. of the same location.Figure 4.3 is a map that graphically demonstrates the performance of the K-means classifier. The outcome was good, thus pointing out the reliability of K-means clustering in remote sensing applications.

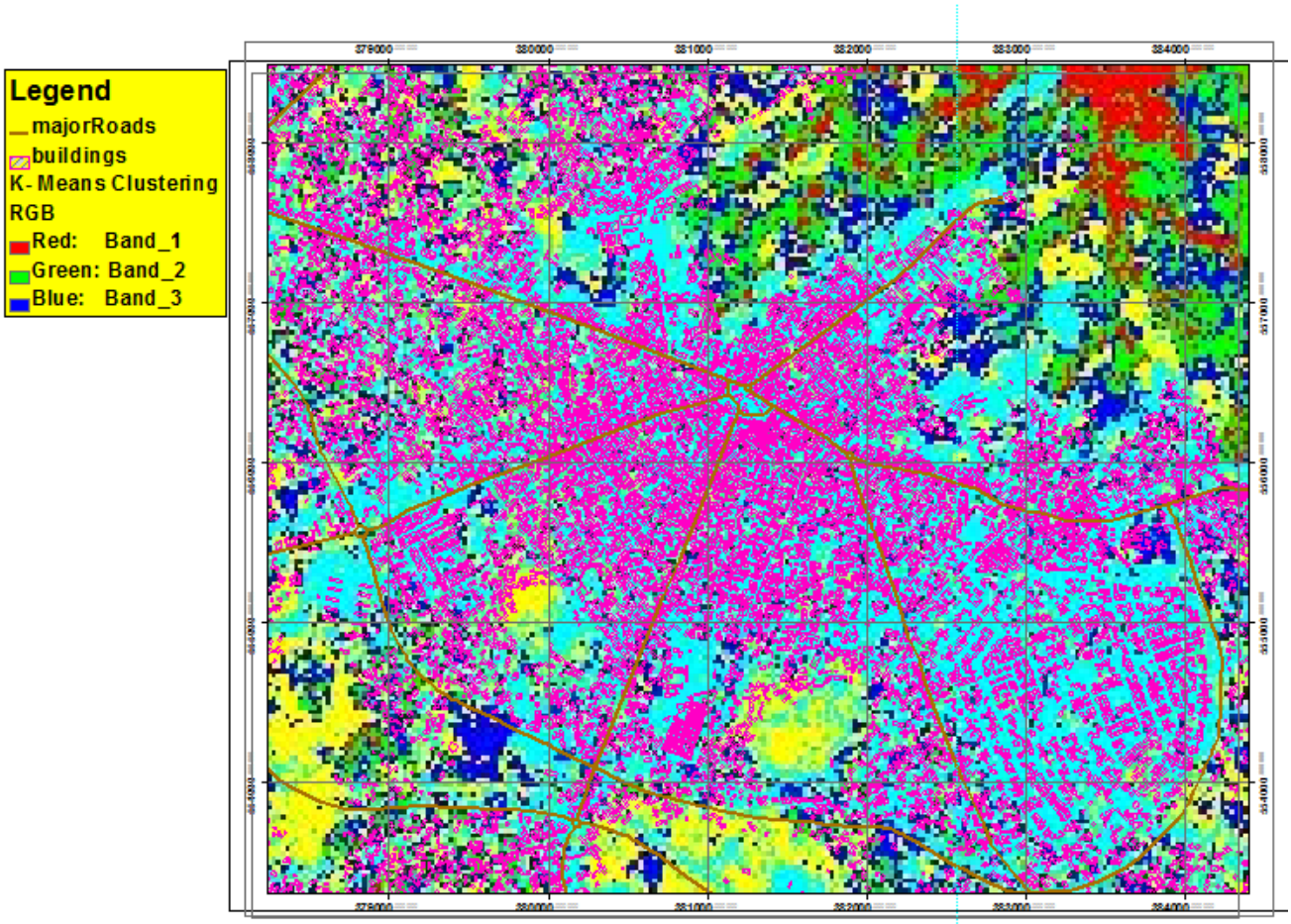


Figure 4.3 Map Showing Digitized Vector Overlay of Uyo Capital City, Nigeria with K-Means Clustering Of Landsat 7 Imagery

Table 5.3b Confusion Matrix Table for 80 Iterations

Ground Truth (Pixel)

Class	*Distribution(%)*	*Sparse vegetation*	*Cultivated areas*	*Built -up*	*Forested areas*	*Agricultural areas*	*Total*
Sparse vegetation	24.263	52703	0	0	0	1733	***54436***
Cultivated areas	21.218	0	50716	319	0	36	***51071***
Built -up	12.385	0	0	30782		0	***30782***
Forested areas	18.349	2858	0	0	39364	0	***42222***
Agricultural areas	23.785	0	676	9	0	54304	***54989***
Total	***100***	***55561***	***51392***	***31110***	***39364***	***56073***	***233500***

Conclusion

This study utilizes data mining in land use classification of Uyo capital city from low resolution satellite imagery. K-means clustering analysis was used as a data mining technique. The area of application was remote sensing, different from the usual data mining studies. K-means clustering was used to generate class labels for the Landsat 7 imagery. The use of K-means clustering analysis in land use classification may provide us with significant findings, and may lead to reliable classification results like the supervised and machine learning algorithms. So far the result of K-means clustering is good enough when the number of clusters is determined in the image and when very large areas are not sampled.

References

1. Alaa el-Halees 2009: *Mining Students Data to Analyze e-learning Behavior*: A Case Study.
2. Alsabti, K., Ranka, S. and Singh, V. 1997:*An Efficient K-Means Clustering Algorithm.* http://www.cise.ufl.edu/ranka/.
3. Borasca, B. L., Bruzzone, L. C., and Zusi,M. 2006: *A fuzzy-input fuzzy- output SVM technique for classification of hyperspectral remote sensing images*. In Proc. 7th NORSIG, 2006, pp. 2–5.
4. Chen Z. and Shixiong, X. 2009: *K-means Clustering Algorithm with Improved Initial center,* in Second International Workshop on Knowledge Discovery and Data Mining (WKDD), pp. 790-792, 2009
5. Chehata, N. and Bretar, F. 2008: *Terrain modeling from lidar data: Hierarchical K-means filtering and Markovian regularization*, in Proc. IEEE ICIP, pp. 1900–1903
6. Deelers, S. and Auwatanamongkol, S. 2010: *Enhancing K-Means Algorithm with Initial Cluster Centers Derived from Data Partitioning along the Data Axis with the Highest Variance*. International Journal of Computer Science, Vol. 2, Number 4.
7. Dinesh, M. S. K.,Chidananda, G. and Nagabhuehan,P. 1997: *Unsupervised classification for remotely sensed data using fuzzy set theory.* In Proc. IEEE IGARSS—A Scientific Vision for Sustainable Development, 1997, vol. 1, pp. 521–523.
8. Ding, Z. J.,Yu, J. and Zhang,Y. Q. 2007: *A new improved K-means algorithm with penalized term,* in Proc. IEEE ICCC, 2007, p. 313.
9. Elena S. 2001:*Using data mining techniques in higher education*. National defence university "carol I" Bucharest 68-72.
10. Elmasri, N. and Gupta, S. 2006: *Fundamentals of Database Systems*, Pearson Education, First edition, 2006.
11. Erdoğan, Ş.and Timor, M., 2005: *A Data Mining Application in a Student Database.* Journal of aeronautics and space technologies, July 2005 Volume 2 Number 2 (53-57)
12. Fahim, A. M., Salem, A. M.,Torkey,F. A. and M. A. Ramadan, 2006. *An Efficient enhanced k-means clustering algorithm,* journal of Zhejiang University, 10(7): 16261633.
13. Fayyad, U.Piatetsky-Shapiro, G. and Smyth,Padhraic 1996*: From Data Mining to Knowledge Discovery in Databases.* AI Magazine Volume 17 Number 3 (1996) (© AAAI)
14. Fayyad, U.M., Piatesky-Shapiro, G., Smyth, P., Uthurusamy, R., 1994:*Advances in data mining and knowledge discovery.*MIT Press, USA,
15. Jia Li 2008:*Prototype Methods: K-Means*. Department of Statistics the Pennsylvania State University. http://www.stat.psu.edu/~jiali
16. Han, J., Kamber, W., 2001:*Data Mining Concepts and Techniques*, Morgan Kaufmann Publishers, USA, 5-10.
17. Han, J. W., Kamber, M., 2006:*Data Mining: Concepts and Techniques*, 2nd Edition, the Morgan Kaufmann Series in Data Management Systems, Gray, J. Series Editor, Morgan Kaufmann Publishers.
18. Kiri,W., Claire C., Seth, R. and Stefan S. 2001:*Constrained K-means Clustering with Background Knowledge.* Proceedings of the Eighteenth International Conference on Machine Learning, 2001, p. 577–584.
19. Koheri A. and Ali R. 2007:*Hierarchical K-means: an algorithm for Centroids initialization for k-means.* department of information science and Electrical Engineering Politechnique in Surabaya, Faculty of Science and Engineering, Saga University, Vol. 36, No.1,
20. Maheshwary, P. and Srivastav N. 2008: *Retrieving similar image using color moment feature detector and K-means clustering of remote sensing images*, in Proc. IEEE Int. Conf. Comput. Elect. Eng., Dec. 20–22, 2008, pp. 821–824.

21. Margaret, H D. 2006: *Data Mining-Introductory and Advanced Concepts*, Pearson Education, 2006.
22. Mc Queen J, 1967:*Some methods for classification and analysis of multivariate observations*. Proceedings. 5th Berkeley Symposium. Math. Statist. Prob., (1): 281–297.
23. Nazeer, K. A. A. and Sebastian, M. P. 2009: *Improving the accuracy and efficiency of the k-means clustering algorithm,*in International Conference on Data Mining and Knowledge Engineering (ICDMKE), Proceedings of the World Congress on Engineering (WCE-2009), Vol 1, July 2009, London, UK.
24. Pandey, U. K. and Pal S. 2011: Data Mining: *A prediction of performer or underperformer using classification*
25. Pham, D T., Dimov, S. S. and Nguyen,C. D. 2004: *Selection of K in K-means clustering*. Proc. Int. Mech Eng. Vol. 219 Part C: J. Mechanical Engineering Science. DOI: 10.1243/095440605X8298
26. Seber, G. A. F. 1984:*Multivariate Observations*, Wiley, New York.
27. Spath, H. 1985: *Cluster Dissection and Analysis*: Theory, FORTRAN Programs, Examples, translated by J. Goldschmidt, Halsted Press, New York.
28. Thearling, K. 1995:*An Introduction to Data Mining*. Discovering hidden value in your data warehouse. DIG White Paper 95/02 October 1995.
29. Tou, J. T. and Gonzalez, R. C. 1974:*Pattern Recognition Principles*, Addison-Wesley Publishing Company, Reading, Massachusetts.
30. Witten, L. H. and Frank E. 2005: *Data Mining: Practical Machine Learning Tools and Techniques*, 2nd ed. New York: Elsevier.
31. Wu, X. D., Kumar, V., Quinlan, J. R.,Ghosh, J. Q., Yang, H.,Motoda, G. McLachlan,J., Ng,A.,Liu, B.,Yu, P. S., Zhou, M., Steinbach, D.,Hand, J. and Steinberg,D. 2008: *Top 10 algorithms in data mining.* Knowledge Information System, vol. 14, no. 1, pp. 1–37.
32. Yang, C.,Bruzzone,L.,Sun, F.,Lu, L. Guan, R. and Liang, Y. 2010: *A Fuzzy-Statistics-Based Affinity Propagation Technique for Clustering in Multispectral Images.* IEEE Transactions On Geoscience And Remote Sensing, VOL. 48, NO. 6, JUNE 2010. Page 2647-2659.
33. Yedla M., Srinivasa,R., and Srinivasa,T. M.2010:*Enhancing K-means Clustering Algorithm with Improved Initial Center*. International Journal of Computer Science and Information Technologies, (IJCSIT) Vol. 1 (2) , 2010, 121-12
34. Yuan, F. Z.,Meng, H. H. X. Zhangz, C., andDong, R. 2004. *A New Algorithm to Get the Initial Centroids.* Proceedings of the 3rd International Conference on Machine Learning and Cybernetics, pp. 26-29.
35. Zheng, J. Z. Z., Cui, A.,Liu, F. and Jia, Y. 2008: *A K-means remote sensing image classification method based on AdaBoost*, in Proc. IEEE ICNC pp. 27–32.

****