Unsupervised Classification of Sentinel-2 Satellite Imagery

(Hungarian: "Sentinel-2 műholdfelvételek felügyelet nélküli osztályozása")

Author:

Godwin Emmanuel Bandawa

Consulents:

Ervin Wirth & Tamás Lovas

2020, Budapest

Contents

ABSTRACT

An effective approach for monitoring global changes is feasible via the use of satellite images. These images provide important information that aids in observing various applications such as change detection, monitoring disasters, natural hazards, or land cover classification. Remote sensing is a key technique used to obtain information related to the Earth's resources and environment.

The Sentinel-2 satellites provide a global coverage of land surfaces with a 5-day revisit time at the equator. These multi-spectral instruments mainly applied in agriculture, such as crop monitoring and management, vegetation and forest monitoring, tracking land cover change for environmental analysis, observation of coastal zones, inland water and glacier monitoring, ice extent, and snow cover mapping.

In this study, high-resolution satellite images were obtained from the Sentinel-2 Copernicus Open Access Hub, and an unsupervised classification algorithm was applied on the images. The K-means clustering algorithm is used particularly as it delivers training results quickly where images are unlabeled. The aim of this study is to investigate sequential (time series) satellite image classifications of Sentinel-2, after running the cluster algorithm on these images relevant conclusions are being made: like the cluster of vegetation changes, calculating surface differences, and changes in land use over time.

Keywords: remote sensing, classification, land use and land cover, unsupervised classification, algorithm, time series analysis, urban sprawl.

1.1 INTRODUCTION

Cities across the world are experiencing rapid changes in landscape and land quality. The changes are originated from the way humans put their land to use and from climatic variability, for example, drought. This has led to desert like conditions in several areas of northern Nigeria causing loss of arable and grazing land and consequently population migration, conflict over the limited resources and economic loss to the people in the area [1]. An important use of satellite observations has been to monitor the changes on the earth surface, new sources of spatial data and innovative techniques offer the potential to significantly improve the analysis, understanding, presentation and modeling of urban dynamics based on remotely sensed data [2]. Remote sensing is acquisition of information without making actual physical contact. It is extraction of some meaningful information in digital form using indirect measurements [3] usually by sensors mounted on terrestrial, airborne, or satellite platforms and this serves as one main data for GIS. High resolution aerial images support a wide range of application fields such as biomass estimation for energy studies, water analysis for pollution detection, environment and ecology investigations, and urban sprawl assessment [4]. Land-cover (LC) and land-use (LU) change information is important because of its practical uses in various applications, including deforestation, damage assessment, disasters monitoring, urban expansion, planning, and land management [5].

If land-use change models are to be used in support of various decision making processes, characterization of urban growth and landscape change involves the procedures of monitoring and modeling, which further require reliable, information based, and robust analytical techniques [6], [7]. However, there is little research conducted on this topic in the particular area of interest suggested in this study, thus the motivation for selecting this region, and also expand this topic at a local scale. In Nigeria, land cover data for natural resource management is missing and where they do exist, they are out of date. Thus, because of the lack of such data, professionals such as town planners and resource managers often make decisions based on false assumptions [1].

In the coming decades most of the world's land cover and land use change (LCLUC) is predicted to take place in the tropics, where population is growing the fastest. United Nations' projections estimate that virtually all of the world's population between now and the middle of this century will emerge in the cities of the developing world, driven by natural increase in both urban and rural areas, and by continued migration from rural to urban areas as people search for economic opportunities. Urbanization is shaping landscapes in and around cities through densification and sprawl, while at the same time increased interaction among cities is creating new hybrid landscapes where rural and urban livelihoods overlap [8], the extensive urbanization induces contentious of high land consumption process, which is the ratio of population and urban area at a given time [9].

The Copernicus Program is an ambitious initiative headed by European Commission in partnership with the European Space Agency (ESA). The Sentinels are a constellation of satellites developed by ESA to operationalize the Copernicus program, which include all-weather radar images from Sentinel-1A and 1B, and high-resolution optical images from Sentinel-2A and 2B [10].

Digital classification of multispectral satellite images is commonly used to obtain information on land cover. In land cover classification, the goal is to obtain relatively few classes. However, a large number of spectral values from individual bands is typically found in the images. The aim of classification techniques is thus to reduce the large number of individual combination to a small number of classes, among the various classification approaches, unsupervised image classification methods are designed to make the best possible use of the overall spectral content of an image [11].

Alternatively, unsupervised learning approach can be applied in mining; image similarities can be derived directly from the image collection, hence inherent image categories can be identified from the image set [12]. Figure 1.1 below illustrates typical unsupervised classification process. Unsupervised classification is a form of pixel based classification and is essentially computer automated classification. The user specifies the number of classes and the spectral classes are created solely based on the numerical information in the data (i.e. the pixel values for each of the bands or indices), clustering algorithms are used to determine the natural, statistical grouping of the data; the pixels are grouped together based on their spectral similarity [13].

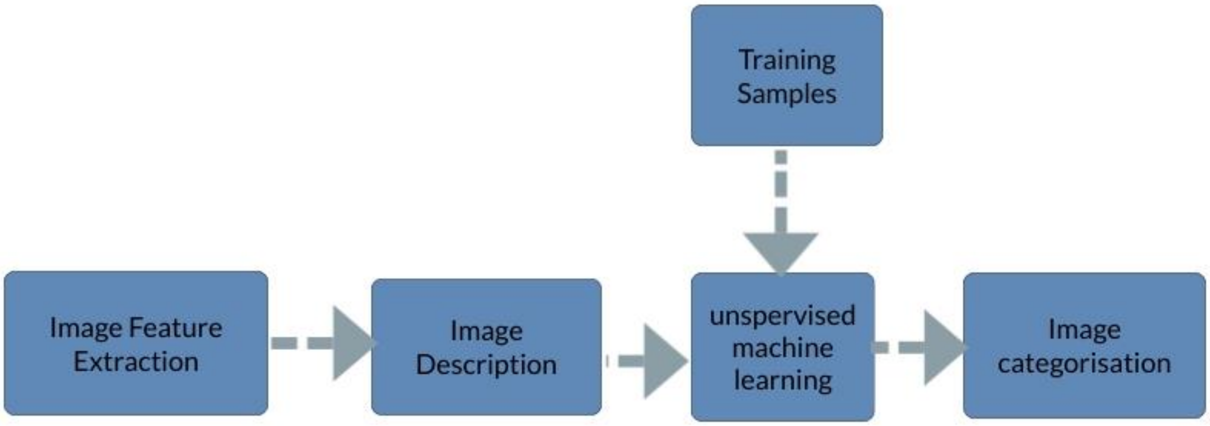

Figure 1.1 Typical unsupervised classification

2. METHODOLOGY

The aim of the methodology used in this study is to be able to acquire meaningful spatial information and evaluate results obtained from the process. The research design as illustrated in fig 2.1 was employed.

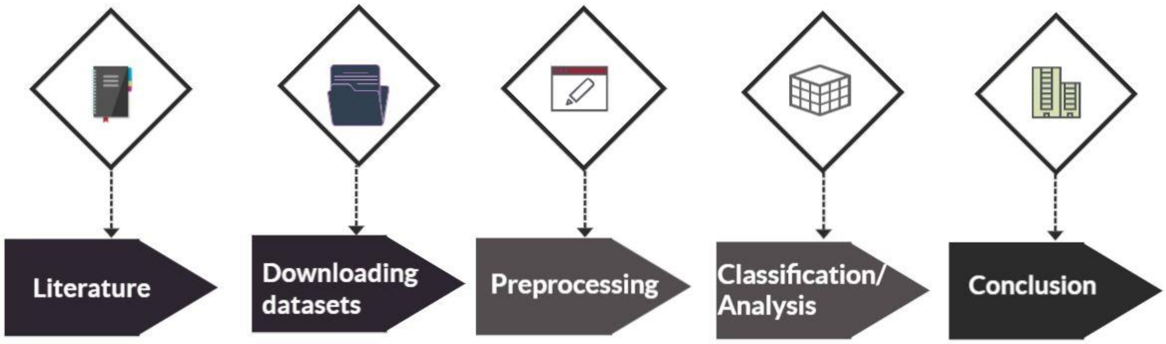

Fig 2.1 Research design/procedure

An evaluation of scientific literatures was carried out on available materials in relation to the topic, Area of Interest (AOI) and based on selected keywords. The literature used were sourced from journal publishing sites and books published about the topics, however, there are only a few or outdated sources available on the selected AOI.

Additionally, I downloaded the High-Resolution satellite image used in this study from the Sentinel-2 Copernicus Open Access Hub by signing up through the United States Geological Survey (USGS) website; the images were obtained within several time intervals to enable the time series analysis. The downloaded images consist of several bands of Sentinel-2 images, these images were further processed to prepare a final input for the classification algorithm. Three different bands were merged together using the QGIS software (a free and open-source cross-platform desktop geographic information system application- and the AOI was clipped out from the merged image tile and saved as a GeoTIFF file.

This library was used to load the raster satellite images into an array like dataset.

Thus, unsupervised classification algorithm was run on the datasets of the final clipped images. The K-means clustering algorithm was implemented using python sciKit-learn library, Numpy library and matplotlib for Visualization. Results from the image classification was analyzed and used to deduce conclusions and propose a direction for future research.

3.1 STUDY AREA AND DATA SOURCE

The study area, which is Jimeta-Yola, is located in the Northeastern part of Nigeria. Jimeta is the largest city, capital city and administrative center of Adamawa State, Nigeria. Located on the Benue River, it has a population of 336,648 (2010) [14]. The town is served by the port of Jimeta (5.5 miles [9 km] north-northwest) on the Benue River, about 500 miles (800 km) north to its confluence with the Niger, and by an airfield [15] (Fig. 3.1).

Adamawa State lies between latitudes 7° 00’N – 11° 00’N and longitude 11°’ 00’E – 14° 00’ E, while the study area approximately lies between latitude 09° 15’N and 9° 16’ N and longitude 12° 25’ E and 12° 24’ E. Like any other Nigerian city, Jimeta comprises of so many land use types ranging from institutional, commercial, and residential (Fig. 3.2). The city is clearly stratified in terms of population densities. These are low, medium and high-density areas. The low density areas are well planned units where government officials reside while medium and high density areas are made up of people with little or unplanned buildings. Land uses in Jimeta can be classified into six categories as study has shown, the results from aerial photographs indicated that six land use/cover were identified thus; built-up land, bare surface, natural vegetation or forest, marshy land, croplands, and water bodies [16].

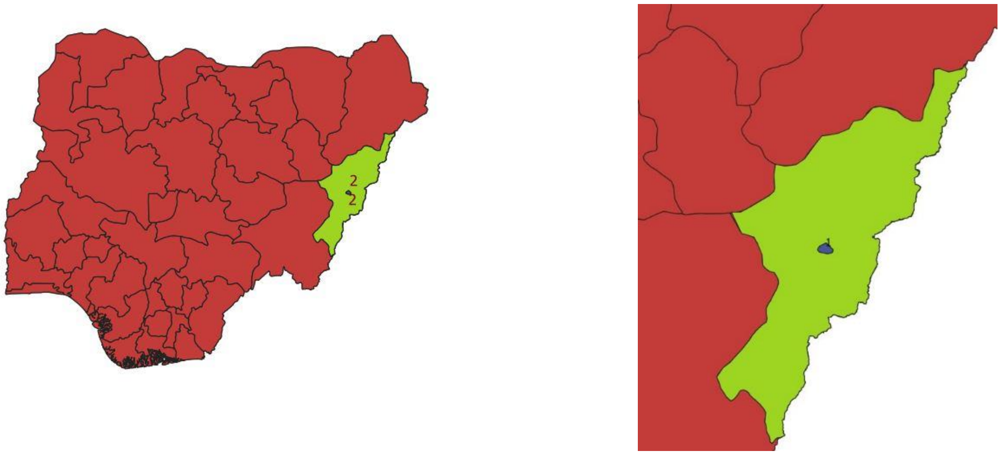

Fig 3.1. Map of Nieria Adamawa state and approximate location of AOI.

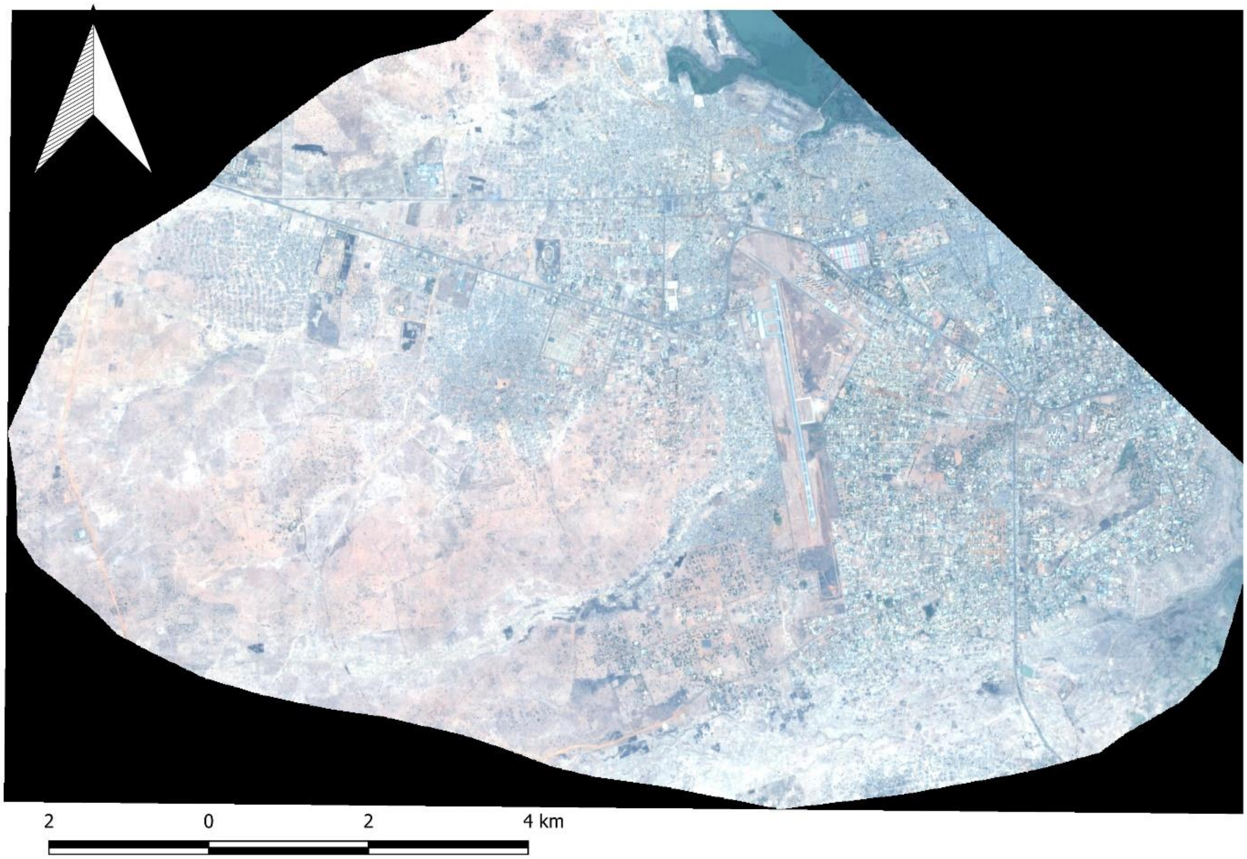

Fig 3.2 Satellite image of AOI.

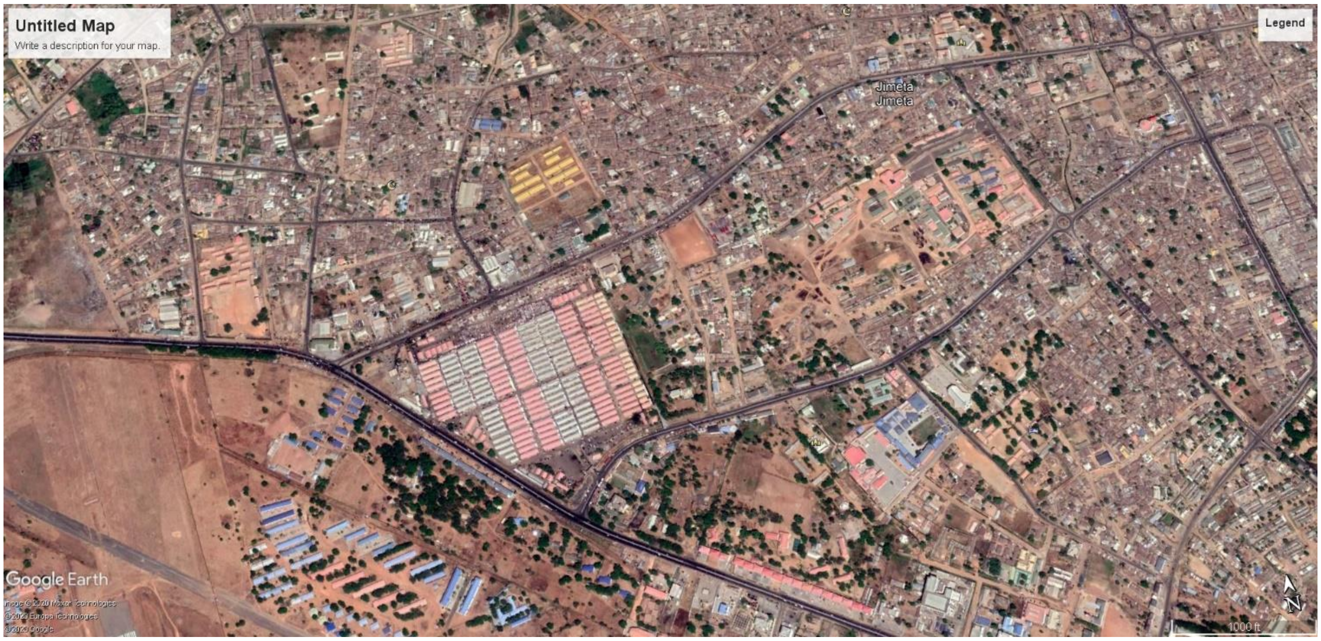
Untitled Map
Write a description for your map.
Legend
Jimeta
Jimeta
Google Earth
Image © 2020 Maxar Technologies
© 2020 Europa Technologies
© 2020 Google
1000 ft

Fig3.2a

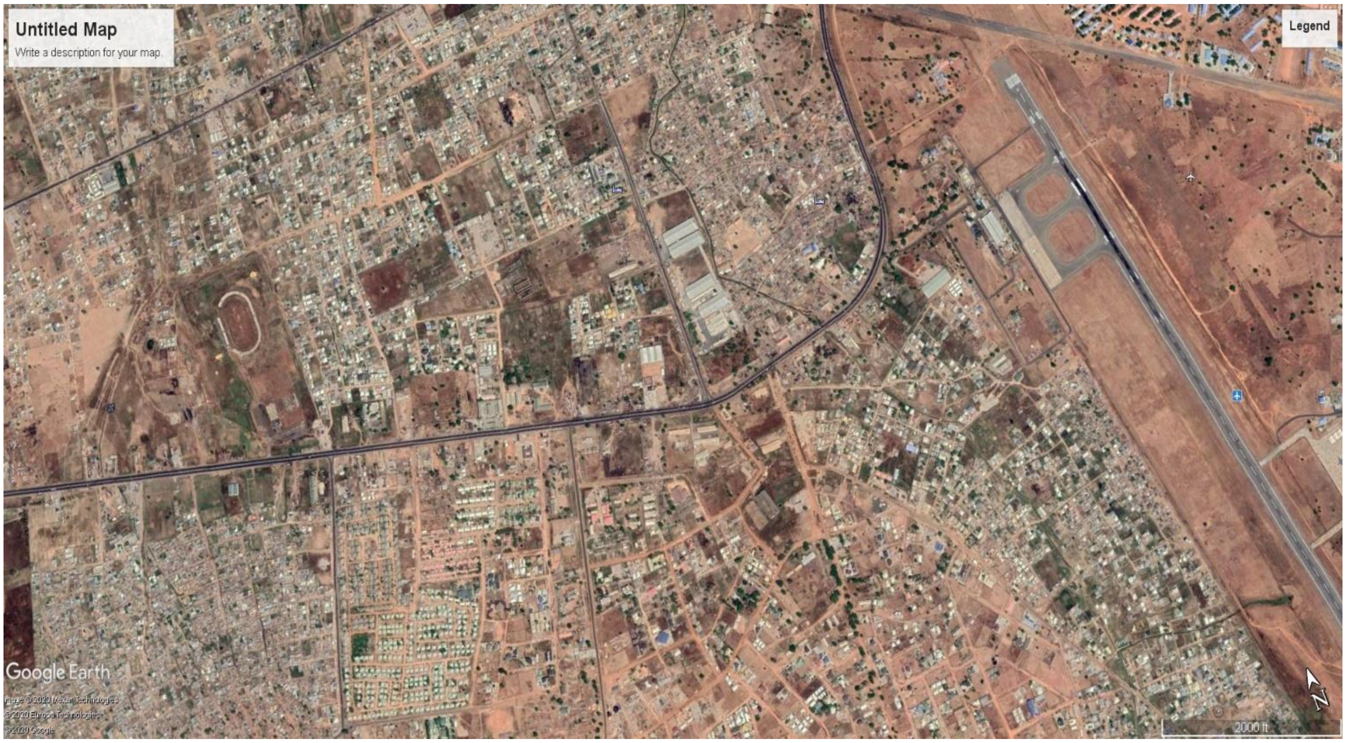
Untitled Map
Write a description for your map.
Legend
Google Earth
Image © 2020 Maxar Technologies
© 2020 Europa Technologies
© 2020 Google
N
2000 ft

Fig3.2b

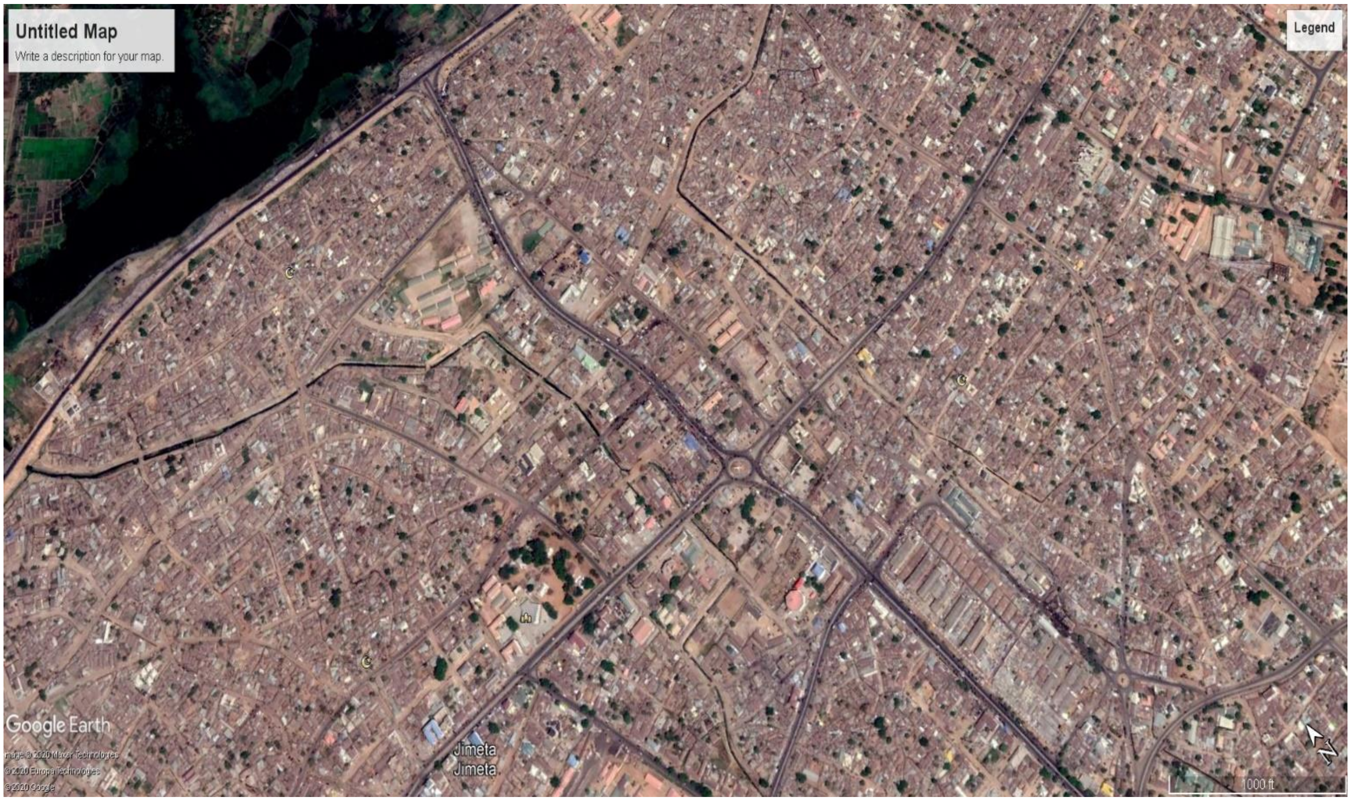
Untitled Map
Write a description for your map.
Legend
Jimeta
Jimeta
Google Earth
Image © 2020 Maxar Technologies
© 2020 Europa Technologies
© 2020 Google
N
1000 ft

Fig3.2c

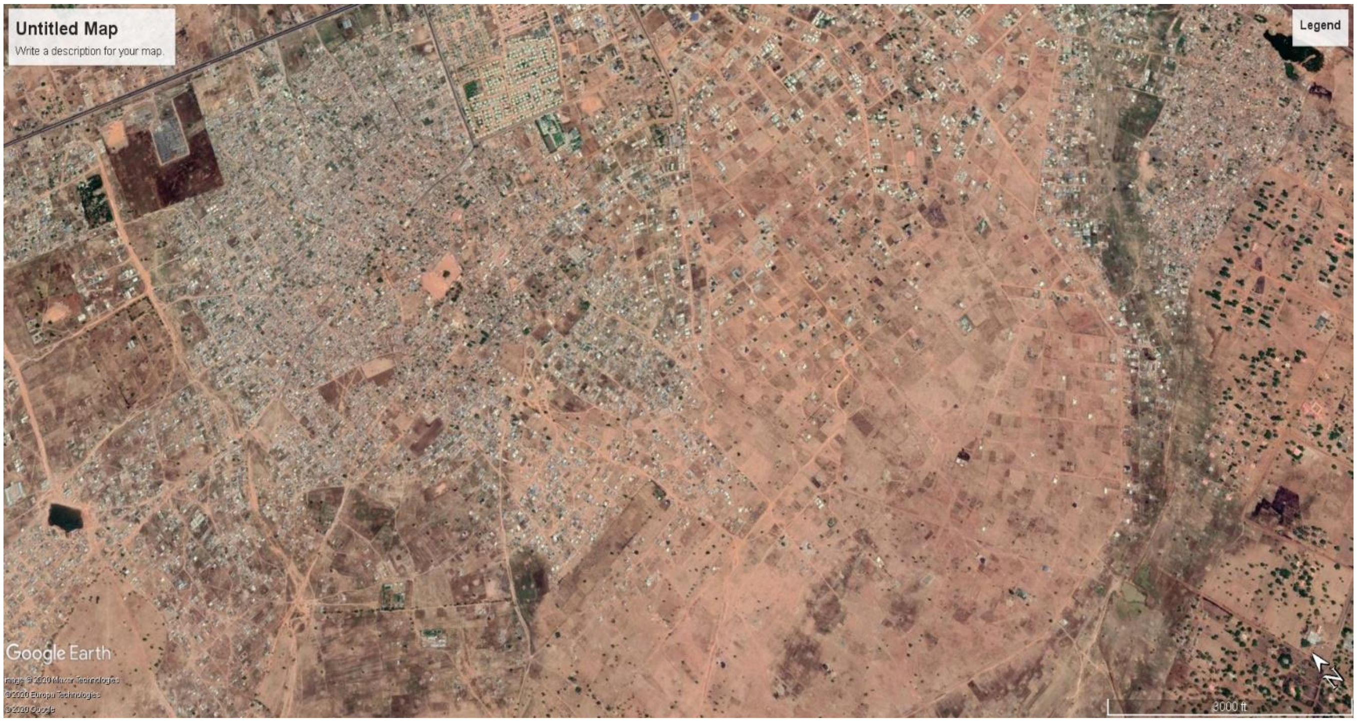
Untitled Map
Write a description for your map.
Legend
Google Earth
Image © 2020 Maxar Technologies
© 2020 Europa Technologies
© 2020 Google
3000 ft

Fig3.2d

Fig 3.2a – fig 3.2d shows a closer look at the AOI

3.2 DATA SOURCE

For the purpose of this research Sentinel-2 Images were downloaded through the USGS. The satellite is equipped with an opto-electronic multispectral sensor for surveying with a resolution of 10 to 60 m in the visible, near infrared (VNIR), and short-wave infrared (SWIR) spectral zones, including 13 spectral channels, which ensures the capture of differences in vegetation state, including temporal changes, and also minimizes impact on the quality of atmospheric photography.

The orbit is an average height of 785 km. The Sentinel-2 mission consists of two satellites developed to support vegetation, land cover, and environmental monitoring. The Sentinel-2A satellite was launched by ESA on June 23, 2015, and operates in a sun-synchronous orbit with a 10-day repeat cycle. A second identical satellite (Sentinel-2B) was launched March 7, 2017 and is operational with data acquisitions available on EarthExplorer. Together they cover all Earth’s land surfaces, large islands, and inland and coastal waters every five days.

The Sentinel fleet of satellites is designed to deliver land remote sensing data that are central to the European Commission’s Copernicus program. The MultiSpectral Instrument (MSI) sensor data are complementary to data acquired by the U.S. Geological Survey (USGS) Landsat 8 Operational

Land Imager (OLI) and Landsat 7 Enhanced Thematic Mapper Plus (ETM+). A collaborative effort between ESA and the USGS provides for the public access and redistribution of global acquisitions of ESA's Sentinel-2 data at no cost through secondary U.S.-based portals, in addition to direct user access from ESA [17], [18].

Sensor	Band number	Band name	Sentinel-2A		Sentinel-2B		Resolution (meters)
			Central wavelength (nm)	Bandwidth (nm)	Central wavelength (nm)	Bandwidth (nm)	
MSI	1	Coastal aerosol	443.9	20	442.3	20	60
MSI	2	Blue	496.6	65	492.1	65	10
MSI	3	Green	560.0	35	559	35	10
MSI	4	Red	664.5	30	665	30	10
MSI	5	Vegetation Red Edge	703.9	15	703.8	15	20
MSI	6	Vegetation Red Edge	740.2	15	739.1	15	20
MSI	7	Vegetation Red Edge	782.5	20	779.7	20	20

Sensor	Band number	Band name	Sentinel-2A		Sentinel-2B		Resolution (meters)
			Central wavelength (nm)	Bandwidth (nm)	Central wavelength (nm)	Bandwidth (nm)	
MSI	8	NIR	835.1	115	833	115	10
MSI	8b	Narrow NIR	864.8	20	864	20	20
MSI	9	Water vapour	945.0	20	943.2	20	60
MSI	10	SWIR – Cirrus	1373.5	30	1376.9	30	60
MSI	11	SWIR	1613.7	90	1610.4	90	20
MSI	12	SWIR	2202.4	180	2185.7	180	20

Table3.1 SENTINEL-2 BANDS

3.3 PREPROCESSING OF SATELLITE IMAGES

After the images were downloaded it was processed using the QGIS software program, the selected bands which are bands 2,3,4,8 (Blue, Green, Red, and Near Infrared respectively) were merged together; hence, a shapefile dataset for the AOI was downloaded from gadm.org (the Database of Global Administrative Areas, (GADM) provides maps and spatial data for all countries and their sub-divisions) [19]. The AOI was clipped out of the new merged raster dataset using the downloaded dataset.

Furthermore, The Geospatial Data Abstraction Library (GDAL) was implemented to read the dataset of the merged raster into an array like dataset before running the classification algorithm. GDAL is a translator library for raster and vector geospatial data formats that is released under an X/MIT style Open Source License by the Open Source Geospatial Foundation. As a library, it presents a single raster abstract data model and single vector abstract data model to the calling application for all supported formats. It also comes with a variety of useful command line utilities for data translation and processing [20].

4. SATELLITE IMAGE CLASSIFICATION

After downloading and processing the sentinel-2 images, the classification algorithm was further implemented on the images. Downloaded datasets have an interval of two years. Python programming language was used during the analysis on the dataset, Libraries used include GDAL for reading and writing the image datasets, Numpy which contains a multi-dimensional array and matrix data structures. It can be utilized to perform a number of mathematical operations on arrays such as trigonometric, statistical, and algebraic routines, Matplotlib for plotting and visualization, and Scikit-learn.

Scikit-learn library was used to deliver the clustering algorithm.it is a free software machine learning library for the Python programming language. It features various classification, regression and clustering algorithms including support vector machines, random forests, gradient boosting, k-means and DBSCAN, and is designed to interoperate with the Python numerical and scientific libraries NumPy and SciPy [21].

4.1 K-MEANS CLUSTERING

k-means clustering is a method of vector quantization, originally from signal processing, The k-means algorithm divides a set of N samples X into K disjoint clusters C, each described by the mean μj of the samples in the cluster. The means are commonly called the cluster “centroids”; note that they are not, in general, points from X, although they live in the same space [22].

K-means clustering has been used as a feature learning, in either (semi-) supervised learning or unsupervised learning. The basic approach is first to train a k-means clustering representation, using the input training data (which need not be labelled). Then, to project any input datum into the new feature space, an "encoding" function, such as the thresholded matrix-product of the datum with the centroid locations, computes the distance from the datum to each centroid, or simply an indicator function for the nearest centroid, or some smooth transformation of the distance [23].

4.2 PROCEDURE

The approach k-means follows to solve problem is called Expectation-Maximization. Below is a breakdown of all the steps on how to solve it mathematically:

$$J = \sum_{i=1}^{m} \sum_{k=1}^{K} w_{ik} \|x^i - \mu_k\|^2 \qquad (1)$$

where *wik*=1 for data point xi if it belongs to cluster k; otherwise, *wik* = 0. Also, μk is the centroid of xi's cluster.

we differentiate J w.r.t. *wik* first and update cluster assignments (*E-step*). Then we differentiate J w.r.t. μk and recompute the centroids after the cluster assignments from previous step (*M-step*). Therefore, E-step is:

$$\frac{\partial J}{\partial w_{ik}} = \sum_{i=1}^{m} \sum_{k=1}^{K} \|x^i - \mu_k\|^2$$

$$\Rightarrow w_{ik} = \begin{cases} 1 & \text{if } k = argmin_j \|x^i - \mu_j\|^2 \\ 0 & \text{otherwise.} \end{cases} \qquad (2)$$

In other words, assign the data point xi to the closest cluster judged by its sum of squared distance from cluster's centroid.

And M-step is:

$$\frac{\partial J}{\partial \mu_k} = 2 \sum_{i=1}^{m} w_{ik}(x^i - \mu_k) = 0$$

$$\Rightarrow \mu_k = \frac{\sum_{i=1}^{m} w_{ik} x^i}{\sum_{i=1}^{m} w_{ik}} \qquad (3)$$

Which translates to recomputing the centroid of each cluster to reflect the new assignments.

$$\frac{1}{m_k} \sum_{i=1}^{m_k} \|x^i - \mu_{c^k}\|^2 \qquad (4)$$

Source [24]

4.3 RESULTS AND ANALYSIS

Two images as shown in Fig4.1 below were processed and analyzed, Results and findings are there thereby shown below as well. These images composed of three bands merged together consisting of red(R), blue (B) and green (G). Both image contain 0% cloud cover.

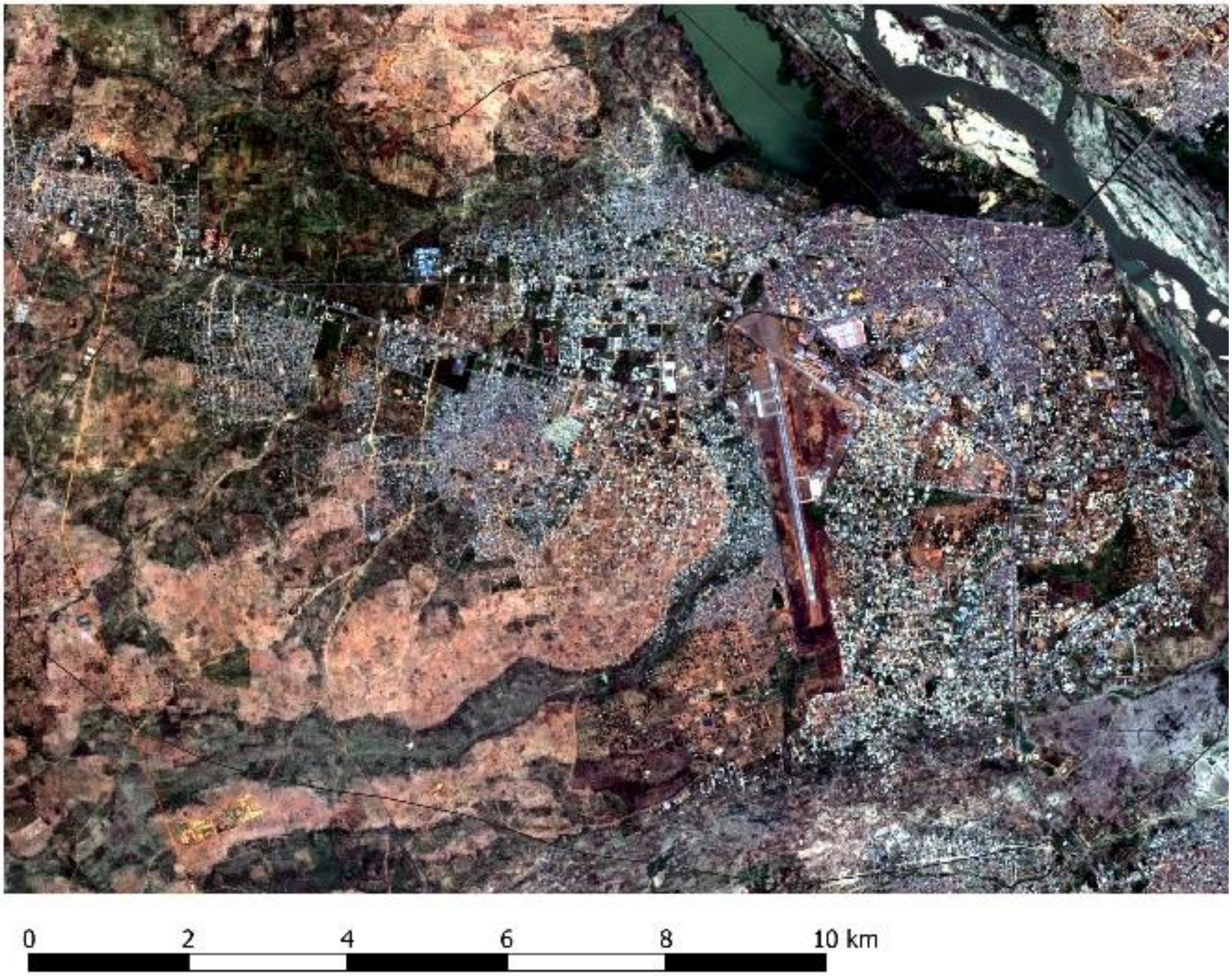

Fig4.1 Image of AOI (November 2019),

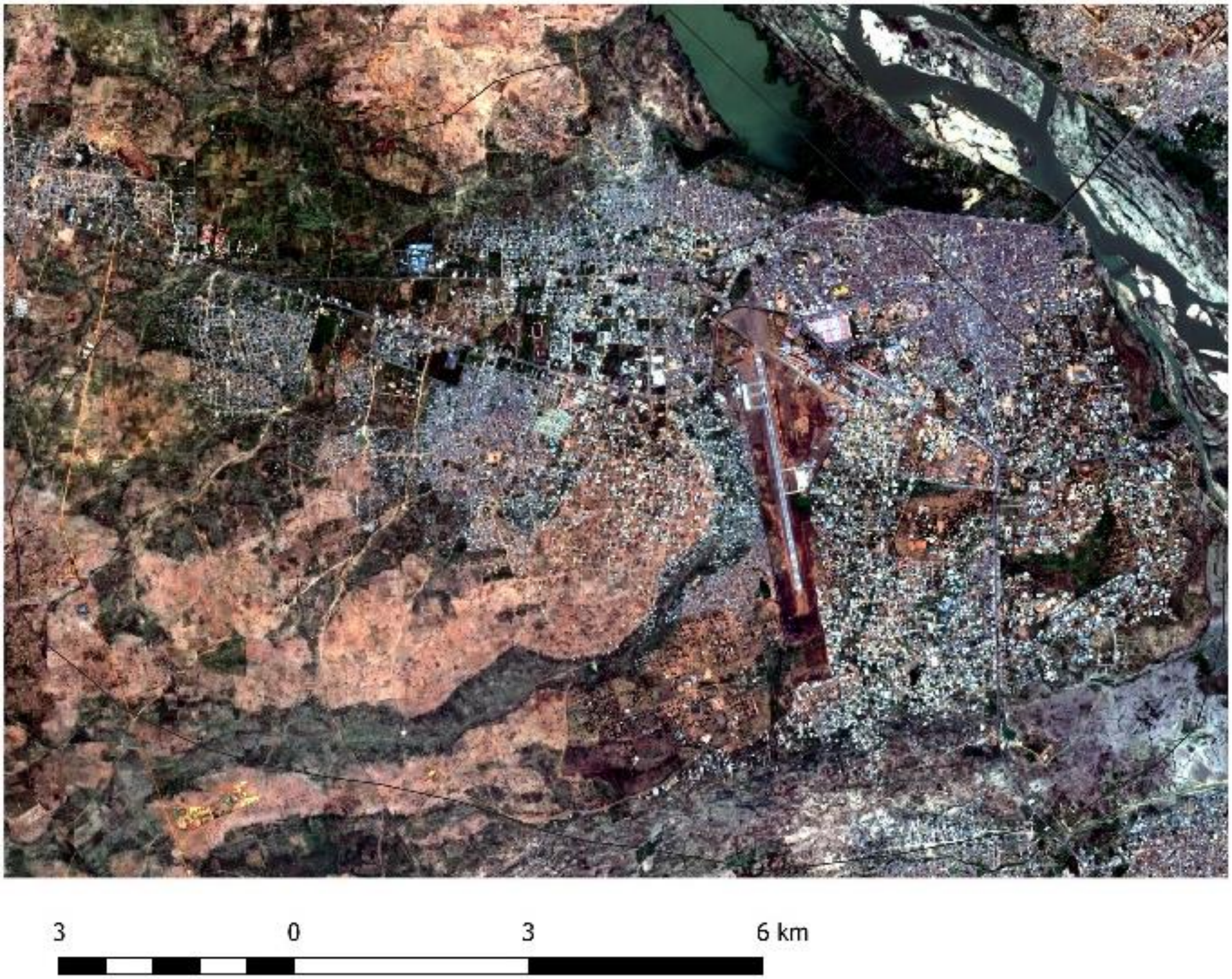

Fig4.1 Image of AOI (November 2017)

Result for Image 1

The dataset for the imagery was captured on 18 November 2019. Below is the histogram of the RGB bands contained in the image. It has a dimension of Size “1566 x 1118 x 3”. The plotted histogram is shown in fig 4.2 below:

Result for Image 2

The dataset for the imagery was captured on 8 November 2017. Below is the histogram of the RGB bands contained in the image (fig 4.3).

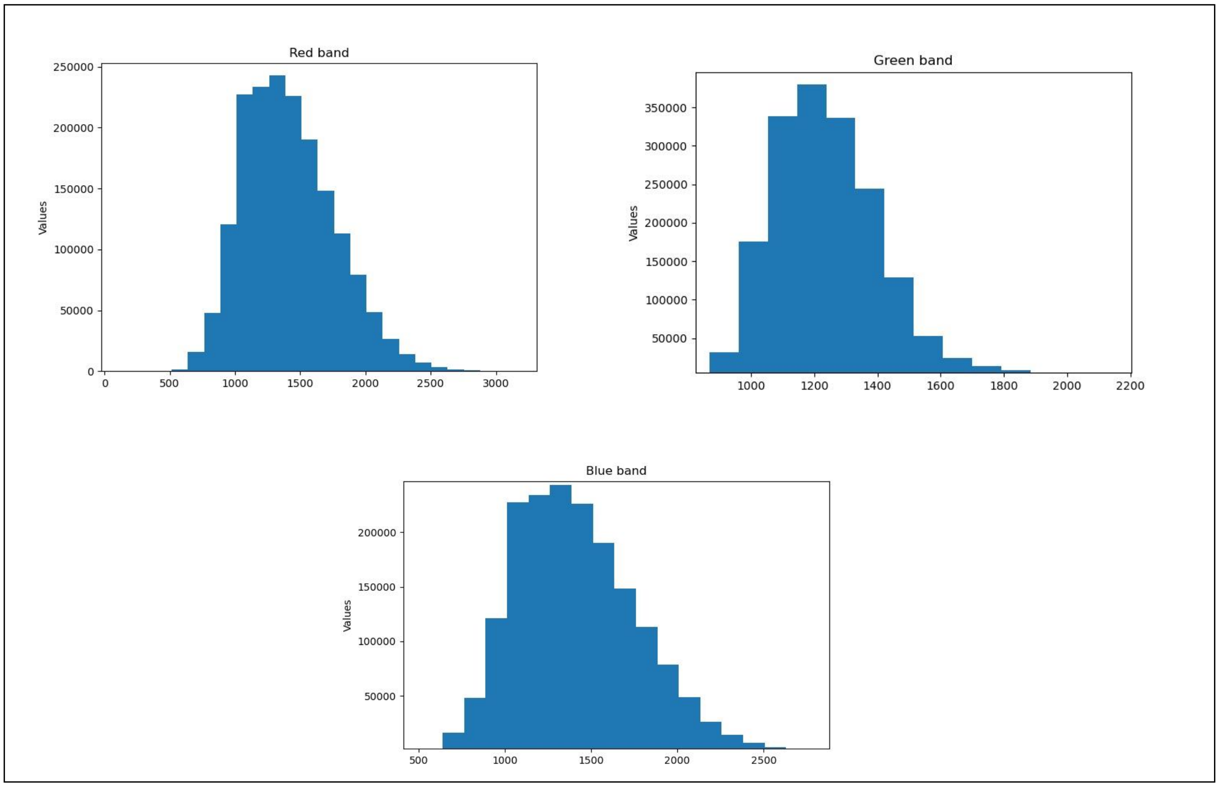

Fig 4.2 RGB composite of first satellite image

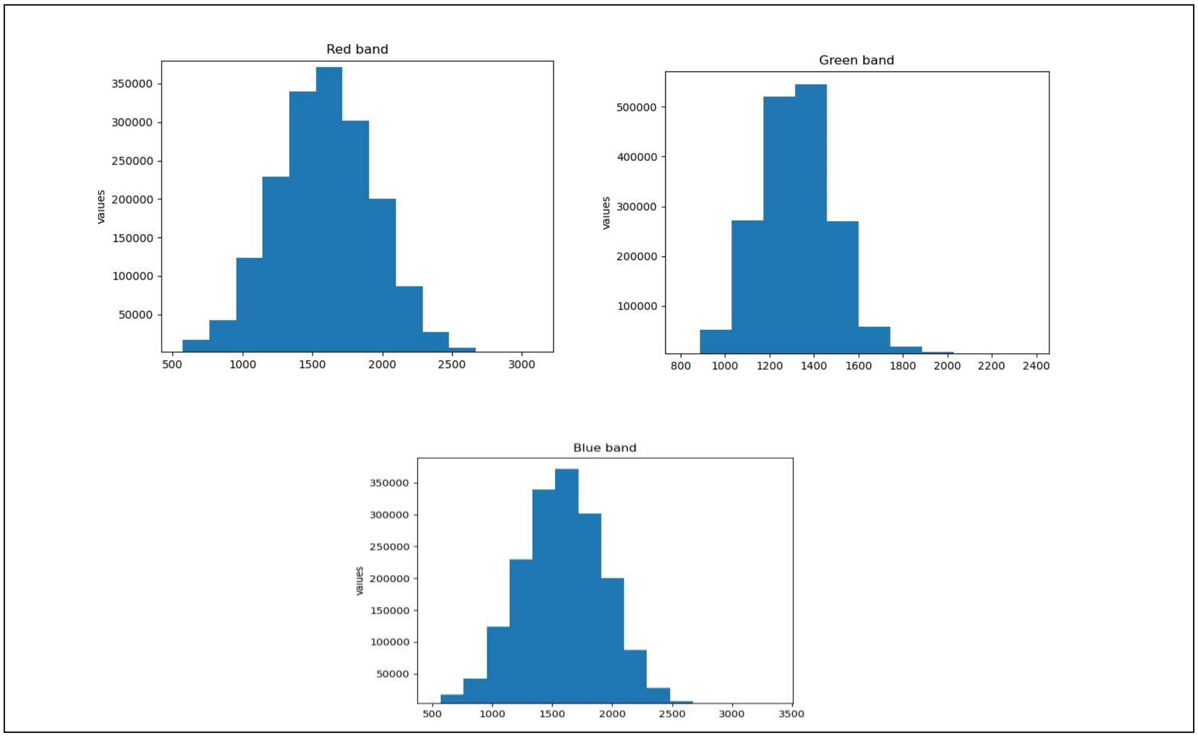

Fig 4.3 RGB composite of second satellite image

When implementing K-means the value of K has to be passed, the number of clusters k is an input parameter: an inappropriate choice of k may yield poor results. That is why, when performing k-means, it is important to run diagnostic checks for determining the number of clusters in the data set [23].

For K = 4 the results of the cluster is shown below:

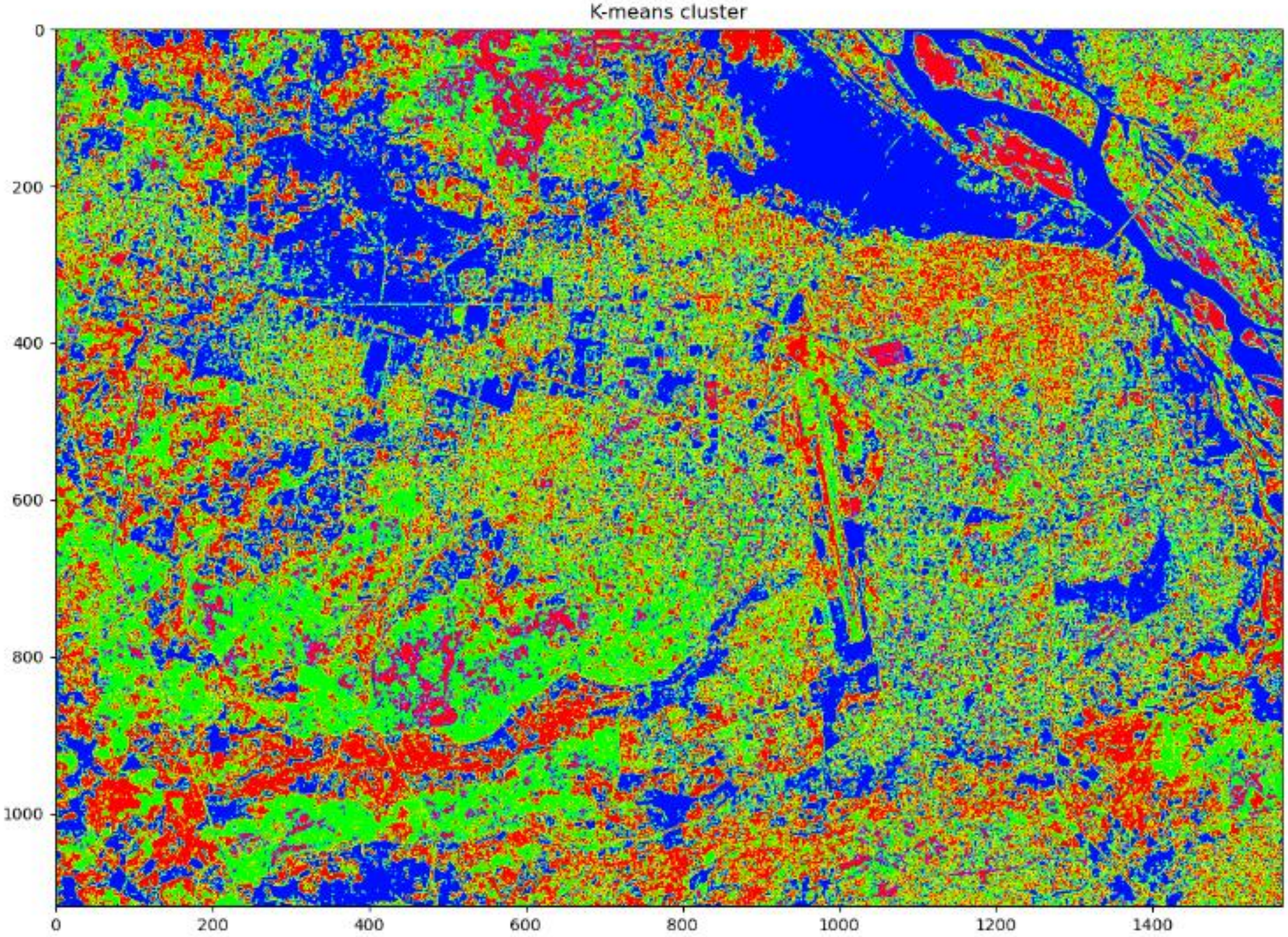

Fig 4.4 cluster image for satellite taken in 2019

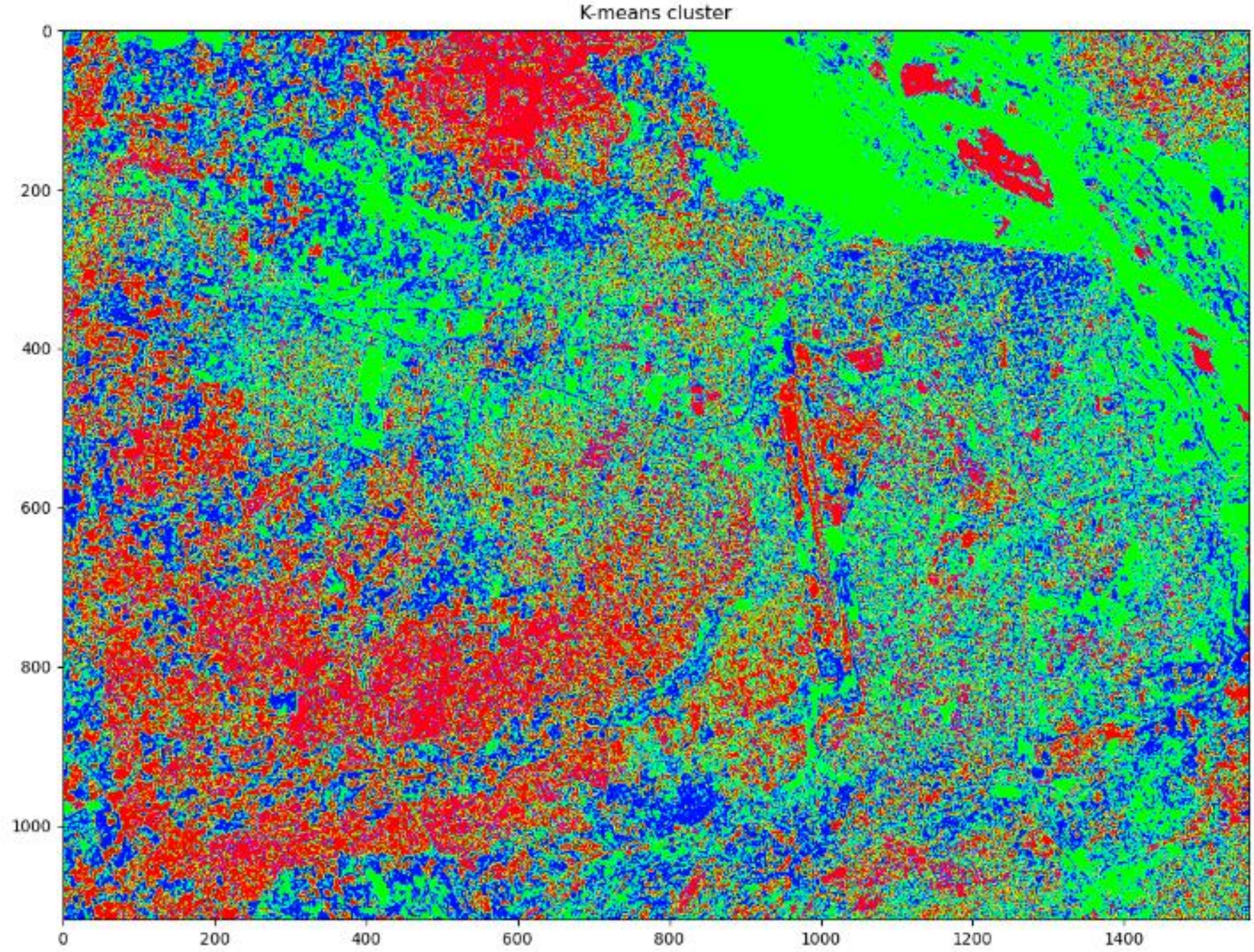

Fig 4.5 cluster image for satellite taken in 2017

From the above result, the label for the 4 cluster is water, built-up area, bare ground

The classification as in fig 4.4

Color	Classification
Blue	Water
Red	Built-up area
green	Bare ground

Pink	shrubs

The classification as in fig 4.5

Color	Classification
green	Water
Red	Built-up area
blue	Bare ground
pink	Shrubs

5. CONCLUSION

In this study, satellites images were downloaded and classified automatically with an unsupervised learning algorithm. The classification was efficient as the output were similar to real life data.

However, from observation and analysis of the downloaded datasets, changes were not much on the land use and the land cover as well.

The major challenge encountered during this study is inadequate accuracy in some of the details regarding the Area of Interest. This can cause conflicting results with regards to future research. Hence, there is need for further research in the field of GIS in this area to obtain a more better result and enhance decision making.

ACKNOWLEDGEMENT

I express my special gratitude to my research supervisors, Dr. Ervin Wirth, and Dr. Tamas Lovas for their special guidance and insightful comments, which is greatly invaluable to the completion of this project.

The faculty of civil engineering and Budapest university of Technology and Economics for this opportunity.

To my parents Mr. Emmanuel and Mrs. Fatima, am grateful for the spiritual support and everything to this time, you have done all what you are supposed to do.

My siblings, Kanwaye, Ahkalola, Catherine and Victor, your encouragement and motivation has always kept me moving; to all my family you are always in my heart thanks for always being there when I needed someone to talk to.
My friends and classmates all others not mention your presence has been a great stir up, motivation, and assisting in any way, thanks for the time spent together.

REFERENCES

[1] S. Garba and T. Brewer, “Assessment of Land Cover Change in the North Eastern Nigeria 1986 to 2005,” *J. Geogr. Geol.*, vol. 5, no. 4, pp. 94–105, 2013.

[2] X. Li, L. Zhang, and C. Liang, “A GIS-based buffer gradient analysis on spatiotemporal dynamics of urban expansion in Shanghai and its major satellite cities,” *Procedia Environ. Sci.*, vol. 2, no. 5, pp. 1139–1156, 2010.

[3] P. Sonde, S. Balamwar, and R. S. Ochawar, “Urban sprawl detection and analysis using unsupervised classification of high resolution image data of Jawaharlal Nehru Port Trust area in India,” *Remote Sens. Appl. Soc. Environ.*, vol. 17, no. October 2019, p. 100282, 2020.

[4] A. Movia, A. Beinat, and F. Crosilla, “Shadow detection and removal in RGB VHR images for land use unsupervised classification,” *ISPRS J. Photogramm. Remote Sens.*, vol. 119, pp. 485–495, 2016.

[5] M. Hussain, D. Chen, A. Cheng, H. Wei, and D. Stanley, “Change detection from remotely sensed images: From pixel-based to object-based approaches,” *ISPRS J. Photogramm. Remote Sens.*, vol. 80, pp. 91–106, 2013.

[6] T. Van De Voorde *et al.*, “Projecting alternative urban growth patterns: The development and application of a remote sensing assisted calibration framework for the Greater Dublin Area,” *Ecol. Indic.*, vol. 60, pp. 1056–1069, 2016.

[7] “Geospatial Analysis and Modelling of Urban Structure and Dynamics (GeoJournal Library) by Bin Jiang, Xiaobai Yao (z-lib.org),” *J. Vis. Lang. Comput.*, vol. 11, no. 3, p. 55, 2559.

[8] M. Benza, J. R. Weeks, D. A. Stow, D. López-Carr, and K. C. Clarke, “A pattern-based definition of urban context using remote sensing and GIS,”

Remote Sens. Environ., vol. 183, pp. 250–264, 2016.

[9] Diksha and A. Kumar, “Analysing urban sprawl and land consumption patterns in major capital cities in the Himalayan region using geoinformatics,” *Appl. Geogr.*, vol. 89, no. October, pp. 112–123, 2017.

[10] “Sentinel Collections.” [Online]. Available: https://developers.google.com/earth-engine/datasets/catalog/sentinel.

[11] J. Cihlar, R. Latifovic, and J. Beaubien, “A comparison of clustering strategies for unsupervised classification,” *Can. J. Remote Sens.*, vol. 26, no. 5, pp. 446–454, 2000.

[12] G. A. Naghdy, C. Todd, A. Olaode, and G. Naghdy, “Unsupervised Classification of Images: A Review,” *Int. J. Image Process.*, vol. 8, no. 5, pp. 325–342, 2014.

[13] “Unsupervised Classification,” 2019. [Online]. Available: http://gsp.humboldt.edu/OLM/Courses/GSP_216_Online/lesson6-1/unsupervised.html.

[14] “Yola, Adamawa.” [Online]. Available: https://en.wikipedia.org/wiki/Yola,_Adamawa#Climate.

[15] Britannica, “Jimeta.” [Online]. Available: https://www.britannica.com/place/Jimeta.

[16] A. B. Mohammed and A. A. Sahabo, “Water Supply and Distribution Problems in Developing Countries : A Case Study of Jimeta-Yola , Nigeria,” *Int. J. Sci. Eng. Appl. Sci.*, vol. 1, no. 4, pp. 473–483, 2015.

[17] “About sentinel-2.” [Online]. Available: https://eos.com/sentinel-2/.

[18] “USGS EROS Archive - Sentinel-2.” [Online]. Available: https://www.usgs.gov/centers/eros/science/usgs-eros-archive-sentinel-2?qt-science_center_objects=0#qt-science_center_objects.

[19] “GADM.” [Online]. Available: https://gadm.org/.

[20] “GDAL.” [Online]. Available: https://gdal.org/.

[21] the free encyclopedia Wikipedia, “Scikit-learn.” [Online]. Available: https://en.wikipedia.org/wiki/Scikit-learn.

[22] “k-MEANS.” [Online]. Available: https://scikit-learn.org/stable/modules/clustering.html#k-means.

[23] “k-means clustering.” [Online]. Available: https://en.wikipedia.org/wiki/K-means_clustering.

[24] I. Dabbura, “K-means Clustering: Algorithm, Applications, Evaluation Methods, and Drawbacks,” 2018. [Online]. Available: https://towardsdatascience.com/k-means-clustering-algorithm-applications-evaluation-methods-and-drawbacks-aa03e644b48a.