UNIVERSITY OF NAIROBI
SCHOOL OF COMPUTING AND INFORMATICS

Density-Based Cluster Analysis of Fire Hot Spots in Kenya's Wildlife Protected Areas

STEPHEN KAMAU KARANJA
P52/73076/2014

SUPERVISOR: DR. ROBERT O. OBOKO

APRIL 2016

A research project report submitted to the School of Computing and Informatics in partial fulfillment of the requirements for the award of the degree of M.Sc. Computational Intelligence of the University of Nairobi.

Declaration

I declare that this research project, as presented in this report, is my original work and has not been presented for a degree in any other institution of higher learning and that all sources I have used or quoted have been indicated and acknowledged by means of complete references.

Stephen K. Karanja Registration Number: P52/73076/2014

Signed: _____ Date: _____

This research project has been submitted for examination with my approval as University Supervisor.

Dr. Robert O. Oboko

Signed: _____ Date: _____

Acknowledgements

I would like to extend my gratitude to my supervisor, Dr. Robert O. Oboko, for his persistent support, guidance, and encouragement during the undertaking of this research project. I am thankful for his willingness to share his time, advice, and expertise in the fields of Data Mining and Machine Learning, which helped me in carrying out my project idea from conception to completion.

I also wish to acknowledge the enormous contribution and support provided by Mr. Wycliffe Mutero, Head of the GIS Section at KWS. In addition to providing domain expertise on fire activity in Kenya's WPAs, he was instrumental in providing access to KWS WPA management plans as well as WPA boundary shapefiles. Further, he played a crucial role in identifying user requirements and testing the web application developed as part of this study.

I acknowledge the use of FIRMS data and imagery from the LANCE system operated by NASA/GSFC/ESDIS with funding provided by NASA/HQ. The MODIS Active Fire Detections were extracted from the MCD14ML fire product distributed by NASA FIRMS.

Finally, I am highly indebted to the larger open source software community for freely providing several software tools without which I would not have successfully completed this project. The entire project was carried out in a Debian GNU/Linux environment. GAWK, QGIS, and MySQL were used to preprocess the MODIS active fire data set. Octave was used to plot most of the graphs. The ELKI software framework was used to perform the density-based cluster analysis. Inkscape was used to edit the ELKI scatterplot diagrams. Mozilla Firefox was the principal web browser used during the web application development. Dia was used to create some of the design diagrams. LaTeX was used to write this project report while LibreOffice Draw was used to create some of the figures in it.

My sincere thanks to all.

Stephen K. Karanja
December 2015

Abstract

Wildfires occurring in Kenya's wildlife protected areas pose a significant risk to wildlife conservation since they cause biodiversity loss and habitat degradation. There is a need for the Kenya Wildlife Service (KWS) to identify the regions in the protected areas that are prone to recurring wildfire outbreaks during the fire season.

This study identified regions that are fire hot spots in Kenya's protected areas by performing a density-based cluster analysis on the Moderate Resolution Imaging Spectroradiometer (MODIS) MCD14ML active fire data set for a 12 year period between 2003 and 2014. Feature subset selection was done using an AWK script written to extract the latitude and longitude fields from the data set. QGIS was used to filter fire points falling outside protected area boundaries. The Environment for Developing Knowledge Discovery in Databases Applications Supported by Index Structures (ELKI) implementation of the Density-Based Spatial Clustering of Applications with Noise (DBSCAN) algorithm was used for the clustering. A sorted k-dist graph estimated the initial DBSCAN parameters. 25 trial runs of DBSCAN with different parameters were used to select the final values: $MinPts = 7$ fire points; $Eps = 700$ meters. A web application with a Google Maps interface was developed to provide an interactive visualization of the fire hot spots.

4,968 fire incidents were observed in 73% of the protected areas. The initial DBSCAN parameters yielded 29 insignificant fire hot spot clusters from these incidents, while the final parameters yielded 43 significant clusters. The 43 clusters were identified in 31% of the protected areas that recorded fire activity. 60% of these clusters occurred in four protected areas.

The findings of this study indicate that density-based cluster analysis is a suitable clustering method for identifying hot spots in geospatial data sets. For DBSCAN, the performance of the sorted k-dist graph heuristic is influenced by the characteristics of a data set. The results also indicate that Chyulu Hills, Dodori, Boni, and Ruma are the protected areas most vulnerable to wildfires in Kenya.

This study recommends the use of density-based cluster analysis for identifying hot spots in geospatial data sets. Experimentation with a wide range of DBSCAN parameters values is advisable. KWS should focus fire management efforts on the identified fire hot spot regions. In addition, it should investigate the impact of wildfire damage in the ecological zones surrounding the hot spots.

Contents

List of Figures

List of Tables

List of Code and Data Files

List of Acronyms

Word	Phrase
ACP	African, Caribbean, and Pacific
AFIS	Advanced Fire Information System
AGNES	AGglomerative NESting
API	Application Programming Interface
ASTER	Advanced Spaceborne Thermal Emission and Reflection Radiometer
AVHRR	Advanced Very High Resolution Radiometer
BIRCH	Balanced Iterative Reducing and Clustering using Hierarchies
CLARANS	Clustering Large Applications based on Randomized Sampling
CLIQUE	CLustering In QUEst
CSV	Comma Separated Values
CURE	Clustering Using REpresentatives
DBSCAN	Density-Based Spatial Clustering of Applications with Noise
DENCLUE	DENsity-based CLUstEring
DIANA	DIvisive ANAlysis
ELKI	Environment for DeveLoping KDD Applications Supported by Index-Structures
EM	Expectation-Maximization
EOS	Earth Observing System
EPSG	European Petroleum Survey Group
ESDIS	Earth Science Data and Information System
ESRI	Environmental Systems Research Institute
ETM+	Enhanced Thematic Mapper Plus
EUMETSAT	European Organisation for the Exploitation of Meteorological Satellites
FIMMA	Fire Identification Mapping and Monitoring Algorithm
FIR	FIRe Monitoring Product
FIRMS	Fire Information for Resource Management System
FRAC	Full Resolution Area Coverage
GAC	Global Area Coverage
GDBSCAN	Generalized DBSCAN
GeoSTAT	Geographic Spatio-Temporal Analysis Tool

Word	Phrase
GIS	Geographic Information System
GNU	GNU's Not Unix!
GOES	Geostationary Operational Environmental Satellites
GRIDCLUS	GRID-based hierarchical CLUStering
GSFC	Goddard Space Flight Center
GUI	Graphical User Interface
HQ	Headquarters
HRPT	High Resolution Picture Transmission
HRV	High Resolution Visible
HTTP	Hypertext Transfer Protocol
ID	Identifier
IEC	International Electrotechnical Commission
IR	Infrared
ISO	International Organization for Standardization
JRC	Joint Research Centre
KBDCA	Kiunga-Boni-Dodori Conservation Area
KDD	Knowledge Discovery in Databases
KFS	Kenya Forest Service
KML	Keyhole Markup Language
KWS	Kenya Wildlife Service
KWSIDS	KWS Integrated Database System
LAC	Local Area Coverage
LANCE	Land, Atmosphere Near Real-Time Capability for EOS
LIS	Lightning Imaging Sensor
MCA	Meru Conservation Area
MODIS	MODerate Resolution Imaging Spectroradiometer
MSG	Meteosat Second Generation
NASA	National Aeronautics and Space Administration
NMF	Nonnegative Matrix Factorization
NOAA	National Oceanic and Atmospheric Administration
NP	National Park

Word	Phrase
NR	National Reserve
NS	National Sanctuary
OGC	Open Geospatial Consortium
OLI	Operational Land Imager
OMI	Ozone Monitoring Instrument
OPTICS	Ordering Points To Identify the Clustering Structure
PNG	Portable Network Graphics
POES	Polar Orbiting Environmental Satellites
RDBMS	Relational Database Management System
SCF	Science Computing Facility
SEVIRI	Spinning Enhanced Visible and InfraRed Imager
SPMF	Sequential Pattern Mining Framework
SQL	Structured Query Language
SSD	Satellite Products and Services Division
SSE	Sum Square Error
STING	STatistical INformation Grid
SVG	Scalable Vector Graphics
TCA	Tsavo Conservation Area
TIRS	Thermal Infrared Sensor
TOMS	Total Ozone Mapping Spectrometer
TRMM	Tropical Rainfall Measuring Mission
URL	Uniform Resource Locator
USGS	United States Geological Survey
VIRS	Visible and InfraRed Scanner
WF-ABBA	Wildfire Automated Biomass Burning Algorithm
WFS	Web Feature Service
WGS	World Geodetic System
WMS	Web Map Service
WPAs	Wildlife Protected Areas
XML	Extensible Markup Language

Chapter 1

Introduction

1.1 Background

The Kenya Wildlife Service (KWS) manages Kenya's Wildlife Protected Areas (WPAs). These include 23 National Parks (NPs), 31 National Reserves (NRs), 6 National Sanctuaries, 4 Marine National Parks and 6 Marine National Reserves. They cover approximately 8% of Kenya's total landmass (KWS, 2013c). Figure 1.1 is a map of the WPAs. There are two types of fires that occur in the WPAs. *Prescribed fires* are controlled by KWS WPA managers. They are applied for conservation purposes to reduce bush encroachment and improve the browsing and grazing conditions in the WPAs. On the other hand, *wildfires* pose a significant risk to wildlife conservation. They cause biodiversity loss and habitat degradation. This study addresses the wildfires due to their negative effects. They have a variety of causes including honey gathering, livestock grazing, and clearing of agricultural land adjacent to the WPAs. Both types of fire are important for wildlife conservation and are included in the KWS WPA management plans (KWS, 2012a,b,c,d, 2013a,b,d).

Satellite remote sensing technologies play a role in detecting, monitoring, and characterizing fires that occur on the Earth's surface. The Moderate Resolution Imaging Spectroradiometer (MODIS) instrument on board the National Aeronautics and Space Administration (NASA) Earth Observing System (EOS) Terra and Aqua satellites is one of the first to include fire monitoring in its design (Giglio, n.d.d). It provides at least four daily observations of active fires detected across the entire globe (NASA Earthdata, 2015). The MODIS active fire data is disseminated by the NASA Fire Information for Resource Management System (FIRMS) service (Davies et al., 2009).

Cluster analysis or **clustering** is the most common unsupervised machine learning task. It groups a set of objects in such a way that objects in the same group are more similar to each other than to those in other groups (Tan, Steinbach & Kumar, 2005). Data clustering has been applied in a wide variety of fields such as data mining, image analysis, bioinformatics, pattern recognition, and information retrieval. There are several methods of data clustering that differ in the way they model the clusters. *Density-based clustering* is a method that defines clusters as connected, dense areas in the data space (Ester, 2014). Objects in low-density regions do not belong to any cluster and are usually considered to be noise or outliers. Density-Based Spatial Clustering of Applications with Noise (DBSCAN) is a classical density-based clustering algorithm.

Research studies have applied clustering algorithms such as K-Means (Vadrevu et al., 2013) and DBSCAN (Usman, Sitanggang & Syaufina, 2015) to the problem of identifying and analyzing hot spot distribution in spatial data. Clustering has also been used in crime analysis to identify hot spots, where there are greater incidences of particular types of crime, in order to manage law enforcement resources more effectively (Divya, Rejimol & Selvan, 2014).

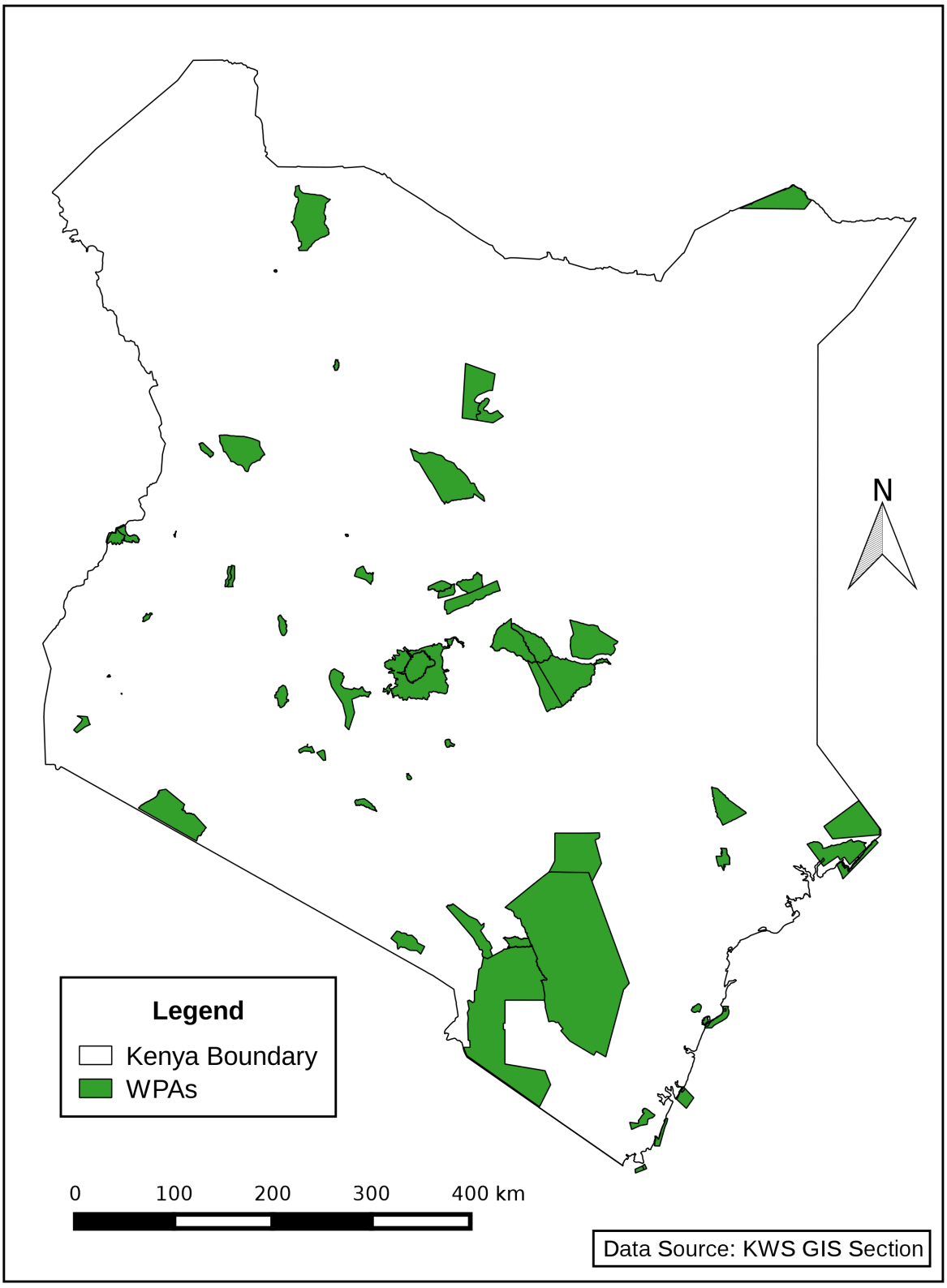


Figure 1.1: Kenya's Wildlife Protected Areas (WPAs)

1.2 Problem Statement

KWS lacks access to accurate information on the regions in the WPAs that are prone to recurring wildfire outbreaks, yet these are the regions where fire damage has the largest impact on wildlife conservation. The usefulness of this information is noted in the Mt. Kenya Ecosystem Management Plan which highlights the need to identify and map all fire hot spots as a strategy for preventing and managing wildfires in the ecosystem (KWS, 2013d). The limited understanding of the distribution pattern of fire incidents in the WPAs prevents the identification of high priority areas where fire management efforts need to be focused to reduce the negative impact of fire damage. This study seeks to support the decision-making process of KWS WPA managers by filling the existing information gap.

1.3 Research Objectives

This study defines the following overall research objective to address the above stated problem:

To provide information on the spatial distribution pattern of fire incidents in Kenya's WPAs

The specific research objectives under this overall objective are:

1. To identify regions that are fire hot spots in Kenya's WPAs by performing a density-based cluster analysis on the MODIS active fire data set, for a 12 year period (2003-2014)

2. To develop a web application that provides an interactive visualization of the fire hot spots in Kenya's WPAs

1.4 Significance

This study provides useful insights into patterns of wildfire occurrence in the WPAs that will help KWS WPA managers to allocate fire management resources effectively. The identified hot spot regions indicate the areas where fire monitoring efforts need to be focused during the fire season. Ground and aerial patrols of these regions can be intensified in an effort to reduce future incidents of wildfires. Further, this information lends support to the decisions of where to locate new fire watchtowers and firebreaks. In addition to the support for fire monitoring, the spatial patterns of fire hot spot clusters are also useful to KWS research scientists in addressing questions relating to causative factors and the ecological impact of the wildfires.

Chapter 2

Literature Review

2.1 Fire Activity in Kenya's Wildlife Protected Areas

Fires occur in most ecosystems of the world affecting an area of about 3 million km^2 every year (Palumbo, 2013). Fire occurrence is important for the biodiversity of many habitats with several positive effects such as the stimulation of new vegetation, the removal of dead vegetation, and the release of nutrients back into the soil. However, fire can also be detrimental and may compromise the survival of ecosystems. There are two types of fires that occur in Kenya's WPAs (KWS, 2012d, 2013d). *Prescribed fires* are controlled by KWS WPA managers. They are applied for conservation purposes to reduce bush encroachment and improve the browsing and grazing conditions for herbivores. On the other hand, *wildfires* pose a significant risk to wildlife conservation through biodiversity loss and habitat degradation.

Fire management is an important objective included in WPA management plans developed by KWS. The Mt. Kenya Ecosystem Management Plan (KWS, 2013d) reports that since 1990, wildfires have been recurring annually during the dry seasons of January-March and June-September. In this ecosystem, damage as a result of fire is highest in plantation forests due to the high tree uniformity and presence of flammable vegetation. Most wildfires are caused by arson and honey gathering. Other important causes are lightning, illegal grazing, clearing of farmland adjacent to the ecosystem, and charcoal burning.

An analysis of the fire outbreaks in the Aberdare ecosystem indicates that 96% of the fires occur in January-March while 4% occur in September (KWS, 2013a). The causes are diverse but the more common ones are honey gathering and clearing of neighbouring farms in preparation for cultivation. Vegetation damage as a result of fire is highest in the moorlands where the accumulated biomass provides the fuel to sustain burning. The Aberdare ecosystem is prone to frequent devastating wildfires. Such fires can lead to the loss of fire intolerant species while encouraging the invasion of fire resistant species. This brings about imbalances in the ecosystem. Measures such as the establishment of firebreaks and construction of fire watchtowers have been put in place to control wildfires.

In Tsavo Conservation Area (TCA), wildfires mainly originate from the surrounding community lands and along the Nairobi-Mombasa highway and railway line (KWS, 2012d). Fires are also an annual feature in the Chyulu Hills where there has been some concern that this may be having an impact on the forest fragments on the crests of the hills. The fire activity in the Chyulu Hills landscape is largely caused by the local communities. The two most important reasons for fire occurrence are honey gathering and livestock grazing whereby burning is conducted to stimulate nutritious pasture for livestock (Kamau, 2013).

The wildfires in TCA have a serious impact on the area's ecology and also reduce the effectiveness of prescribed fire as a management tool by undermining the ability of WPA managers to assess the overall effectiveness of the prescribed fires. The TCA Management Plan (KWS, 2012d) outlines fire management strategies such as raising the awareness of the risk and impact of wildfires among users of the Nairobi-Mombasa highway and also among communities living adjacent to TCA.

Kiunga Marine National Reserve (NR), Boni NR, and Dodori NR make up the Kiunga-Boni-Dodori Conservation Area (KBDCA) which is a KWS management planning unit. According to the KBDCA Management Plan (KWS, 2013b), the wildfires in this conservation area occur in the coastal forests. They are mainly caused by honey gatherers who use inappropriate honey harvesting methods and herders who use fire as a means of controlling pests and for pasture improvement. In addition, farmers also use fire to clear forest areas for cultivation.

Ruma National Park (NP) is a relatively small park covering an area of 120 km^2. The Ruma National Park Management Plan (KWS, 2012c) notes that the wooded grasslands covering 68% of the park increase its risk of wildfire outbreaks. The wildfires occuring in the park are both intentional and accidental. Poachers start wildfires to stimulate green flush, especially after the rains, to attract grazers. Accidental fires occur before the planting season when farmers burn farm litter during preparation of land for planting (KWS, 2012c).

Table 2.1 summarizes the common causes of wildfires in Kenya's WPAs as identified in the KWS WPA management plans (KWS, 2012a,b,c,d, 2013a,b,d; KWS & KFS, 2012). Table 2.2 summarizes the wildfire management measures applied in the WPAs. These measures were also identified from the KWS WPA management plans.

2.2 Satellite Remote Sensing Technologies for Monitoring Fires

Satellite remote sensing technologies play a role in detecting, monitoring, and characterizing fires that occur on the Earth's surface. The MODIS Active Fire and Burned Area Products website (Giglio, n.d.d) notes that there are several satellite systems currently in orbit that provide information on different fire characteristics such as location and timing of active fires, burned area, areas that are dry and susceptible to wildfire outbreaks, and pyrogenic trace gas and aerosol emissions. Further, the satellite systems have different capabilities in terms of spatial resolution, sensitivity, spectral bands, and times and frequencies of overpasses. However, none of the sensing systems prior to MODIS included fire monitoring in their design. The following satellite systems have applications for fire monitoring.

Table 2.1: Common causes of wildfires in Kenya's WPAs

Type	**Cause**	**Affected WPA**
Illegal activity	Inappropriate honey gathering methods	Aberdare NP, Chyulu Hills NP, KBDCA, Mt. Kenya Ecosystem
	Herders start fires to improve pasture for livestock grazing	Chyulu Hills NP, Kakamega NR, KBDCA, MCA, Mt. Kenya Ecosystem
	Poachers start fires to improve pasture which attracts grazers	Ruma NP, Kakamega NR
	Charcoal burning	Aberdare NP, Mt. Kenya Ecosystem
	Farmers start fires to clear forest areas for cultivation	KBDCA
	Highway and/or railway line users start fires along the transport system	TCA
	Arson	Aberdare NP, Mt. Kenya Ecosystem
Accidental	Farmers start fires to clear farmland adjacent to the WPAs in preparation for cultivation	Aberdare NP, Kakamega NR, MCA, Mt. Kenya Ecosystem, Ruma NP
	Fires originate from adjacent community / pastoral lands	Hell's Gate NP, Mount Longonot NP, TCA
	Tourists start fires through inappropriate disposal of cigarette butts	Aberdare NP, Mt. Kenya Ecosystem
Natural	Lightning	Mt. Kenya Ecosystem

2.2.1 Advanced Very High Resolution Radiometer

The Advanced Very High Resolution Radiometer (AVHRR) is flown on National Oceanic and Atmospheric Administration (NOAA) Polar Orbiting Environmental Satellites (POES). It measures electromagnetic radiation (light reflected and heat emitted) from Earth. The AVHRR was originally intended only as a meteorological satellite system but it does have applications for fire monitoring. Its visible bands can detect smoke plumes from fires as well as burn scars. The middle-infrared band can detect actual hot spots and active fires. Its ability to detect fires is greater at night, since the system can confuse active fires with heated ground surfaces, such as beach sand and asphalt (Giglio, n.d.a).

Table 2.2: Wildfire management measures applied in Kenya's WPAs

Measure	**WPA**
Establishment and maintenance of firebreaks e.g. roads, fire resistant plants	Aberdare NP, Arabuko Sokoke NP, Hell's Gate NP, Kakamega NR, KB-DCA, MCA, Mt. Kenya Ecosystem, Mt. Longonot NP, Ruma NP
Construction of fire watchtowers	Aberdare NP, Arabuko Sokoke NP, Kakamega NR, TCA
Establishment of a fire station / fire and rescue center / rapid response unit	Aberdare NP, TCA
Collection and analysis of information on wildfire occurrences	Aberdare NP, Mt. Kenya Ecosystem
Procurement of fire-fighting equipment	Aberdare NP, Kakamega NR, KB-DCA, Mt. Kenya Ecosystem, Ruma NP
Training of staff / community forest associations on fire fighting	Kakamega NR, KBDCA, Mt. Kenya Ecosystem, Ruma NP
Raising awareness of the risk and impact of wildfires among communities living adjacent to the WPAs	Aberdare NP, Arabuko Sokoke NP, Kakamega NR, Ruma NP, TCA
Raising awareness of the risk and impact of wildfires among users of the Nairobi-Mombasa highway	TCA
Raising awareness of the risk and impact of wildfires among tourists e.g. campers	Mt. Kenya Ecosystem
Development and maintenance of communication systems to support fire management activities e.g. radio, telephone	KBDCA, Mt. Kenya Ecosystem
Development / review of a fire management plan	Kakamega NR, Mt. Kenya Ecosystem, Ruma NP
Establishment of fire fighting operational guidelines	Mt. Kenya Ecosystem, TCA
Establishment of an elaborate fire detection and reporting system	Mt. Kenya Ecosystem

For Local Area Coverage (LAC), High Resolution Picture Transmission (HRPT), and Full Resolution Area Coverage (FRAC) data, the instantaneous field-of-view of each channel is approximately 1.4 milliradians leading to a resolution of 1.1 km at the satellite subpoint for a nominal altitude of 833 km. Global Area Coverage (GAC) data has a 4 km resolution. POES satellite orbits are timed to allow complete global coverage twice per day, per satellite, in swaths of about 2,600 km in width (NOAA, 2012).

The Fire Identification Mapping and Monitoring Algorithm (FIMMA) is an automated algorithm used to detect fires from AVHRR data. The algorithm is only accurate over forested regions. Fire pixels occurring within landcover types with some tree cover are kept as possible fires. As a result, the algorithm may miss real fires over urban areas, as well as agricultural burns (NOAA SSD, 2015).

2.2.2 Geostationary Operational Environmental Satellites

The Geostationary Operational Environmental Satellites (GOES) are operated by NOAA. They house a 5-channel (1 visible, 4 infrared) imaging radiometer designed to sense radiant and solar reflected energy from sample areas of the Earth. They are stationed in orbits that remain fixed over one spot on the equator, providing continuous coverage of the Western Hemisphere. GOES satellites acquire images every 15 minutes, at up to 1 km resolution in visible light, for the detection of smoke, and 4 km resolution in thermal infrared to directly detect the heat of fires (Giglio, n.d.b).

The Wildfire Automated Biomass Burning Algorithm (WF-ABBA) is a contextual multi-spectral thresholding algorithm which utilizes dynamic local thresholds derived from the GOES satellite imagery and ancillary databases to locate fire pixels (NOAA SSD, 2014). It also provides very rough estimates of the sub-pixel area and mean temperature of fires. Fire locations represent the approximate location of the fire pixel and do not represent the actual fire size. The minimum detectable fire size at the sub-satellite point, and smoldering at 450 K, is approximately 0.5 to 1 acre in size in relatively non-cloudy conditions. WF-ABBA is also able to identify hot spots through smoke.

2.2.3 Landsat

The Landsat series of Earth-observing satellites monitor characteristics and changes on the surface of the Earth at high resolution (Giglio, n.d.c). The Landsat 7 satellite uses the Enhanced Thematic Mapper Plus (ETM+) to acquire images of the Earth which provide land surface information (USGS, 2015b). Landsat 8 carries two instruments: the Operational Land Imager (OLI) and the Thermal Infrared Sensor (TIRS) (USGS, 2015a). The 11 spectral bands on Landsat 8 (and 8 spectral bands on Landsat 7) monitor different types of Earth resources over a wide area (81° North to 81° South). The thermal band enables the system to detect "hot spots". Landsat 8's TIRS provides two thermal bands. In both Landsat 7 and Landsat 8, all spectral bands except band 8 provide a spatial resolution of 30 m. The panchromatic band 8 has a resolution of 15 m. Landsat 7 and 8 provide impressive high-resolution images but only infrequently, revisiting an area every 16 days.

2.2.4 Total Ozone Mapping Spectrometer

The Total Ozone Mapping Spectrometer (TOMS) is a measuring device that provides data regarding ozone levels. It produces a complete data set of daily ozone levels around the world. This instrument is the first to show aerosols (airborne dust and smoke particles) over land. It also provides the ability to distinguish aerosols that absorb light from aerosols that reflect it. TOMS makes 35 measurements every 8 seconds, each covering an area 50-200 km wide on the ground. Close to 200,000 daily measurements cover almost every spot on the Earth except for areas near the poles. These data make it possible to observe a variety of Earth events including forest fires, dust storms and biomass burning (Giglio, n.d.e).

The Ozone Monitoring Instrument (OMI) on board the NASA Aura satellite records total ozone and other atmospheric parameters related to ozone chemistry and climate (Wilson, 2007). It employs hyperspectral imaging in a push-broom mode to observe solar backscatter radiation in the visible and ultraviolet ranges of the electromagnetic spectrum. The instrument views the Earth in 740 wavelength bands along the satellite track with a swath large enough to provide global coverage in 14 orbits (1 day). The nominal 13 $\times$ 24 km spatial resolution can be zoomed to 13 $\times$ 13 km for detecting and tracking urban-scale pollution sources.

2.2.5 Meteosat Second Generation

The Meteosat Second Generation (MSG) geostationary weather satellites house the optical imaging radiometer called the Spinning Enhanced Visible and InfraRed Imager (SEVIRI). The sensor features 12 spectral channels and provides cloud imaging and tracking, fog detection, measurement of the Earth surface and cloud top temperatures, tracking ozone patterns, as well as active fire monitoring (NASA Earthdata, 2015). The nominal coverage of the satellites includes Europe, Africa, and adjacent seas. They provide full disc imagery data every 15 minutes. The various channels provide measurements with a resolution of 3 km at the sub-satellite point. The High Resolution Visible (HRV) channel provides measurements with a resolution of 1 km (ESA, 2015).

The European Organisation for the Exploitation of Meteorological Satellites (EUMETSAT) provides the active fire monitoring product (FIR). FIR is a fire detection product that indicates the presence of fire within a pixel. The underlying concept of the algorithm takes advantage of the fact that SEVIRI channel IR 3.9 is very sensitive to fire hot spots. The algorithm distinguishes between potential fires and active fires (EUMETSAT, 2015).

2.2.6 Moderate Resolution Imaging Spectroradiometer

The MODIS instrument on board the NASA EOS Terra and Aqua satellites detects fires that are burning at the time of overpass, under relatively cloud-free conditions. It acquires data continuously providing global coverage every 1-2 days. There are at least 4 daily MODIS observations for almost every area on the equator, with the number of

overpasses increasing (due to overlapping satellite orbits) closer to the poles. A MODIS active fire detection represents the center of a 1 km^2 pixel flagged as containing one or more actively burning fires (NASA Earthdata, 2015).

Fire detection is performed using a contextual algorithm that exploits the strong emission of mid-infrared radiation from fires (Giglio et al., 2003). Thresholds are first applied to the observed mid-infrared and then the thermal infrared brightness temperature after which false detections are rejected by examining the brightness temperature relative to neighboring pixels. Validation of the Terra MODIS active fire product has primarily been performed using coincident, high resolution fire masks derived from Advanced Spaceborne Thermal Emission and Reflection Radiometer (ASTER) imagery (Giglio, 2013).

The NASA FIRMS service distributes: (i) Near Real-Time MODIS active fire data (MCD14DL) that is processed by Land, Atmosphere Near Real-Time Capability for EOS (LANCE) using the standard MODIS MOD14/MYD14 Fire and Thermal Anomalies product, and (ii) Standard MODIS active fire data (MCD14ML) that is processed by the MODIS Fire Science Computing Facility (SCF) at the University of Maryland (NASA FIRMS, n.d.). The Near Real-Time MODIS active fire data is available within 3 hours of satellite overpass while the Standard MODIS active fire data is generally available after 3 months. The Standard data is quality checked and sometimes reprocessed at a later date if some problems are found with specific granules (NASA Earthdata, 2015).

The MODIS active fire data includes 12 attribute fields labeled as follows: Latitude, Longitude, Brightness, Scan, Track, Acquisition Date, Acquisition Time, Satellite, Confidence, Version, Bright_T31, and FRP. Each of these attributes is fully described in NASA FIRMS (n.d.). FIRMS has developed a global fire e-mail alert system based on the Near Real-Time data (Davies et al., 2009). The system notifies subscribed users when a fire is detected in, or near, a specified area of interest, country or protected area. In addition to this, active fire locations for the last 24 hours, 48 hours, and 7 days are available for download in shapefile, Keyhole Markup Language (KML), Web Map Service (WMS), or Comma Separated Values (CSV) formats. Older data can be obtained through the Archive Download Tool (NASA FIRMS, 2015). The tool provides Near Real-Time data and, as it becomes available (usually after 2 months), it is replaced with data extracted from the Standard MODIS fire product. The tool allows download of archived data by specifiying a region of interest and time period (NASA Earthdata, 2015).

2.2.6.1 Precision of the Active Fire Data

The precision of the Latitude and Longitude attribute fields in the MODIS active fire data is 3 decimal places. Assuming a mean Earth radius of 6,371 km, 1° of longitude and latitude at the equator represents a distance of about 111 km as per equation (2.1). This is because the Earth's circumference is about 40,030 km and there are 360° of longitude.

$$\text{Distance of } 1^\circ \text{ of longitude (km)} = \frac{2 \times \pi \times 6,371}{360} \tag{2.1}$$

The 3 decimal places used in the MODIS data set coordinates means that the smallest possible value for latitude or longitude is 0.001°. This represents a precision of about 111 m at the equator. The maximum rounding error is half this value (55.5 m). Equation (2.2) illustrates this.

$$\text{Maximum rounding error (m)} = \frac{\text{Distance of } 1^\circ \text{ of longitude (km)} \times 0.001 \times 1000}{2} \tag{2.2}$$

The spatial resolution of the MODIS data set is approximately 1 km^2 per pixel. Each active fire detection represents the center of a pixel flagged as containing one or more fires, or other thermal anomalies. The location is the center point of the pixel and not necessarily the coordinates of the actual fire on the ground (NASA Earthdata, 2015). Since a fire could have happened at any point inside the 1 km^2 pixel, the farthest locations from the center are at the corners of the pixel. This distance is 707.1 m. Equation (2.3) shows how it is determined using the Pythagorean Theorem for a right triangle whose short sides are 500 m long.

$$\text{Farthest distance from pixel center (m)} = \sqrt{500^2 + 500^2} \tag{2.3}$$

Taking into account the rounding error and the spatial resolution, the maximum possible error for the location of a MODIS fire point is therefore 762.6 m (707.1 m + 55.5 m). However, the flat-Earth distance of 707.1 m will differ slightly from the equivalent great-circle distance of a spherical Earth, computed with the haversine formula.

2.2.7 Comparative Summary

Table 2.3 provides a comparative summary of the satellite systems that have applications for fire monitoring. It shows that fire data from the GOES satellites cannot be used for this study because their coverage is restricted to the Western Hemisphere. TOMS provides data on smoke particles which is an indicator of fire activity. However, data on active fires is more suitable for this study since it is the most direct indicator of fire activity.

Although Landsat imagery data is exceptionally high-resolution, it is available very infrequently. The 16-day repeat-cycle will result in several active fires going unobserved. MSG-SEVIRI data has a much higher temporal resolution than MODIS data. However, it has a poorer spatial resolution. A higher spatial resolution is considered a better trade-off for this study. In addition, the MODIS data provides fire locations extracted using an algorithm whose results have been validated (Giglio, 2013).

The table shows that AVHRR data has both a lower spatial and temporal resolution than the MODIS data. Based on this comparison of the various remote sensing fire products available, MODIS appears to be the most appropriate for this study.

Table 2.3: Comparison of fire monitoring satellite systems

System	Coverage	Spatial Resolution	Temporal Resolution	Data
AVHRR	Global	1.1 km and 4 km	Twice per day	Fire locations
GOES	Western Hemisphere	4 km	Every 15 minutes	Fire locations
Landsat	81°N-81°S	15 m and 30 m	Every 16 days	Imagery on fires
TOMS	Global	13 × 24 km	Once per day	Imagery on smoke particles
SEVIRI	Europe, Africa	3 km	Every 15 minutes	Imagery on fires
MODIS	Global	1 km	4 times per day	Fire locations

2.3 Data Clustering Methods

According to Russell and Norvig (2010), the three main categories of machine learning are determined by three types of feedback available to learn from. In *supervised learning* an agent observes some example input-output pairs and learns a function that maps from input to output. In *unsupervised learning* the agent learns patterns in the input without any explicit feedback. In *reinforcement learning* the agent learns from a series of rewards or punishments. The most common unsupervised learning task is data clustering.

Aggarwal (2014) has broadly defined the basic problem of clustering to be one in which, given a set of data points, the objective is to partition them into a set of groups that are as similar as possible. More specifically, data points in the same group are more similar to each other than to those in other groups (Tan, Steinbach & Kumar, 2005). Data clustering has been applied in a wide variety of data mining and machine learning tasks such as image segmentation, pattern recognition, information retrieval, and also in the field of bioinformatics. It is important due to its capabilities in determining the intrinsic grouping in a set of unlabeled data. The increasing availability of large data sets has rendered manual labeling difficult and expensive (Alelyani, Tang & Liu, 2014). This has created a growing interest in the use of data clustering as a technique for automatic data labeling in the data mining process.

Clustering can be seen as either an exploratory task or preprocessing step (Alelyani, Tang & Liu, 2014). If the goal is to explore and reveal the hidden patterns in the data, clustering becomes a stand-alone exploratory task by itself. On the other hand, if the generated clusters are going to be used to facilitate another data mining or machine learning task, clustering will be a preprocessing step. During preprocessing, clustering can be used to handle noisy data by detecting and removing outliers. It can also be used

at this stage to label the data for subsequent classification tasks.

There are several clustering methods described in the literature. Key methods are: hierarchical clustering, partitional clustering, probabilistic clustering, density-based clustering, grid-based clustering, Nonnegative Matrix Factorization (NMF), and spectral clustering. The rest of this section presents a brief overview of the characteristics of each of these methods while highlighting their strengths, weaknesses, and the type of clustering problems for which they are suitable.

2.3.1 Hierarchical Clustering

Hierarchical clustering creates a hierarchy of clusters that are represented through a binary tree-based data structure called a *dendrogram*. *Agglomerative* algorithms implement a bottom-up approach that begins with each data point as a singleton cluster and successively merges them into larger clusters. A suitable distance metric is used to measure similarity between data points and a linkage criterion based on this distance determines the choice of clusters to merge at each step. Some popular choices are *single-linkage* clustering, *complete linkage* clustering, and *average linkage* clustering. On the other hand, *divisive* algorithms implement a top-down approach that begins with all the data points in a huge macro-cluster and successively splits it into smaller clusters. Partitional clustering algorithms can be used to perform this splitting (Reddy & Vinzamuri, 2014). Examples of algorithms for this method of clustering are: Balanced Iterative Reducing and Clustering using Hierarchies (BIRCH), CHAMELEON, AGglomerative NESting (AGNES), and DIvisive ANAlysis (DIANA).

Hierarchical clustering is not very robust to outliers since it does not have a notion of noise. Such data points appear as additional clusters and may cause other clusters to merge, a phenomenon known as *chaining*. In addition, it is ineffective at capturing arbitrarily shaped clusters. This method would therefore not be ideal for performing density-based clustering in which outliers define low-density regions in the data space. Further, the run-time complexity of hierarchical algorithms is quadratic hence making them undesirable especially for large-scale problems (Reddy & Vinzamuri, 2014).

2.3.2 Partitional Clustering

Partitional clustering divides the data set into several clusters at once rather than hierarchically, using partitioning representatives. In each iteration, every data point is assigned to the cluster whose partitioning representative is nearest, as determined by a distance function, and then each representative is adjusted according to the data points assigned to it. K-Means is a widely used algorithm where the partitioning representatives are cluster centroids corresponding to the mean of all the data points in each cluster. It typically uses the Euclidean distance metric. Variants such as K-Medians which uses the median rather than the mean and K-Medoids which samples the partitioning representative from the data, also exist. In addition, the Fuzzy C-Means algorithm considers each data point as having a degree of membership to each cluster (Aggarwal, 2014; Reddy & Vinzamuri, 2014).

As noted by Reddy and Vinzamuri (2014), K-Means has the advantage of simplicity and computational efficiency. This allows it to be used on large data sets. However, it does not yield the same clustering on different runs due to the initial random assignment of centroids. The requirement to specify the number of clusters (K) in advance also proves to be a significant drawback in certain practical use cases where this can only be determined experimentally. It is difficult for K-Means to detect non-spherical clusters or clusters of different sizes and densities because it partitions the data space into Voronoi cells which produces clusters of equal size and convex shape.

2.3.3 Probabilistic Clustering

Probabilistic clustering defines clusters as data points most likely belonging to the same probability distribution. A generative model, such as a mixture of Gaussians, is often *assumed* on the data set (Deng & Han, 2014). Here, the Expectation-Maximization (EM) algorithm models the data set with a fixed number of Gaussian distributions that are initialized randomly. It then iteratively optimizes the parameters of the model to achieve a maximum likelihood fit to the data set (Aggarwal, 2014). This clustering method is highly effective for clustering artificial data sets generated by sampling random objects from a distribution. While it provides a strong statistical foundation for modeling clusters, many real data sets might not be generated from the mathematical model that it assumes. For instance, in data sets where clusters are density-based, assuming Gaussian distributions on the data will yield ineffective results.

2.3.4 Density-based Clustering

Density-based clustering defines clusters as connected, dense areas in the data space separated from each other by sparser areas (Ester, 2014). Data points in the sparse regions are considered to be noise or border points to more than one cluster and are not assigned to any cluster. Mean-shift clustering is a related method which produces density-based clusters but does not guarantee that clusters are connected.

DBSCAN and Ordering Points To Identify the Clustering Structure (OPTICS) are two typical density-based clustering algorithms that can discover arbitrarily shaped clusters. They both expect a density drop in the data space to effectively detect cluster borders and will yield arbitrary results where the cluster density decreases continuously. Density-based clustering methods also have the drawback of not being able to detect intrinsic cluster structures such as the mixture of Gaussian distributions commonly found in artificially generated data sets. Another challenge of these methods is that they are naturally defined on data points in a continuous space and therefore cannot be meaningfully used in a discrete or non-Euclidean space without specialized transformations (Aggarwal, 2014). In addition, processing high-dimensional data may pose a challenge since density is more difficult to define for such data.

According to Ester (2014), clustering algorithms such as EM and K-Means produce spherical clusters due to the assumption that data are generated from a probability distribution of a given type. However, spatial data with a reference to a two or three-dimensional

concrete space corresponding to our real world naturally contains non-spherical clusters. These clusters may have arbitrary shapes due to constraints imposed by geographic features such as mountains and rivers. Ester further notes that the paradigm of density-based clustering has been proposed to not only meet the requirement to discover clusters of arbitrary shape but also to scale to large databases and detect and remove noise and outliers in the data. It can be considered a non-parametric method, since it makes no assumptions about the number of clusters or their distribution.

2.3.5 Other Clustering Methods

Grid-based clustering algorithms partition the data space into a finite number of cells to form a grid structure and then form clusters from these cells. The clusters correspond to regions that are more dense in data points than their surroundings. Since, grid-based clustering algorithms cluster cells rather than individual data points, they present a significant reduction in time complexity and are efficient in mining large data sets (Cheng, Wang & Batista, 2014).

NMF factorizes an input nonnegative matrix into two nonnegative matrices of lower rank. It can be applied in solving data mining and machine learning problems such as pattern recognition and text mining. NMF with the sum of squared error cost function is equivalent to a relaxed K-Means clustering (Li & Ding, 2014).

Spectral clustering algorithms construct a similarity graph for all the data points after which the data points are embedded in a space with the use of the eigenvectors of the graph Laplacian. Finally, a classical clustering algorithm such as K-Means is applied to partition the embedding. The name *spectral* denotes the fact that the clustering results are obtained by analyzing the spectrum of the graph Laplacian. Spectral clustering has been applied to the problems of image segmentation, text mining, and speech processing. Studies have shown that it is theoretically closely related to kernel K-Means and NMF (Liu & Han, 2014).

2.3.6 Common Clustering Algorithms

Table 2.4 presents a summary of common clustering algorithms grouped under the clustering method they belong to.

2.4 Density-based Clustering Algorithms

2.4.1 Density-Based Spatial Clustering of Applications with Noise

DBSCAN is a classical density-based data clustering algorithm that was proposed by Ester et al. (1996). It requires two parameters: *Eps* (ε) and *MinPts*. A point is called a *core point* if its neighborhood of radius *Eps* contains at least *MinPts* points. A point q is *directly density-reachable* from a core point p if q is within the *Eps*-neighborhood of p. Density-reachability is given by the transitive closure of direct density-reachability. Two

Table 2.4: Common clustering algorithms

Method	**Algorithm**	**Comment**
Hierarchical	AGNES	Agglomerative
	DIANA	Divisive
	CURE	
	CHAMELEON	
	BIRCH	
Partitional	K-Means	Most widely used partitional clustering algorithm
	K-Medians	Variants of K-Means
	K-Medoids	
	K-Modes	
	Fuzzy C-Means	Points have fuzzy membership to clusters
	Mean shift	
	CLARANS	Designed for large-scale data sets
Probabilistic	EM	
Density-based	DBSCAN	
	GDBSCAN	Generalized extension of DBSCAN
	OPTICS	Produces a cluster-ordering of points
	DENCLUE	Based on the concept of influence functions
Grid-based	GRIDCLUS	
	STING	
	CLIQUE	Performs subspace clustering of high-dimensional data

points p and q are called *density-connected* if there is a third point o from which both p and q are density-reachable. A cluster is then a set of density-connected points which is maximal with respect to density-reachability. *Noise* is defined as the set of points in the data set not belonging to any of its clusters. The definition of density-based clusters assumes a distance function *dist(p, q)* for pairs of points. More formal definitions are provided in Ester (2014).

DBSCAN performs one region query per data point to retrieve its *Eps*-neighborhood. With the use of a spatial index such as an R-tree or X-tree, the runtime complexity for n

data points is *O(n log n)*. Without the use of an accelerating index structure, the runtime complexity increases to $O(n^2)$.

DBSCAN provides several advantages over other clustering algorithms. Unlike K-Means, it does not require the specification of the number of clusters beforehand. In addition, it can find clusters of arbitrary shape and size. It also has a built-in notion of noise which makes it robust to outliers. The major weakness of DBSCAN is that its global parameters cannot effectively cluster data sets that have large differences in local densities for different regions of the data space. The OPTICS algorithm addresses this challenge but it only visualizes the cluster structure without actually determining explicit clusters. Another weakness of DBSCAN is experienced in defining density for high-dimensional data sets.

When DBSCAN is applied to the problem of identifying hot spots in spatial data sets, its parameters directly influence the number and density of hot spot clusters. If the chosen value of *Eps* is too small, a large part of the data will not be clustered. A large cluster representing a single hot spot may be fragmented yielding two or more smaller hot spots. On the other hand, a very large value of *Eps* will cause smaller clusters to merge, resulting in fewer hot spots. Larger values of *MinPts* are usually better for data sets with noise and will yield more significant clusters. If the chosen value of *MinPts* is too small, closely spaced noise points will be incorrectly clustered while if it is too large, small clusters are likely to be labeled as noise.

2.4.2 Ordering Points To Identify the Clustering Structure

While DBSCAN can find clusters of arbitrary shapes, it cannot handle data with clusters of different densities, due to its use of a single density threshold. Many real-life data sets have an intrinsic cluster structure that cannot be characterized by global density parameters, and very different local densities may be needed to reveal clusters in different regions of the data space. The OPTICS algorithm addresses this challenge by producing a cluster-ordering of a data set with respect to its density-based clustering structure. This contains the information about every clustering level of the data set up to a generating distance *Eps*. OPTICS works in principle like an extended DBSCAN algorithm for an infinite number of distance parameters Eps_i which are smaller than the generating distance *Eps*. However, it does not assign cluster memberships but stores the order in which the points are processed. The clustering structure of a data set can be visualized graphically by a *reachability plot* that shows the reachability-distance values for all points sorted according to the clustering order (Ester, 2014).

2.4.3 Density-based Clustering

DENsity-based CLUstEring (DENCLUE) is another density-based clustering algorithm which generalizes the basic idea of density-based clusters beyond the distance-based *Eps*-neighborhood. It uses the concept of *influence functions* which mathematically model the influence of a data point in its neighborhood. Typical examples of influence functions are square wave functions or Gaussian functions. The density at some point is estimated

by the sum of the influences of all data points. A point is said to be *density-attracted* to a density-attractor if they are connected through a path of high-density points. The efficient implementation of the DENCLUE algorithm is based on the observation that most data points do not contribute to the density function at any given point of the data space. This can be exploited by computing only a local density function, while guaranteeing tight error bounds (Ester, 2014).

2.5 DBSCAN Implementations

There are a number of existing DBSCAN implementations provided by various software packages and environments. This section reviews the most common implementations and provides a comparative summary in table 2.5.

Table 2.5: Comparison of DBSCAN implementations

	Geo Distance Function	**Language**	**Index Structure**
ELKI	Yes	Java	R*-Tree, M-Tree, KD-Tree
R fpc package	No	R	-
R dbscan package	No	R, C++	KD-Tree
Weka	No	Java	-
SPMF	No	Java	KD-Tree
scikit-learn	No	Python	KD-Tree, Ball-Tree

The open-source Environment for DeveLoping KDD-Applications Supported by Index-Structures (ELKI) software framework provides an efficient implementation of DBSCAN. It has full native support for geographical distance functions (ELKI Development Team, 2014a) and the WGS 84 coordinate reference system (ELKI Development Team, 2014b). Further, ELKI provides the R*-Tree, M-Tree (also known as Ball-Tree) and K-Dimensional Tree (KD-Tree) index structures for accelerating geographical distance functions (ELKI Development Team, 2014c), which significantly improves algorithm performance.

In contrast, none of the other available DBSCAN implementations provides geographical distance functions and index structures to support them. The implementations in the R statistical software 'fpc' package (Oehlschlaegel, 2015) and 'dbscan' package (Hahsler, Arya & Mount, 2015) can only use either a pre-computed distance matrix or the Euclidean distance function, which is unsuitable for geospatial data. While the distance matrix is suitable and relatively fast, it requires pre-computation and is also expensive in memory usage. The 'fpc' package implementation has a quadratic runtime complexity ($O(n^2)$) since it does not use any index structures, while the 'dbscan' package implementation is significantly faster and can work with larger data sets since it uses the KD-Tree index structure for the nearest neighbor search. Both of the R implementations are less suitable for geospatial data when compared to the ELKI implementation.

Weka is an open-source machine learning and data mining toolkit which contains a basic implementation of DBSCAN that is not intended to be used as a reference for runtime benchmarks (Schubert, Melnikova-Albrecht & Holzmann, 2014). It does not support geographical distance functions (Kibriya, 2014). Sequential Pattern Mining Framework (SPMF) is an open-source data mining library that offers a minimalistic implementation of DBSCAN for Euclidean distance only. It uses a KD-Tree to store points internally in order to avoid having $O(n^2)$ runtime complexity (Fournier-Viger, 2015). The scikit-learn open source machine learning library includes a Python implementation of DBSCAN which also does not provide native support for geographical distance functions. However, it has support for the KD-Tree and Ball-Tree index structures for computing pointwise distances and finding nearest neighbors (scikit-learn developers, 2014).

2.6 Related Work

2.6.1 The Use of Cluster Analysis for Identifying Hot Spots in Spatial Data Sets

A number of studies have focused on the application of cluster analysis to the problem of identifying hot spots in spatial data sets. Recent research by Vadrevu et al. (2013) used the K-Means algorithm to identify hot spot regions of fire clusters in diverse geographical regions of India based on MODIS active fire data for the years 2010 and 2011. Their study restricted the number of clusters (K) to eight to depict major biomass burning regions and used the standard deviation ellipses (with one standard deviation) to identify the cluster locations. Although their findings indicate that K-Means was useful in identifying fire hot spots in diverse geographical regions, the hot spots were not defined by density of fire points. They covered geographically vast regions of the Indian subcontinent and included widely varying fire point densities. The study is also limited in that the heuristic applied in selecting the number of clusters *a priori* is not applicable to the general case of identifying hot spots from data. This work provides the opportunity to use density-based clustering as an alternative for identifying fire hot spots and to further evaluate its performance in this task.

In their study on the suitability of clustering algorithms for crime hot spot analysis, Divya, Rejimol and Selvan (2014) define a hot spot as a high concentration area of some activity. They conducted an experiment to evaluate the suitability of hierarchical clustering, K-Means, and DBSCAN in crime hot spot analysis. They ran the algorithms over a spatial data set of about 300 crime incidents and compared them on criteria such as number of clusters, average number of elements in the clusters, running time, and the Davies-Bouldin index. Their results show that DBSCAN achieved the best performance which they attribute to its density-based model. Further, they note its ability to identify noise and automatically discover the number of clusters as attractive qualities. These results provide an opportunity to investigate the suitability of DBSCAN in other application domains. In particular, it appears that further investigation of the performance of the algorithm on the MODIS active fire data set would be useful in assessing its effectiveness as a tool for hot spot analysis of spatial data.

Usman, Sitanggang and Syaufina (2015) applied the DBSCAN algorithm to determine the areas that have a high density of hot spots in the peat land area of Sumatra, Indonesia for the years 2002 and 2013. The study defined a hot spot as an area of pixels in satellite imagery that had higher temperature than surrounding areas. In this sense, "hot spot" refers to raw data rather than information derived from raw data analysis. The objective of their study was therefore not to discover hot spots in the data, but to analyze the pattern of distribution of these hot spots. This study analyzed the hot spot distribution in the clustered regions based on the physical characteristics of the peat land. The R statistical software was used for clustering the data. The *Eps* and *MinPts* parameters were determined using the *k-dist* method. The clustering result was evaluated using the Sum Square Error (SSE) method. The results report that the study found changes in the pattern of distribution of hot spots in the peat land between the two years. This work used fire data from the Ministry of Forestry of Indonesia. It also provides an opportunity to apply DBSCAN to the MODIS active fire data set.

A study by Palumbo (2013) discussing the relation between fire activity and the change of land cover in Kenya over 20 years used the MODIS active fire data for 2002 to 2012 to determine the fire activity. This was indicated by fire density, defined as the number of fire counts per area, on a scale from 0 to 1. The areas were plotted on a map as concentric circular regions with incremental radii of 25 km. Unlike density-based clustering, this definition of density does not take into account the connectedness of fire points and does not have an explicit notion of noise based on a threshold measure. Further, the results provide a summary of areas in Kenya with high fire activity over the 11 years under study at a low spatial resolution. Since the focus of analysis was the whole of Kenya rather than WPAs, the results are insufficient for identifying areas within WPAs that are fire hot spots. This work therefore preserves the need for a study to identify fire hot spots from MODIS data, within Kenya's WPAs.

2.6.1.1 Definition of a Hot Spot

A common theme emerging from the review of related studies is the varying definition of a "hot spot". The two major concepts of a hot spot identified in the literature are: (i) a raw data element representing a fire event that is observed in the environment, and (ii) a region of high intensity of an activity that is discovered after analysis of raw data concerning that activity. For the purposes of this study, we concretely define a fire hot spot as: *a connected, dense region in the data space that indicates a spatial concentration of fire incidents*. This definition is purposely aligned with the definition of clusters in density-based clustering and therefore, a density-based cluster identified with appropriate parameters indicates a fire hot spot.

2.6.2 Web Applications Providing Visualization of Fire Data

There are a number of web-based applications that provide visualization of MODIS active fire data. The FIRMS service provides the Web Fire Mapper (NASA FIRMS, 2014) which

integrates global MODIS active fire locations with other geospatial layers and delivers this combined information through web mapping services. It allows users to query active fire locations derived from MODIS data approximately 2-4 hours after satellite overpass. Davies et al. (2008) have also developed a web mapping service for South Africa, known as Advanced Fire Information System (AFIS). It allows users to view and query fire detections by time period.

The European Commission's Joint Research Centre (JRC) developed a web application for monitoring vegetation fires in the protected areas of the African, Caribbean, and Pacific (ACP) countries, including Kenya (Palumbo et al., 2013). The tool provides locations of active fires and burned area extent derived from MODIS data. It is designed to produce graphs, tables, and maps of the fire activity for a selected protected area and period of time. Sentinel (Geoscience Australia, 2015) is a national bushfire monitoring system that provides timely information about hot spots to emergency service managers across Australia. The mapping system allows users to identify fire locations with a potential risk to communities and property. It extracts hot spot information from images acquired by MODIS and other remote sensing satellites and provides visualization via a Google Maps-driven web interface.

None of these web-based applications provides the visualization of fire hot spot clusters generated using data clustering methods. They have all focused on the mapping of active fire locations. The JRC application provides a fire density map for which the density is defined as the number of fire counts per specified area. This method is different from density-based clustering. It does not take into account the density-connectedness of fire points. However, it provides a useful basis for comparison.

Oliveira and Souza Baptista (2013) implemented GeoSTAT (Geographic Spatio-Temporal Analysis Tool), a web-based application for the analysis and visualization of spatio-temporal data. The application uses the Google Maps Application Programming Interface (API) to offer a dynamic map and accesses spatial or spatio-temporal data published in servers that implement the Open Geospatial Consortium (OGC) WMS (Web Map Service) and WFS (Web Feature Service) services. Spatial and temporal filters are used to select and query the visualized data. In addition, a data mining component provides access to seven clustering algorithms from the Weka toolkit. The DBSCAN algorithm implemented in the system was used to cluster two data sets of power line failures and fire hot spots. The visualization of the generated clusters could be used to confirm the hypothesis that some of the fire hot spots are the cause of failures in power transmission lines.

The application developed in this study focused on providing a general solution for spatio-temporal analysis which was as flexible as possible with regards to application domain independence, diverse data sources, and a wide array of data clustering algorithms. The interactiveness of the application's data mining component was impacted by the fact that it takes a long time to cluster huge data sets in real-time.

GeoClustering is a geospatial clustering web service developed by Wang, Wang and Liang (2011). It enables users to cluster online data sources using the DBSCAN algorithm.

The clustering results can be visualized through a web mapping interface powered by the Google Maps API. The clusters are represented by convex hull polygons. GeoClustering emphasizes openness and interoperability by adopting the Extensible Markup Language (XML) as the data interchange format. It also provides an API through which web applications and services can access the geospatial clustering service. This is accomplished using a Hypertext Transfer Protocol (HTTP) GET request specifying the XML data file and DBSCAN parameters.

GeoClustering appears to meet the requirements for visualizing the MODIS fire hot spot clusters. However, the developed prototype was not available for evaluation and use. In addition, the study did not provide the implementation details of DBSCAN. The study reports on an evaluation of GeoClustering that was conducted for two datasets. A crime data set from the Calgary Police Online Crime Map had 414 data points. The average clustering time in one test environment was 6.4 seconds. A second data set of earthquake data from the USGS National Earthquake Information Center contained 204 data points. The average runtime was approximately 1.1 seconds. Both of these data sets are relatively small in size. It remains unknown how the runtime complexity of the DBSCAN implementation used in GeoClustering changes for larger data sets such as MODIS.

Chapter 3

Methodology

The system development methodology used in this study was **Incremental with Iterative Prototyping**. This was conducted in two phases: (i) Preliminary Analysis which covered Research Objective 1, and (ii) Web Application Development which covered Research Objective 2. The activities carried out in the first phase were: sourcing of the MODIS active fire data, preprocessing the data, and performing the density-based cluster analysis using the DBSCAN algorithm. The activities of the second phase were: requirements definition for the application, design, coding, testing, and deployment of the completed application. Figure 3.1 illustrates the methodology and figure 3.2 presents a schematic diagram of the system architecture.

This methodology was selected because it combined linear and iterative system development approaches. The first phase was linear while the second was iterative. In addition, the iterative prototyping of the application facilitated a user-centered development process while providing the flexibility to accommodate changes in the user requirements. By improving user participation, it became easier to validate the requirements and identify missing functionality at an early stage.

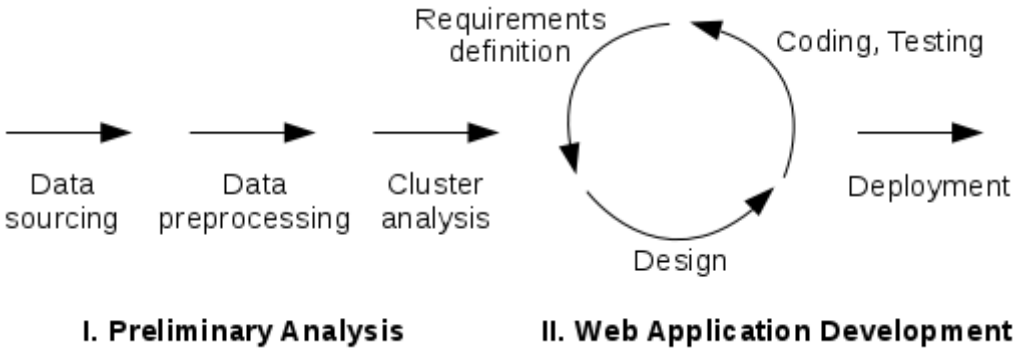


Figure 3.1: System development methodology

3.1 Preliminary Analysis

3.1.1 Data Sourcing

The study used the Standard MODIS active fire data (MCD14ML) for identifying the fire hot spots in Kenya's WPAs. This data set was selected because it is processed by the MODIS Fire SCF at the University of Maryland to provide a higher quality suitable for scientific publications as compared to the near real-time data (MCD14DL) (NASA Earthdata, 2015).

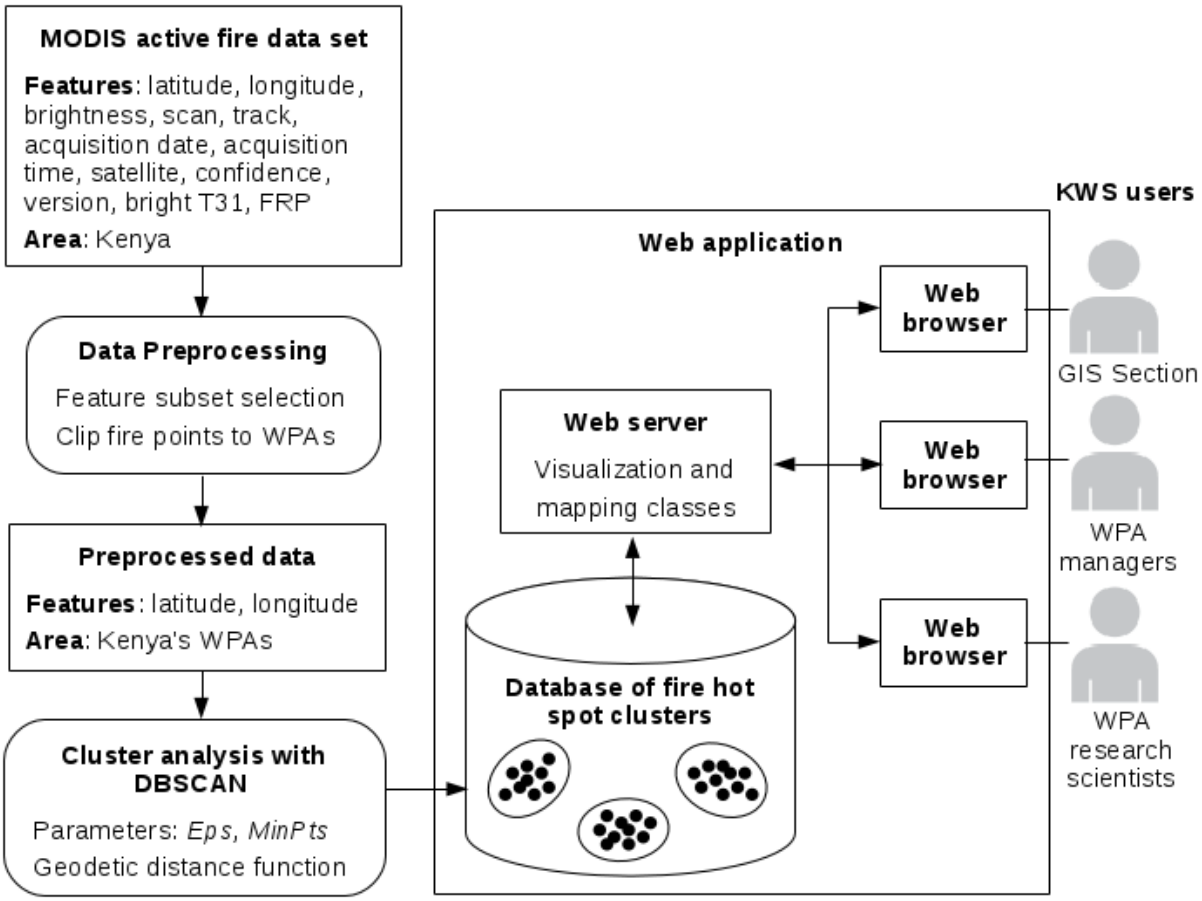


Figure 3.2: Schematic diagram of the system architecture

The publicly accessible MCD14ML data set was downloaded from the NASA FIRMS server using the Archive Download Tool available online (NASA FIRMS, 2015). A NASA FIRMS Archive Download request was placed with a custom polygon around Kenya's borders indicating the area of interest, while the time period extended from January 2003 to May 2015. The data format was selected as CSV text file to provide the raw fire points in a form that was easy to process using scripting tools. The request was processed in about 42 minutes after which an e-mail was sent from NASA FIRMS with a Uniform Resource Locator (URL) providing the link from which to download a ZIP file of the data. The text file containing the MCD14ML data set was named firms2186714310171011_MCD14ML.csv.

Although MODIS data is available beginning in November 2000, data from 1st January 2003 to 31st December 2014 was used in the study so as to include only complete calendar years during which high quality data from both the Aqua and Terra satellites is available. Complete years were selected so as to reflect the full Kenya fire season (January-March and June-September).

3.1.2 Data Preprocessing

3.1.2.1 Feature Subset Selection

A number of tasks were carried out on the MODIS fire data set during the preprocessing stage. The feature subset selection was done with the open-source GAWK software, version 4.1.1. GAWK is the GNU implementation of the AWK programming language. It was selected for this task because it excels at performing complicated text processing tasks with short, single-line programs.

Before preprocessing, the downloaded MODIS MCD14ML active fire data set had 12 attribute fields (features) and 107,848 fire points. Some of the fire data in the CSV text file named firms2186714310171011_MCD14ML.csv is presented in listing A.1.

Only the latitude and longitude were needed for the cluster analysis in this study. These geographic coordinates were used by the DBSCAN algorithm to identify the fire hot spot clusters. The following GAWK command was executed on a Debian GNU/Linux system to extract the latitude and longitude from the input CSV file. In the output file (MCD14ML.txt), the longitude was saved first followed by the latitude in order for the axes to appear correctly in the ELKI visualization graph. In addition, fire data for the year 2015 was excluded from the output file since it did not yet contain a complete fire season.

Note: The '$' character in the following and subsequent commands indicates a command prompt in a Linux terminal program.

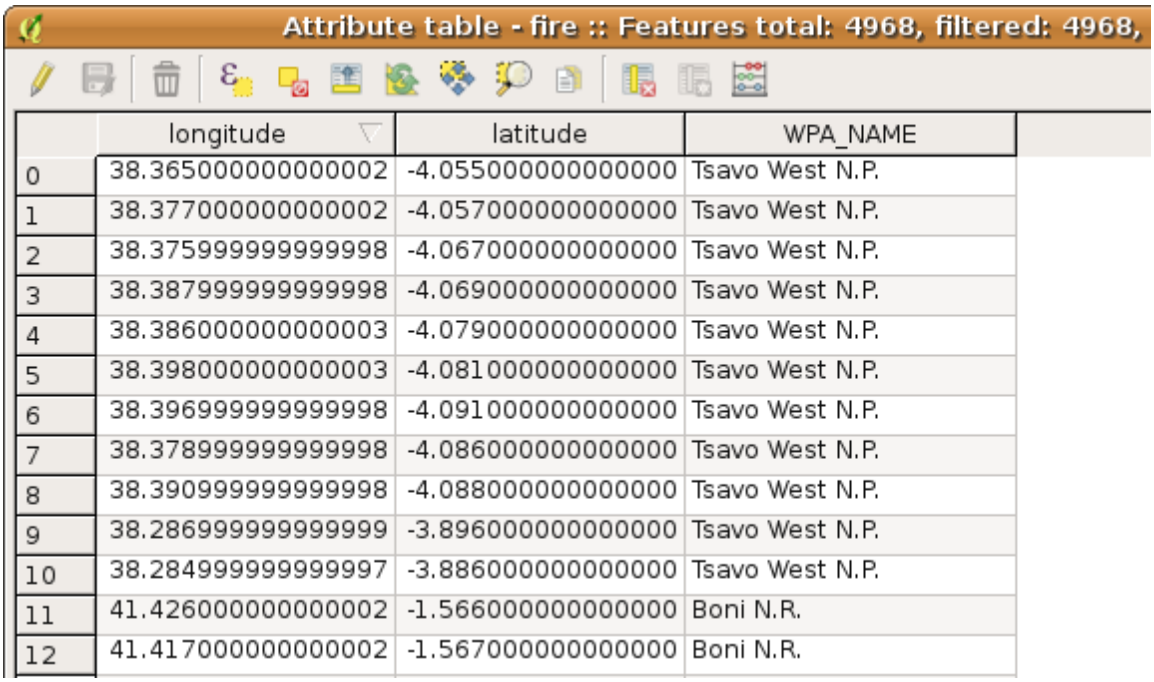
```
$ echo "longitude latitude" > MCD14ML.txt
$ awk -F, -v 'OFS= ' 'NR > 1 && $6 !~ /^2015/ { print $2, $1 }' \
firms2186714310171011_MCD14ML.csv >> MCD14ML.txt
```

After the feature subset selection was performed, the MODIS MCD14ML active fire data set was reduced to 2 relevant attribute fields (features) and 104,239 fire points. The data in the text file named MCD14ML.txt was in the format shown in listing A.2.

3.1.2.2 Geoprocessing

The open-source QGIS software, version 2.8.2, was used to create a map with a vector layer of Kenya's WPAs. The data for this layer was acquired from the KWS GIS Section in Environmental Systems Research Institute (ESRI) shapefile format. This WPA shapefile was produced in November 2008. The WPAs layer had the WPA_NAME and WPA_ID as the fields in its attribute table. After this layer was added, the MCD14ML.txt file produced by GAWK was also added to the map as a delimited text layer. World Geodetic System (WGS) 84 (European Petroleum Survey Group (EPSG):4326) was selected as the coordinate reference system for the two layers.

MODIS fire points falling outside the WPAs layer were clipped using the Intersect geoprocessing tool under the Vector menu in QGIS. For this operation, the MCD14ML.txt layer was selected as the input vector layer while the WPAs layer was selected as the intersect layer. The resulting shapefile layer, named 'fire', had 4,968 fire points. It had the longitude, latitude, WPA_NAME, and WPA_ID as the fields in its attribute table. These fields resulted from the intersection of the fields in the two contributing layers. The WPA_ID field was not useful since it was not a unique key. It was therefore deleted using the 'Delete Column' feature in the attribute table. Figure 3.3 shows the three fields and some of the records in the attribute table. Figure 3.4 shows the map of the fire points overlaid on the WPAs.



	longitude	latitude	WPA_NAME
0	38.365000000000002	-4.055000000000000	Tsavo West N.P.
1	38.377000000000002	-4.057000000000000	Tsavo West N.P.
2	38.375999999999998	-4.067000000000000	Tsavo West N.P.
3	38.387999999999998	-4.069000000000000	Tsavo West N.P.
4	38.386000000000003	-4.079000000000000	Tsavo West N.P.
5	38.398000000000003	-4.081000000000000	Tsavo West N.P.
6	38.396999999999998	-4.091000000000000	Tsavo West N.P.
7	38.378999999999998	-4.086000000000000	Tsavo West N.P.
8	38.390999999999998	-4.088000000000000	Tsavo West N.P.
9	38.286999999999999	-3.896000000000000	Tsavo West N.P.
10	38.284999999999997	-3.886000000000000	Tsavo West N.P.
11	41.426000000000002	-1.566000000000000	Boni N.R.
12	41.417000000000002	-1.567000000000000	Boni N.R.

Figure 3.3: Fire layer attribute table in QGIS

All the data in the attribute table was copied to a temporary text file named tmp. This tab-separated file had an additional first field called wkt_geom which indicated the longitude and latitude as being of the POINT shape. Since it was unnecessary, it was removed with the following GAWK command to produce a CSV text file, named MCD14ML_WPA.csv, which contained only the 4,968 MODIS fire points falling within Kenya's WPAs. The names of the WPAs in this file were edited to standardize acronyms and correct spelling mistakes. Listing A.3 shows some of the records in this file.

```
$ awk -F '\t' -v 'OFS=,' '{ print $2, $3, $4 }' tmp > MCD14ML_WPA.csv
$ rm tmp
```

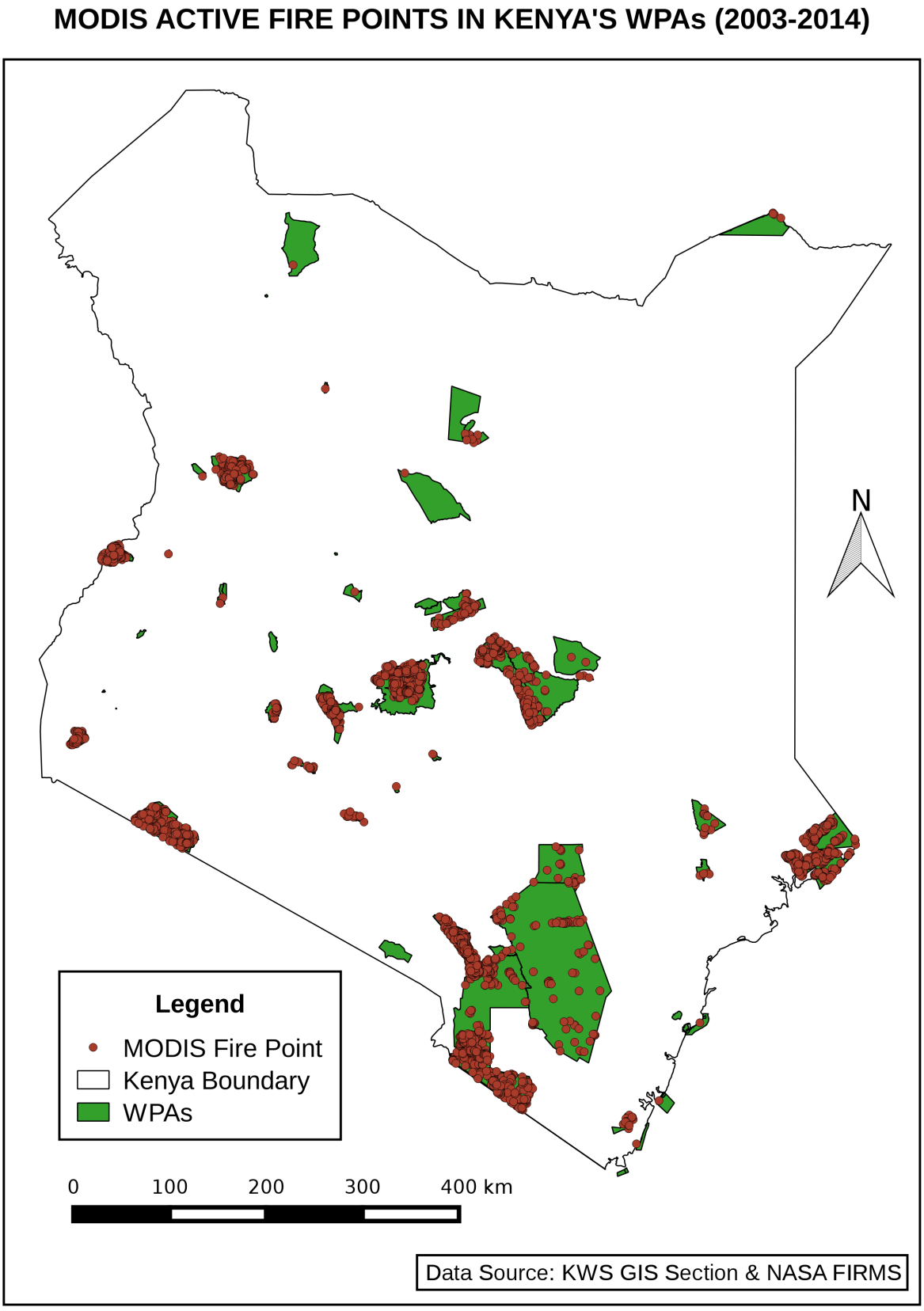


Figure 3.4: MODIS active fire points falling within Kenya's WPAs (2003-2014)

3.1.2.3 Computing the Frequency Distribution of MODIS Fire Points

After the MCD14ML_WPA.csv text file had been created, Structured Query Language (SQL) statements were used to process it to create an output CSV text file containing a frequency distribution of the preprocessed MODIS fire points. This included the number of fire points (absolute frequency) and the corresponding percentage of fire points (relative frequency) in each WPA. The WPA was used as a categorical variable in the frequency distribution.

The SQL statements used to perform this task were written in a text file named fdist.sql. Its entire contents are reproduced in listing A.4. The SQL statements were then executed by the MySQL open-source Relational Database Management System (RDBMS), version 5.6.25-4, under the Debian GNU/Linux environment. They imported the 4,968 fire points from the input CSV text file into a database table, computed the frequencies for each WPA that recorded fire activity, and stored them in descending order in an output CSV text file named fdist.csv. The /tmp/ directory was used to temporarily store both the input and output CSV files in order to prevent permission errors resulting from the use of other directories.

The command used to invoke the MySQL command-line tool to execute the SQL statements in the fdist.sql script file is shown below. After this execution, two GAWK commands were used to verify the total values of the frequencies. The result of each command appears below it.

```
$ mysql -u root -p < fdist.sql
Enter password:
@n := COUNT(lat)
4968
$ awk -F, '{ sum += $2 } END {print sum }' fdist.csv
4968
$ awk -F, '{ sum += $3 } END { print sum }' fdist.csv
99.98
```

The fdist.csv file contained 45 lines of data, one for each of the 45 WPAs that had fire points falling within it. Each line had three fields: the WPA Name, the Absolute Frequency, and the Relative Frequency (as a percentage). The data was sorted in descending order, from the largest to the smallest frequency. Some of the records in the file are presented in listing A.5.

Octave was used to plot a horizontal bar graph of the frequency distribution using the data in the fdist.csv text file. An Octave script file named fdist.m contained the code shown in listing A.6, which created a Portable Network Graphics (PNG) image file of the plot. The following command line was used to execute the code in the Octave script file:

```
$ octave fdist.m
```

Octave was also used to produce a scatterplot diagram of the preprocessed MODIS fire points contained in the MCD14ML_WPA.csv text file. An Octave script file named scatterplot.m contained the code shown in listing A.7, which created a PNG image file of the scatterplot. The following command line was used to execute the code in the Octave script file:

```
$ octave scatterplot.m
```

3.1.3 Estimation of DBSCAN Parameters with a Sorted *k*-dist Graph

The density-based cluster analysis of the MODIS fire data was performed using the DBSCAN clustering algorithm as implemented in the ELKI software framework, version 0.6.5∼20141030 (Achtert et al., 2013). DBSCAN was selected because it did not require the specification of the number of clusters beforehand. In addition, it could find clusters of arbitrary shape and size and had a built-in notion of noise which made it robust to outliers. Unlike DENCLUE, it uses a straight-forward distance-based notion of the range parameter *Eps*. The algorithm was also preferred over OPTICS because it could create a hard clustering of the data. Since the data for this study consisted of 2-dimensional geographic coordinates, DBSCAN was suitable because the definition of density and distance was straightforward.

The ELKI software framework was selected because its DBSCAN implementation provided a number of advantages over other existing implementations. It had full native support for geographical distance functions and the WGS 84 coordinate reference system. Further, it provided the R*-Tree, M-Tree, and KD-Tree index structures for accelerating geographical distance functions to achieve improved algorithm performance.

Before the cluster analysis could be performed, it was necessary to determine the initial values of the two parameters of the DBSCAN algorithm: *Eps* and *MinPts*. The range parameter *Eps* was initially estimated using the *sorted k-dist graph* heuristic, with $k = 4$, as proposed by Ester et al. (1996) in the original DBSCAN paper. This graph provided a plot of the sorted distances of each point in a data set to its k-th (4-th) nearest neighbor in order to indicate the density distribution in the data set. This method of estimating *Eps* specified setting it to the k-dist value for the threshold point p in the first "valley" (at the bend or "knee") of the graph.

The paper proposed setting k to 4 for 2-dimensional data since their experiments indicated that the k-dist graphs for $k > 4$ did not significantly differ from the 4-dist graph but required considerably more computation. This was considered appropriate for the preprocessed MODIS fire data set in which the two geospatial dimensions were latitude and longitude. As a result, *MinPts* was set to 5 ($k + 1$) for the ELKI implementation. This is because the version of ELKI used in this study defined *MinPts* as the minimum number of points in the *Eps*-neighborhood of a point. This included the point itself and its neighbors, hence the smallest possible cluster contained 5 fire points.

ELKI was used to produce the sorted k-dist values for the graph because it provided the necessary geographical distance function. Its "KNNDistancesSampler" algorithm was run on the MCD14ML_WPA.csv input text file with the parameters specified as follows: the "LngLatDistanceFunction" distance function was selected because the input file had the longitude and latitude as the first and second columns respectively; WGS 84 was selected as the spheroid earth model; k was set to 4; and the output was saved in a text file named knn-distances.txt within the k_dist directory. Figure 3.5 shows the ELKI MiniGUI window running the algorithm. The following Linux command line achieves the same effect:

```
$ java -cp /usr/share/java/elki.jar \
de.lmu.ifi.dbs.elki.application.KDDCLIApplication \
-dbc.in MCD14ML_WPA.csv \
-algorithm KNNDistancesSampler \
-algorithm.distancefunction geo.LngLatDistanceFunction \
-geo.model WGS84SpheroidEarthModel \
-knndistanceorder.k 4 \
-resulthandler ResultWriter \
-out k_dist
```

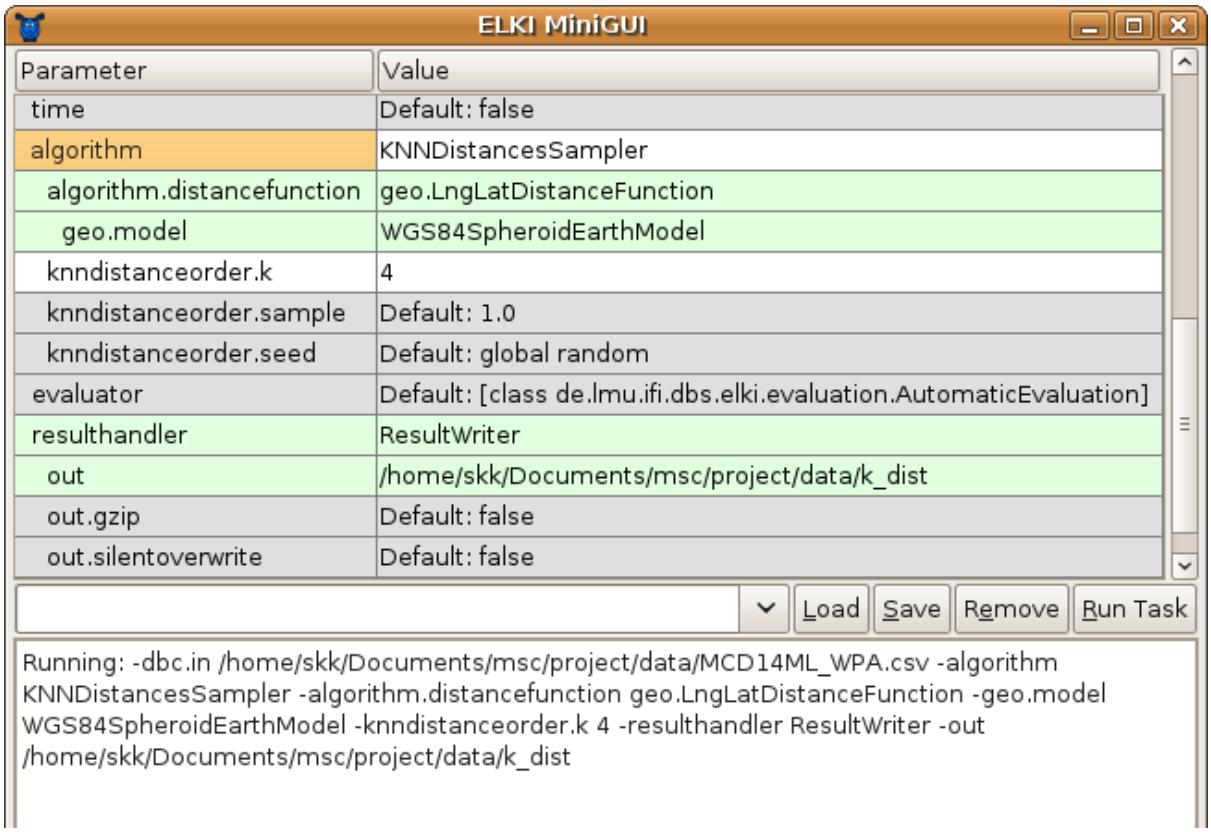


Figure 3.5: Computing the sorted k-dist values in ELKI

The file knn-distances.txt contained a list of the distances of each of the 4,968 MODIS fire points to its k-th (4-th) nearest neighbor sorted in ascending order from smallest to largest. The distances ranged from 111.28 m to 387.47 km. Some of the smallest and largest distances are provided in listing A.8 as they appear in the file, in meters.

GNU Octave, version 4.0.0, was used to plot the sorted k-dist graph with the k-Nearest Neighbor (KNN) distances listed in the knn-distances.txt text file. The series of Linux and Octave commands used to achieve this is provided below:

```
$ octave
>> x = load('-ascii', 'knn-distances.txt');
>> size(x)
ans =

   4968      1

>> plot(x)
>> xlabel("MODIS Fire Points");
>> ylabel("Distance to 4-th Nearest Neighbor (m)");
>> title("Sorted KNN-Distance Graph (k = 4)");
>> grid("on");
```

The following Octave commands were appended to those shown above to produce a plot of the threshold point with annotations showing the point itself, the noise region, and the clusters region.

```
>> line([4906, 4906], [0, 13564]);
>> annotation("textarrow", [0, 0], [0, 0], "position", [0.75, \
0.45, -0.1, 0.1], "headstyle", "vback3", "string", "threshold \
point", "fontsize", 11);
>> annotation("textbox", [0.73, 0.2, 0.1, 0.1], "string", "noise", \
"fontsize", 11, "edgecolor", "white");
>> annotation("textbox", [0.4, 0.2, 0.1, 0.1], "string", \
"clusters", "fontsize", 11, "edgecolor", "white");
```

3.1.4 Initial Clustering with Estimated DBSCAN Parameters

The ELKI DBSCAN implementation was executed under the Debian GNU/Linux environment on a computer with a first generation Intel Core i5 processor and 4 GB of main memory. The initial parameters were as estimated from the sorted k-dist graph. *Eps* was set to 13,564 m (13.564 km) while *MinPts* was set to 5. The algorithm was run without any index structure and with the R*-Tree, M-Tree, and KD-Tree index structures for performance comparison. The parameters used for each case are outlined below.

1. No index structure:

```
-dbc.in MCD14ML_WPA.csv
-algorithm clustering.DBSCAN
-algorithm.distancefunction geo.LngLatDistanceFunction
-geo.model WGS84SpheroidEarthModel
-dbscan.epsilon 13564.0
-dbscan.minpts 5
```

2. R*-Tree index structure (shown in figure 3.6):

```
-dbc.in MCD14ML_WPA.csv
-db.index tree.spatial.rstarvariants.rstar.RStarTreeFactory
-pagefile.pagesize 4096
-rtree.reinsertion-distancce geo.LngLatDistanceFunction
-geo.model WGS84SpheroidEarthModel
-spatial.bulkstrategy SortTileRecursiveBulkSplit
-algorithm clustering.DBSCAN
-algorithm.distancefunction geo.LngLatDistanceFunction
-geo.model WGS84SpheroidEarthModel
-dbscan.epsilon 13564.0
-dbscan.minpts 5
```

3. M-Tree index structure:

```
-dbc.in MCD14ML_WPA.csv
-db.index tree.metrical.mtreevariants.mtree.MTreeFactory
-pagefile.pagesize 4096
-mtree.distancefunction geo.LngLatDistanceFunction
-geo.model WGS84SpheroidEarthModel
-algorithm clustering.DBSCAN
-algorithm.distancefunction geo.LngLatDistanceFunction
-geo.model WGS84SpheroidEarthModel
-dbscan.epsilon 13564.0
-dbscan.minpts 5
```

4. KD-Tree index structure:

```
-dbc.in MCD14ML_WPA.csv
-db.index tree.spatial.kd.MinimalisticMemoryKDTree$Factory
-algorithm clustering.DBSCAN
-algorithm.distancefunction geo.LngLatDistanceFunction
-geo.model WGS84SpheroidEarthModel
-dbscan.epsilon 13564.0
-dbscan.minpts 5
```

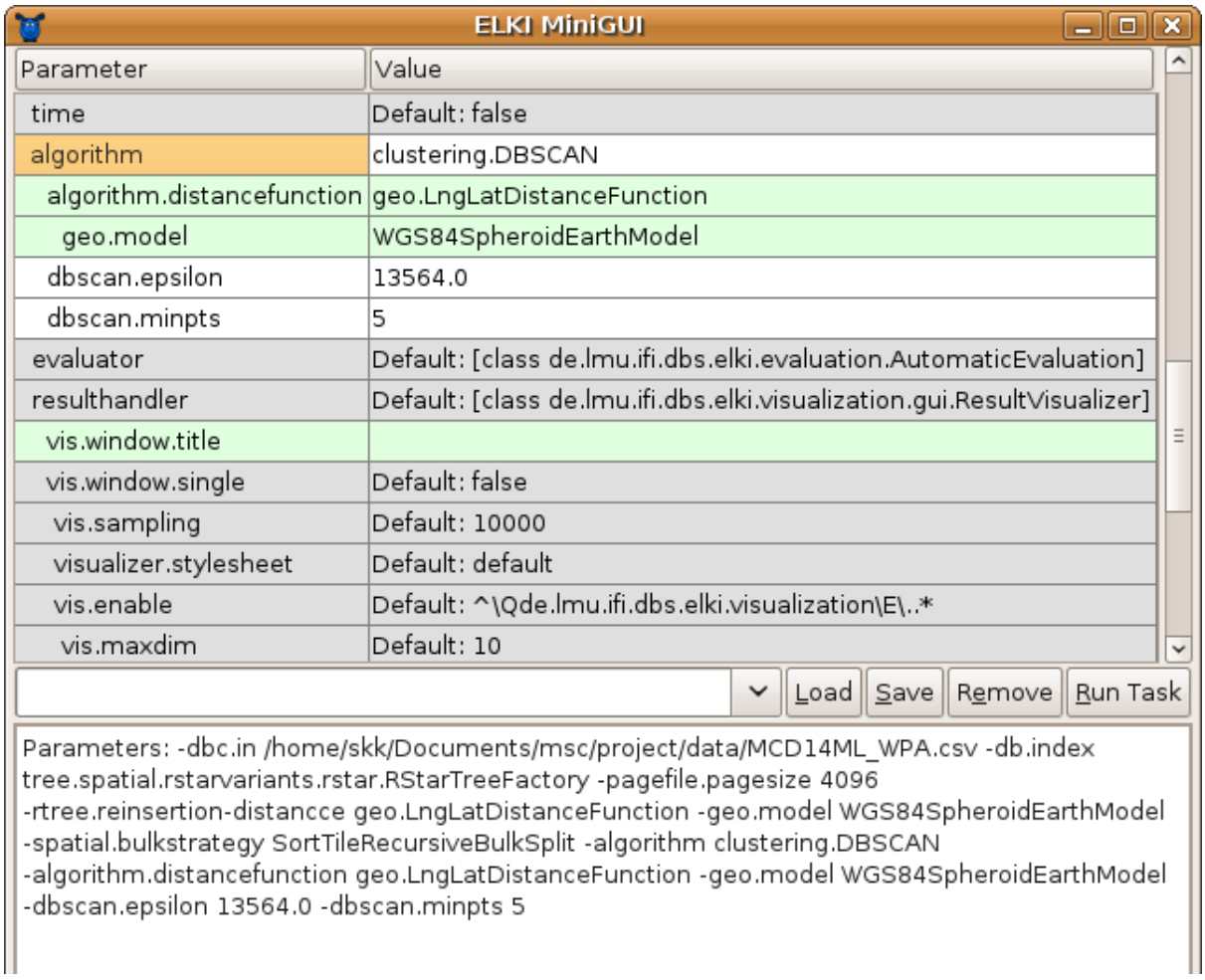


Figure 3.6: Running ELKI DBSCAN with the R*-Tree index and initial parameters

Each of the above cases was executed three times and the time taken for each run was recorded. The three runtimes for each case were then averaged to produce the mean runtime. The scatterplot diagram of the fire points produced by ELKI for the initial DBSCAN parameters was exported from the ELKI visualization window in Scalable Vector Graphics (SVG) format after which it was edited for presentation with the open-source Inkscape vector graphics editor software, version 0.91.

3.1.5 Trial Runs with Different DBSCAN Parameters

After DBSCAN had been executed with the initial parameters, 25 trial runs with different values were performed to determine the pair of parameters (*Eps* and *MinPts*) that gave a clustering result that was suitable for the MODIS fire data in Kenya's WPAs. All the trial runs were conducted with the R*-Tree index structure for efficiency reasons. The values used for *MinPts* were: 5, 7, 8, 10, and 12. Each of these values was paired with the following values for *Eps*: 500 m, 600 m, 700 m, 800 m, and 900 m. Since *Eps* is a radius, the area of the *Eps*-neighborhood was calculated (in square kilometers) as the

area of the circle with the radius *Eps* (in meters) as shown in equation (3.1).

$$Eps\text{-neighborhood} = \pi \times \left(\frac{Eps}{1000}\right)^2 \tag{3.1}$$

The clustering result of each pair of DBSCAN parameter values used in the trial runs was evaluated by examining its ELKI scatterplot diagram and the number of clusters produced. The most suitable parameter values were selected based on their ability to produce significant clusters in the presence of noise. In addition, the clusters had to be small enough in size to be identified *within* WPAs.

3.1.6 Final Clustering with the Most Suitable DBSCAN Parameters

The trial run parameters selected as the most suitable for the MODIS fire data in Kenya's WPAs were $MinPts = 7$ and $Eps = 700$ m. DBSCAN was executed with the selected parameter values and two different ELKI result handler parameters. The first result handler parameter was set to the *default* "ResultVisualizer" class which produced a Graphical User Interface (GUI) visualization for the scatterplot diagram. The parameters for this case are presented below.

```
-dbc.in MCD14ML_WPA.csv
-db.index tree.spatial.rstarvariants.rstar.RStarTreeFactory
-pagefile.pagesize 4096
-rtree.reinsertion-distancce geo.LngLatDistanceFunction
-geo.model WGS84SpheroidEarthModel
-spatial.bulkstrategy SortTileRecursiveBulkSplit
-algorithm clustering.DBSCAN
-algorithm.distancefunction geo.LngLatDistanceFunction
-geo.model WGS84SpheroidEarthModel
-dbscan.epsilon 700.0
-dbscan.minpts 7
```

The second result handler parameter was set to the "ResultWriter" class which created an output text file for each identified cluster, plus noise, in a specified directory named cluster. There were no visualized results for this case. The parameters used are presented below. Figure 3.7 shows the ELKI MiniGUI window running DBSCAN with these parameters.

```
-dbc.in MCD14ML_WPA.csv
-db.index tree.spatial.rstarvariants.rstar.RStarTreeFactory
-pagefile.pagesize 4096
-rtree.reinsertion-distancce geo.LngLatDistanceFunction
```

```
-geo.model WGS84SpheroidEarthModel
-spatial.bulkstrategy SortTileRecursiveBulkSplit
-algorithm clustering.DBSCAN
-algorithm.distancefunction geo.LngLatDistanceFunction
-geo.model WGS84SpheroidEarthModel
-dbscan.epsilon 700.0
-dbscan.minpts 7
-resulthandler ResultWriter
-out cluster
```

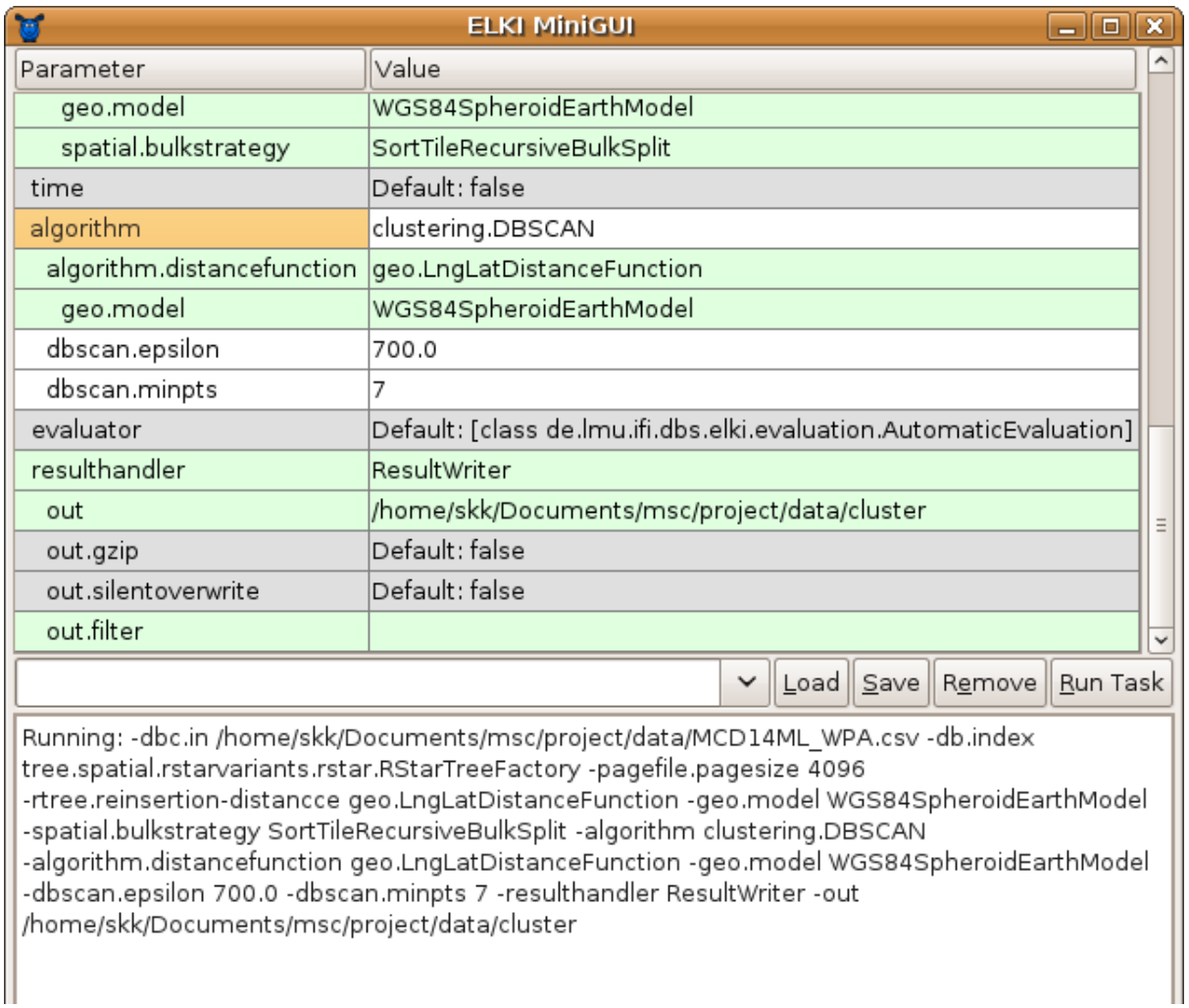


Figure 3.7: Running ELKI DBSCAN with $MinPts = 7$ and $Eps = 700$ m

The cluster directory created by executing DBSCAN with the ELKI result handler parameter set to the "ResultWriter" class had a total of 46 text files. 43 of these files were for each identified fire hot spot cluster. They were named cluster_id.txt where 'id' ranged from 0 to 42. For example, the text file for the first cluster was named cluster_0.txt. A text file named noise.txt contained the list of noise points. Statistical indices evaluating

the clustering result were written to a text file named cluster-evaluation.txt. Finally, a text file named settings.txt contained the ELKI parameters used to acquire the result. Figure 3.8 shows the list of the text files in the cluster directory.

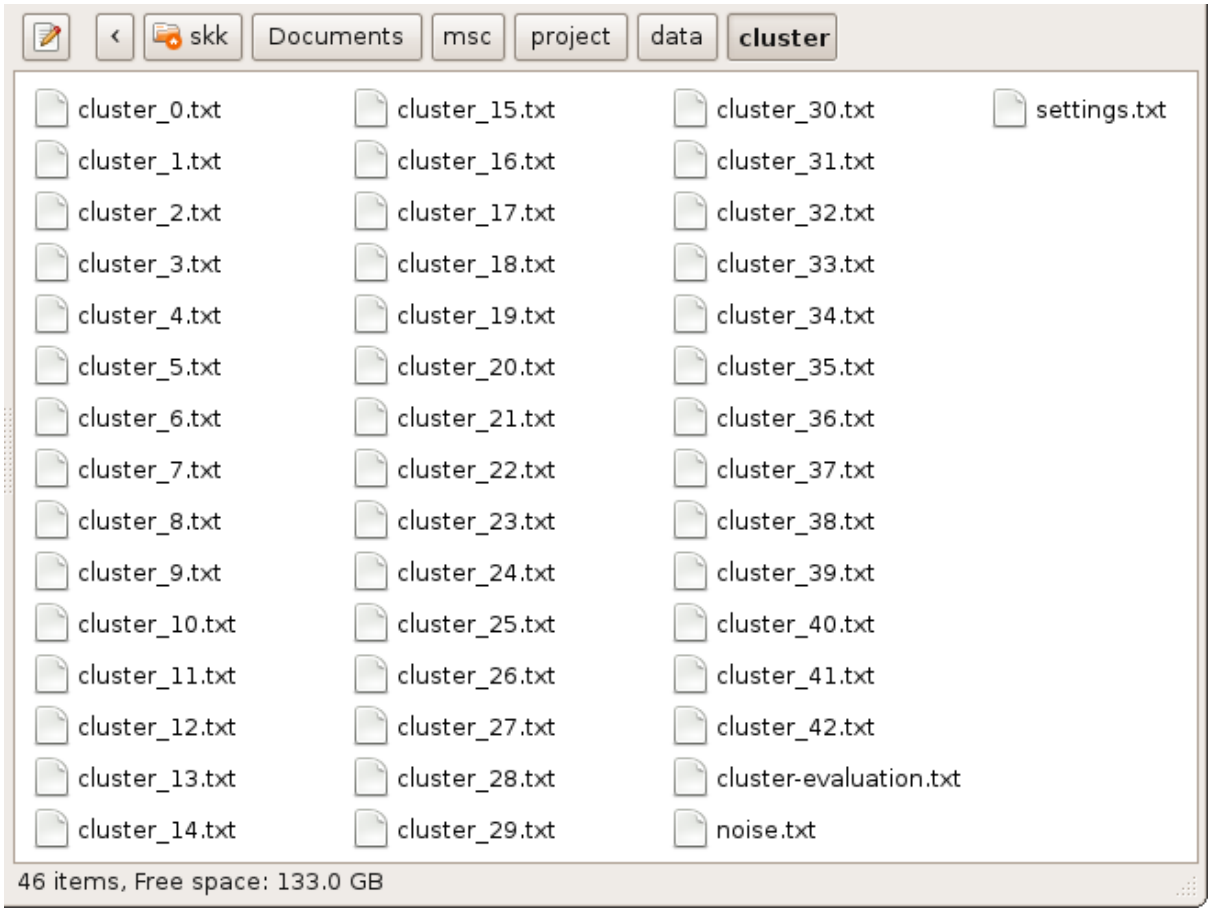


Figure 3.8: List of text files in the cluster directory

The text files containing the clusters and noise data all followed a standard format. The first four lines contained metadata describing the contents of the file. They began with a '#' character to mark them as comments. The metadata lines included: the cluster identifier (e.g. Cluster 0), cluster name (Noise or Cluster), cluster noise flag (a true or false value indicating whether the file contained noise or cluster data), and cluster size (number of fire points in the cluster).

The rest of the lines listed the fire points contained in the cluster, in four fields. ELKI prepended an identifier (ID) field to the three original fields that were present in the input file. This field showed the unique integer ID ELKI assigned to each fire point. It was an integer value ranging from 1 to 4,968 (the total number of fire points). The four fields were separated by a single space. Some of the data in the noise.txt and cluster_0.txt text files is shown in listings A.9 and A.10 respectively.

3.1.7 Computing the Frequency Distribution and Sizes of Identified Fire Hot Spot Clusters

A number of SQL queries were prepared and executed in the Microsoft SQL Server 2008 R2 database implemented for the web application developed in this study. The SQL query shown in listing A.11 was executed to produce a frequency distribution table of the fire hot spot clusters. This included the absolute and relative frequencies of the hot spot clusters that were identified in the WPAs.

The SQL query in listing A.12 was executed to produce a table of the number of fire points in each fire hot spot cluster. The result of the query was sorted from largest to smallest. Another SQL query shown in listing A.13 was executed to produce a table which showed the average number of fire points per km^2 for each WPA that had recorded fire activity.

3.2 Web Application Development

3.2.1 Requirements Definition

The functional requirements for the web application were gathered through unstructured interviews with the Head of the Geographic Information System (GIS) Section at KWS. After the initial requirements had been gathered, they were be used to build the first prototype which was tested by the KWS GIS Section for requirements validation. Subsequent iterations of development were used to refine the requirements specification. Table 3.1 summarizes the functional requirements. It lists four software attributes and the functional requirements identified under each of these attributes.

The primary requirement on accessibility resulted in the application being developed as a module of the KWS Integrated Database System (KWSIDS). KWSIDS is a web-based database system running on the Apache web server in a server computer hosted at the KWS headquarters. It is accessible to KWS staff both at the headquarters and in the field stations via the KWS Intranet.

3.2.2 Design

The web application was designed using a three-tier model. The bottom tier contained the relational database storing the application's data. The middle tier contained the application logic for handling queries on the data. The top tier contained the user interface through which the users provided input and received processed output from the application. Figure 3.9 shows the web application's three-tier model.

The database of fire hot spot clusters was modeled using an entity relationship diagram during this stage of development. Two tables containing data on the WPAs and the MODIS fire points and clusters were identified. Figure 3.10 shows the entity relationship model of the database.

Table 3.1: Web application functional requirements

Attribute	**Functional Requirement**
Accessibility	The application shall be developed as a module of the KWSIDS system to enable wide access by KWS staff via the KWS Intranet.
	The application shall be accessible to KWS staff both at the KWS headquarters and in the field stations connected to the KWS wide area network.
	The application shall be publicly accessible to all KWS staff without requirement for a log-in account on the KWSIDS system.
User Interface	The application shall display the location of each MODIS fire point on the map.
	The application shall display each fire hot spot cluster on the map.
	The application shall display all the WPA boundaries on the map.
	Users shall be able to query the fire points and hot spots by WPA.
	The application shall display summary statistics on the number of fire points and hot spot clusters, in all WPAs and in each selected WPA.
	The application should enable the user to print the map.
	The application should display fire points against satellite imagery for comparison with background vegetation.
Usability	The application shall enable the user to interact with the map through panning and zooming actions.
	The application shall provide a consistent user interface across all the major desktop web browsers used at KWS (these are Mozilla Firefox, Microsoft Internet Explorer, Google Chrome, and Opera).
Interoperability	The application shall enable the MODIS fire data set to be exported in CSV format for advanced spatial analysis with the ArcGIS software used by the KWS GIS Section.

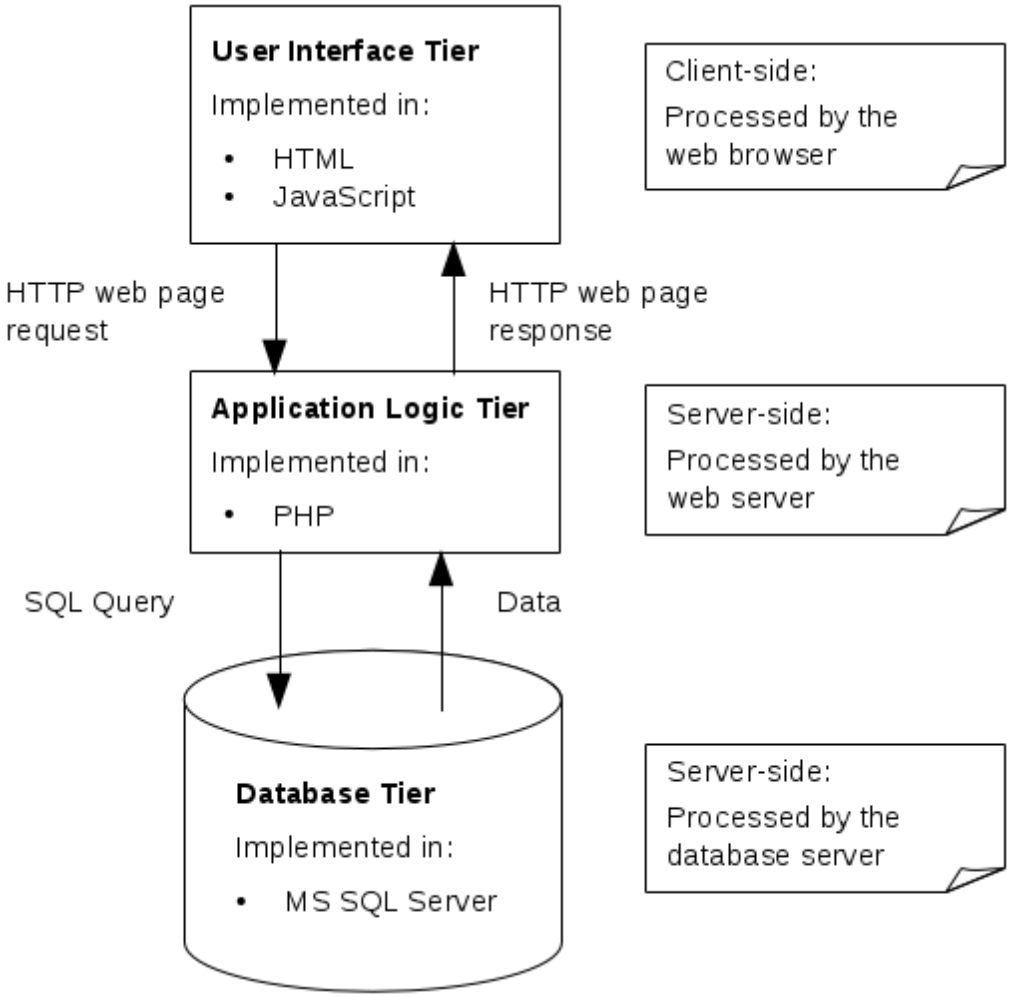


Figure 3.9: The three-tier model of the web application

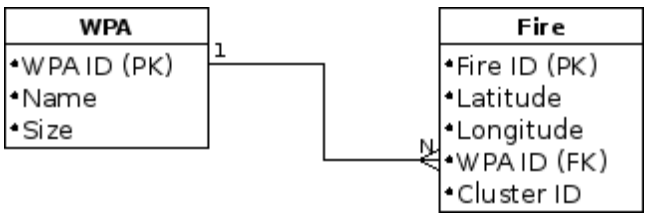


Figure 3.10: The entity relationship model of the database showing the WPA and Fire tables

The application modules for visualizing the fire points and hot spots were designed after the database design since this was a data-driven application. Seven modules were identified for the web application. A component diagram was used to highlight the dependencies between these modules. Figure 3.11 is the component diagram for the web application. It shows the seven application modules and their dependencies.

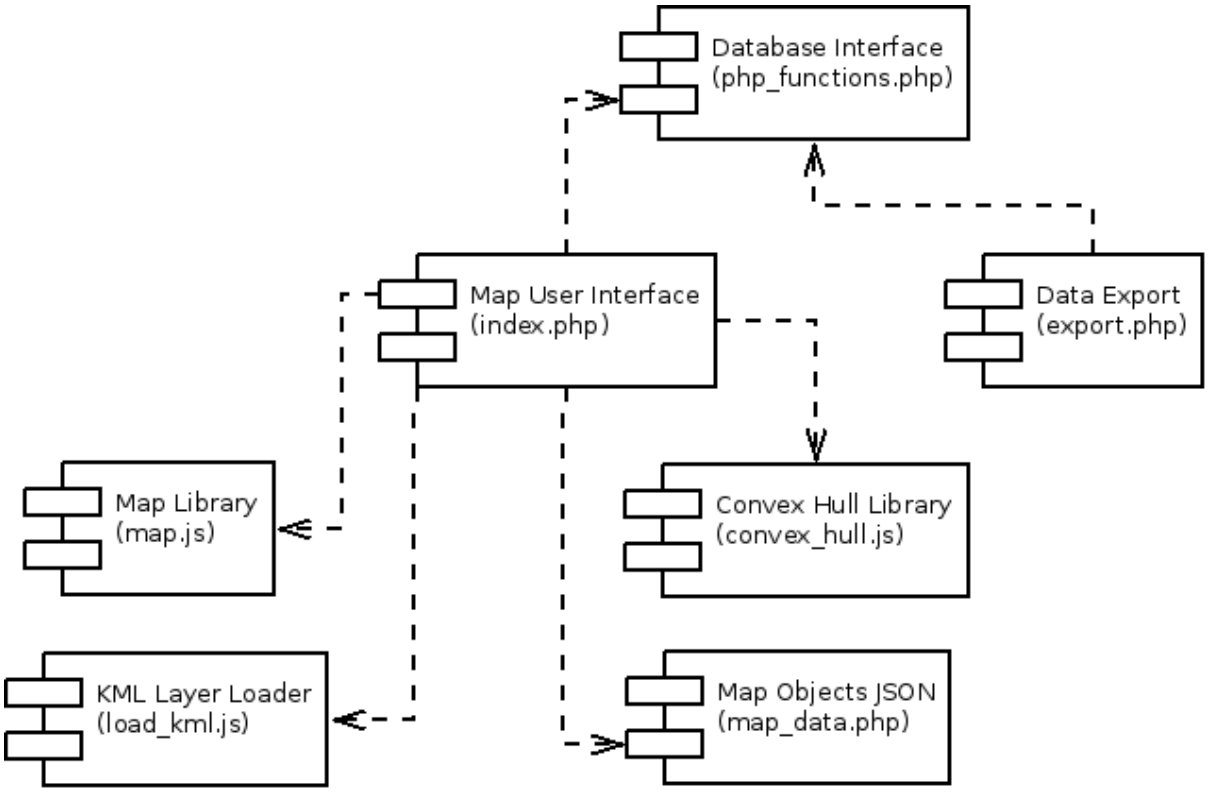


Figure 3.11: Component diagram for the web application

3.2.3 Coding

3.2.3.1 Database Implementation

In this stage of development, the database design was implemented using the Microsoft SQL Server 2008 R2 database that is used by the KWSIDS system. The two tables identified during the design stage were the WPA table and the Fire table. Listing A.14 and A.15 show the SQL scripts that were used to create the two tables. Figure 3.12 shows the SQL Server database diagram based on the entity relationship model in figure 3.10.

After the database tables had been created, the GAWK script shown in listing A.16 was written to extract the data for the Fire database table from the cluster and noise text files produced by ELKI. The following GAWK command was executed to produce a CSV text file named fire.csv using the script. This file had data for the Latitude, Longitude, WPA Name, and Cluster ID fields of the database table. The Fire ID field was generated automatically by SQL Server during the data import process.

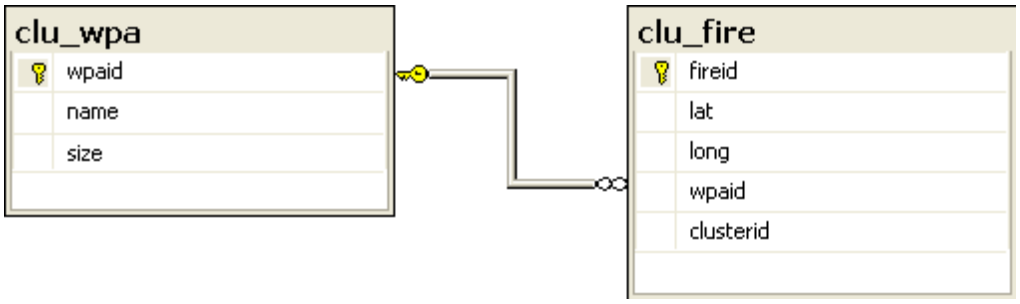


Figure 3.12: The SQL Server database diagram showing the database implementation

```
$ awk -v OFS=, -f fire.awk cluster_*.txt noise.txt > fire.csv
```

The unique set of WPA names in the fire.csv file were extracted with the following GAWK command. The size of each WPA was entered manually from KWS GIS shapefiles. This was the data to be imported into the WPA database table. It was saved in a CSV text file named wpa.csv. The WPA ID field was also generated automatically by SQL Server during the data import process.

```
$ awk -F, '{ a[$3] } END { for (k in a) print k }' fire.csv | sort \
> wpa.csv
```

After this, the GAWK script in listing A.17 was written to replace the WPA names in fire.csv with their respective WPA IDs. The WPA IDs were the serial numbers of the WPA records in wpa.csv. These records had already been sorted alphabetically by WPA name. The following commands were executed to replace the WPA names with the respective WPA IDs.

```
$ awk -F, -v OFS=, -f wpa.awk wpa.csv fire.csv > fire.new.csv
$ mv fire.new.csv fire.csv
```

The data in wpa.csv and fire.csv was then imported into the SQL Server WPA and Fire database tables respectively using the SQL Server Import and Export Wizard. Some of the records in the wpa.csv and fire.csv CSV text files are presented in listings A.18 and A.19. Figures 3.13 and 3.14 show some of the records in the WPA and Fire tables in the KWSIDS SQL Server database.

3.2.3.2 User Interface and Application Logic Implementation

The seven modules implementing the user interface and application logic layers were programmed using the PHP, JavaScript, and HTML languages. These modules were responsible for responding to user queries with the requested data.

wpaid	name	size
1	Aberdare NP	765.7000
2	Amboseli NP	392.0000
3	Arabuko Sokoke NP	6.0000
4	Arawale NR	533.0000
5	Bisanadi NR	606.0000
6	Boni NR	1339.0000
7	Buffalo Springs NR	131.0000
8	Central Island NP	5.0000
9	Chepkitale NR	178.2000
10	Chyulu Hills NP	734.2700

Figure 3.13: Records in the WPA table in the KWSIDS SQL Server database

fireid	lat	long	wpaid	clusterid
1	-2.5680	37.8400	10	0
2	-2.5680	37.8410	10	0
3	-2.5660	37.8370	10	0
4	-2.5720	37.8420	10	0
5	-2.5690	37.8350	10	0
6	-2.5700	37.8350	10	0
7	-2.5620	37.8410	10	0
8	-2.5700	37.8340	10	0
9	-1.8350	41.0690	12	10
10	-1.8350	41.0680	12	10

Figure 3.14: Records in the Fire table in the KWSIDS SQL Server database

The two main features of the web application were the visualization of the fire hot spot clusters on the Google Maps interface and the facility for exporting the MODIS fire data in CSV format. Each feature was accessible through a link in the KWSIDS navigation panel. Both links were provided under a parent link for the application.

Mapping Feature

The mapping feature provided a toolbar, the Google Maps interface, and a sidebar. The toolbar contained a select box allowing users to select a specific WPA or all the WPAs (the default option). The toolbar also provided three controls for showing or hiding the sidebar, resetting the map to its default state, and printing the map. The sidebar contained the map legend, including the fire point symbols and WPA boundaries layer. It also provided fire summary statistics for the selected WPA. Figure 3.15 shows the mapping web page.

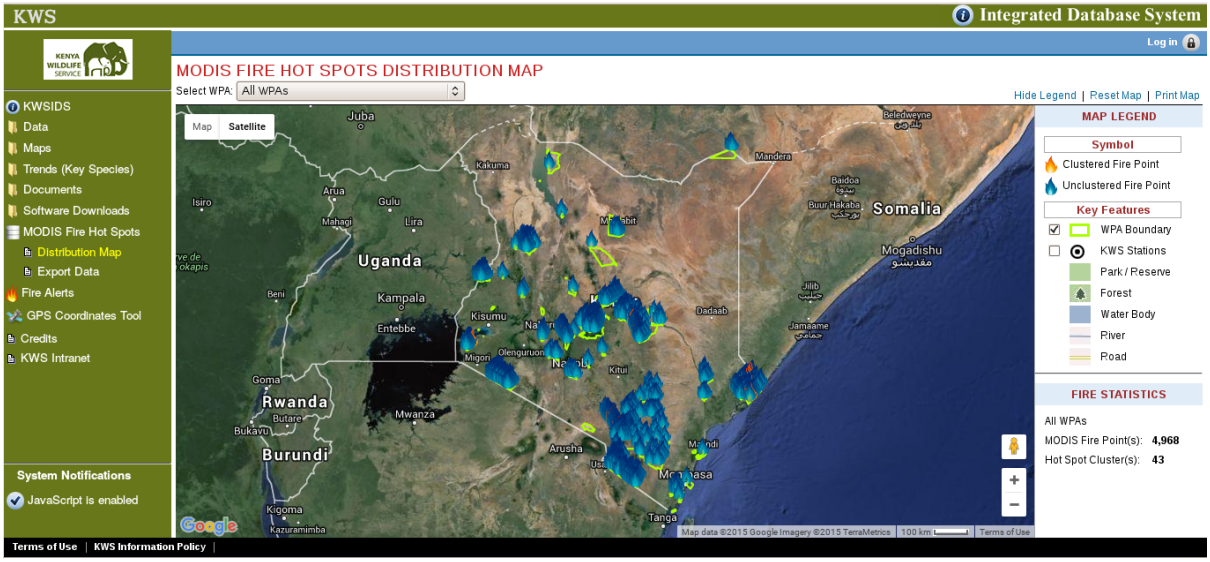


Figure 3.15: Mapping feature showing the Google Maps interface, toolbar, and sidebar

Each fire point in the selected WPA was plotted on the Google Maps interface. Clustered fire points were plotted with a red flame symbol while unclustered (noise) fire points were plotted with a blue flame symbol. In addition, the fire hot spot clusters identified in that WPA were also displayed as convex hull polygons around the fire points constituting that hot spot cluster, as shown in figure 3.16. Clicking on each fire point opened an information window displaying its coordinates, WPA, and cluster ID. Clicking on a fire hot spot cluster also opened an information window displaying its ID, WPA, and the number of fire points contained in the cluster. Figure 3.17 shows this feature.

Upon initial loading, the Google Maps interface displayed the fire points and hot spot clusters in all the WPAs. When the user selected a particular WPA in the toolbar, the application hid the fire points and clusters that did not fall in that WPA in order to allow the user to focus on visualizing the selected WPA. In addition to this, the application also updated the fire summary statistics displayed in the sidebar to correspond to the selected WPA.

The interactive features of Google Maps such as panning, zooming, and switching between different map types were all available to the user. The map loaded with the Satellite map type and a zoom level and map center that displayed the whole of Kenya. In addition, the WPA boundaries layer was also displayed. As a result, the user would initially get an overview of all the fire points against the WPA boundaries.

Data Export Feature

The data export feature enabled users to export the MODIS fire data set from the database in CSV format. The export data set could then be imported into the ArcGIS software, used at KWS, for advanced spatial analysis. As shown in figure 3.19, the export interface provided a form for selecting the WPA and specifying whether noise

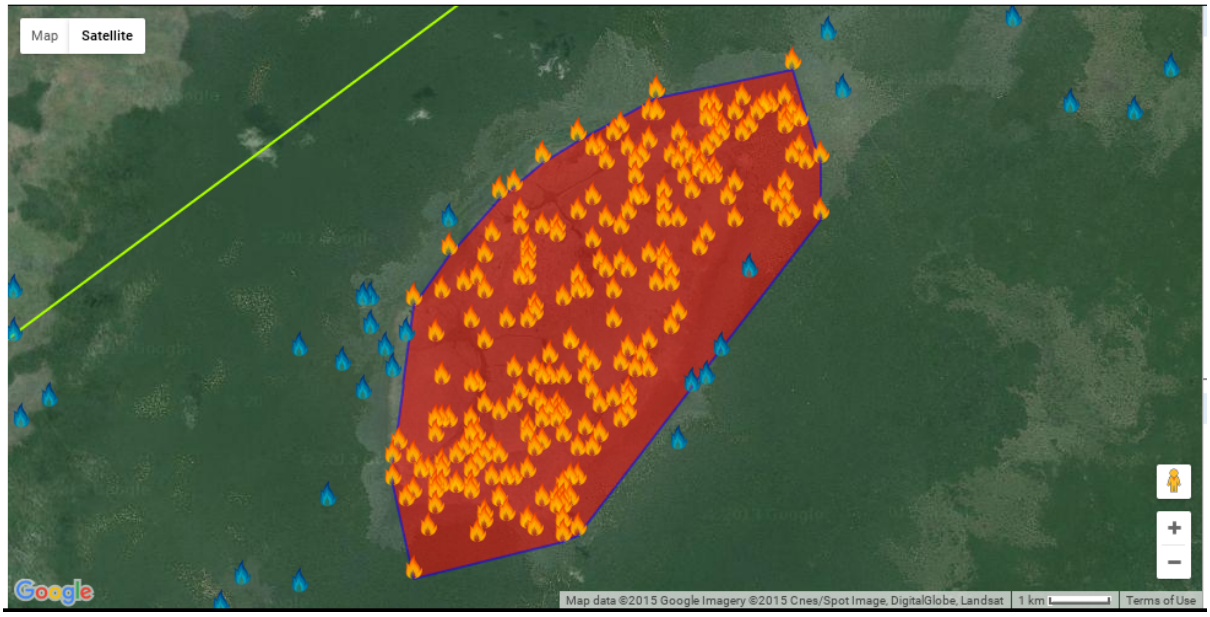


Figure 3.16: The largest fire hot spot cluster in Boni National Reserve

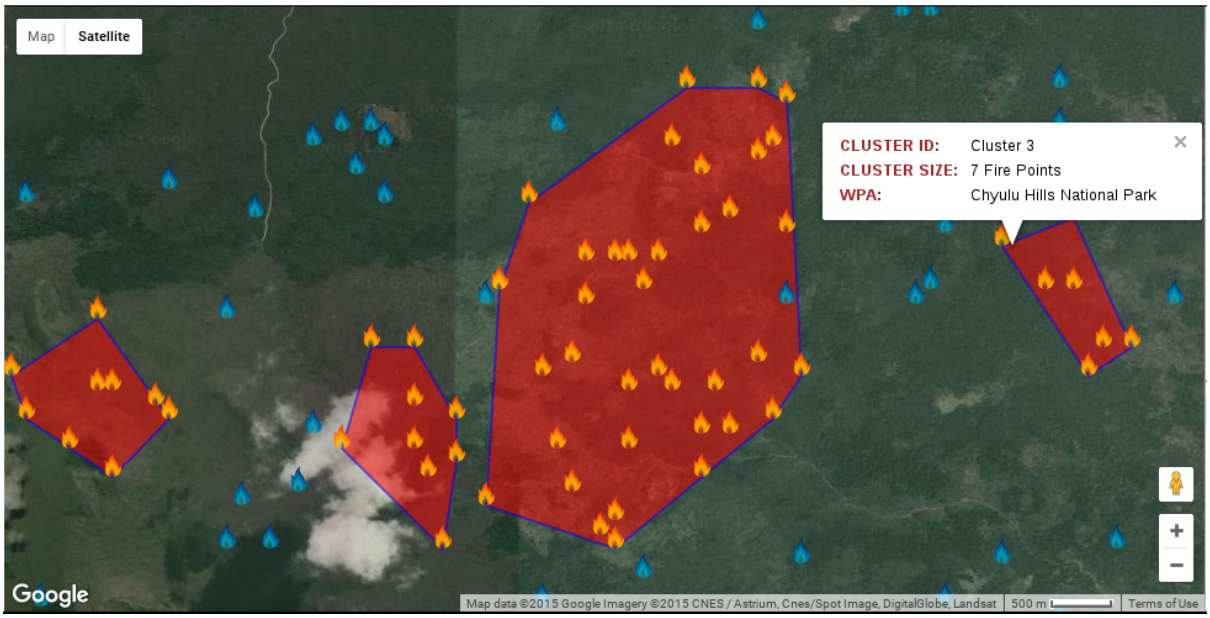


Figure 3.17: Information window of a fire hot spot cluster with 7 fire points in Chyulu Hills National Park

points should be included in the export data set. The application would then generate the export data set and prompt the user to download the CSV file containing the data.

3.2.4 Testing

The software quality attributes and software quality metrics to be used as performance indicators during testing of the web application were identified at this stage. The software

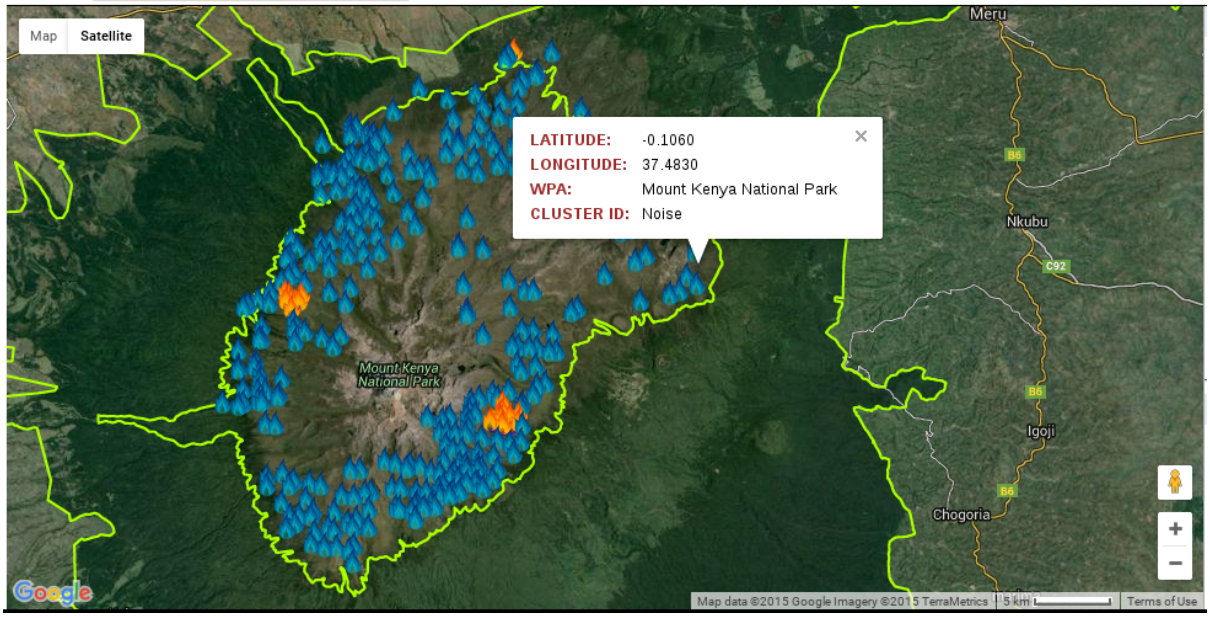


Figure 3.18: Information window of an unclustered (noise) fire point in Mount Kenya National Park

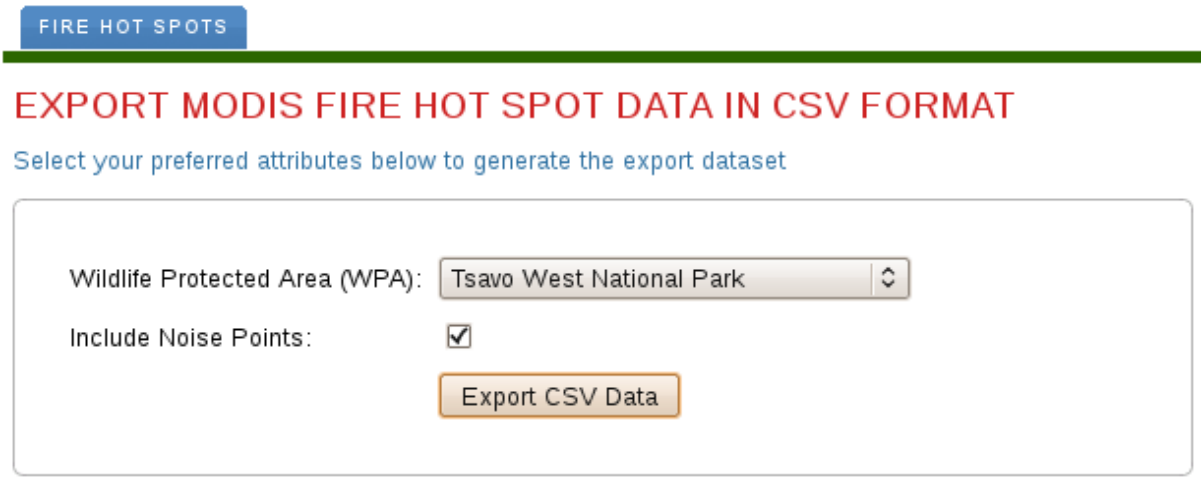


Figure 3.19: Form for exporting the MODIS fire data set

quality attributes were based on the International Organization for Standardization/International Electrotechnical Commission (ISO/IEC) 25010:2011 Software Quality Model (ISO, 2011). The attributes and metrics are presented in table 3.2.

The results of the web application testing were as follows:

- 100% of the user input was validated correctly.
- There were no errors logged by the KWSIDS system during execution of the database queries.

Table 3.2: Software quality attributes and metrics

Requirement	**Software Quality Attribute**	**Software Quality Metric**
Functional	Correctness	% of user input validated correctly.
		No. of errors logged during execution of all database queries
	Completeness	% of specified functional requirements met by the application
	Interoperability	The application exports fire data in the CSV format suitable for processing by the ArcGIS software used at KWS
Non-functional	Efficiency	Average web page load time for each database query
	Usability	The application provides a consistent user interface across all web browsers used by KWS staff
	Reliability	No. of system failures logged by the web and database servers

- 100% of the specified functional requirements were met by the application.
- The application exported fire data in the CSV format suitable for processing by the ArcGIS software used at KWS.
- The average initial web page load time for the mapping query was 4 seconds. After the initial map had loaded, zooming and panning actions took an average of 2 seconds to complete loading the Google Maps tiles. The selection of specific WPAs took less than 1 second to reflect on the map.
- It took less than 1 second to generate the export data set.
- The application provided a consistent user interface across all the web browsers used by KWS staff.
- There were no system failures logged by the Apache web server and SQL Server database server during the testing period.

3.2.5 Deployment

The web application was deployed on the KWS Intranet as a module of the KWSIDS system. It was made available for access by all KWS staff connected to the KWS wide area network.

Chapter 4

Results and Discussion

4.1 Frequency Distribution of MODIS fire points

Table 4.1 shows the frequency distribution table of the MODIS fire points. The 4,968 MODIS fire points occurred in 45 WPAs out of a total of 62. Therefore, fire activity was present in 73% of the WPAs during the 12 years under study (2003-2014). There was no fire activity observed in 17 WPAs (27%) during the same period. The average number of fire points observed per WPA was 110 (4,968 fire points / 45 WPAs) while the standard deviation of the number of fire points per WPA was 171.09.

These statistics indicate a relatively high level of fire activity in the WPAs. The number of fire points observed per WPA had a high standard deviation indicating large differences in the amount of fire activity occurring in the various WPAs. More than half of this fire activity (51.24%) was recorded in only 5 WPAs (Tsavo West NP, Chyulu Hills NP, Boni NR, Masai Mara NR, and Dodori NR).

Table 4.1: Frequency distribution table of the MODIS fire points

No.	**WPA Name**	**No. of Fire Points**	**Relative Frequency**
1.	Tsavo West NP[2]	693	13.95%
2.	Chyulu Hills NP	581	11.69%
3.	Boni NR[3]	443	8.92%
4.	Masai Mara NR	415	8.35%
5.	Dodori NR	414	8.33%
6.	Mount Kenya NP	340	6.84%
7.	South Turkana NR	328	6.60%
8.	Aberdare NP	264	5.31%
9.	Mount Kenya NR	229	4.61%
10.	North Kitui NR	204	4.11%
11.	Chepkitale NR	198	3.99%
12.	Ruma NP	166	3.34%

[2] NP - National Park
[3] NR - National Reserve

No.	WPA Name	Absolute Frequency	Relative Frequency
13.	Tsavo East NP	119	2.40%
14.	Meru NP	118	2.38%
15.	Mount Elgon NP	118	2.38%
16.	Kiunga Marine NR	75	1.51%
17.	Nyambene NR	53	1.07%
18.	Bisanadi NR	34	0.68%
19.	South Kitui NR	27	0.54%
20.	Shimba Hills NR	21	0.42%
21.	Kora NP	18	0.36%
22.	Lake Nakuru NP	18	0.36%
23.	Nairobi NP	17	0.34%
24.	Marsabit NP	13	0.26%
25.	Mount Longonot NP	11	0.22%
26.	Arawale NR	11	0.22%
27.	Tana River Primate NR	6	0.12%
28.	Shaba NR	6	0.12%
29.	Hell's Gate NP	4	0.08%
30.	Malka Mari NP	4	0.08%
31.	Ngai Ndethya NR	3	0.06%
32.	Kamnarok NR	2	0.04%
33.	Rahole NR	2	0.04%
34.	Mwea NR	2	0.04%
35.	Saiwa NP	1	0.02%
36.	Losai NR	1	0.02%
37.	Diani Chale Marine NR	1	0.02%
38.	Sibiloi NP	1	0.02%
39.	Nasolot NR	1	0.02%
40.	Marsabit NR	1	0.02%

No.	WPA Name	Absolute Frequency	Relative Frequency
41.	Malindi Marine NR	1	0.02%
42.	Laikipia NR	1	0.02%
43.	South Island NP	1	0.02%
44.	Ol Donyo Sabuk NP	1	0.02%
45.	Mombasa Marine NR	1	0.02%
	Total	**4,968**	**99.98%**[1]

Figure 4.1 shows the frequency distribution bar graph for the absolute frequencies presented in table 4.1.

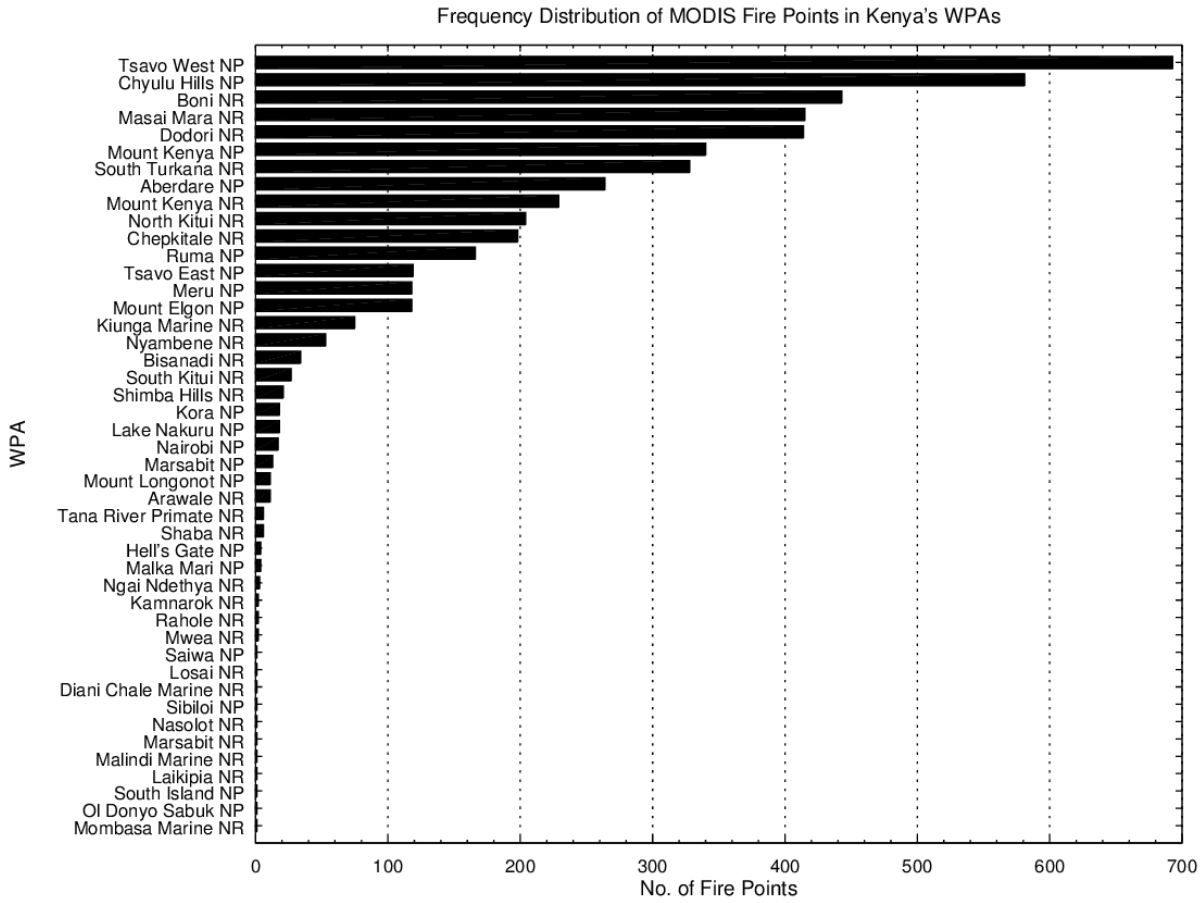


Figure 4.1: Frequency distribution bar graph

[1] The Relative Frequency total is slightly less than the expected 100% due to rounding errors introduced by the MySQL RDBMS.

Figure 4.2 shows the scatterplot diagram of the preprocessed MODIS fire points. Each '+' symbol represents a fire point identified by its latitude-longitude coordinate pair. This diagram shows the overall distribution of fire points in all WPAs at a low spatial resolution. From this, we can see that it is impossible to tell where the fire hot spot regions *within* Kenya's WPAs are. This necessitates the density-based cluster analysis that was performed in this study.

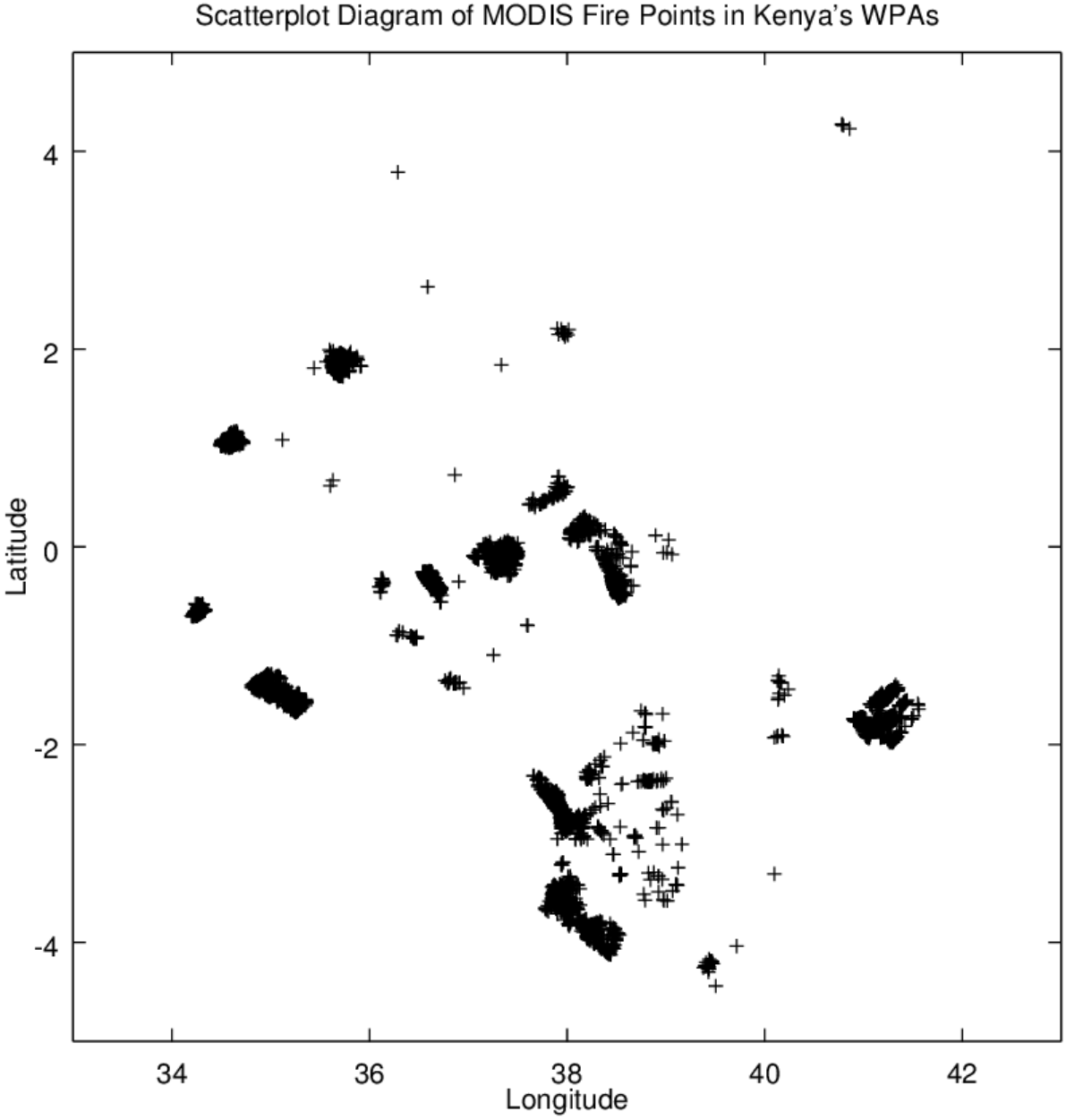


Figure 4.2: Scatterplot diagram of the MODIS fire points

4.2 DBSCAN Parameter Estimates from the Sorted *k*-dist Graph

The sorted k-dist graph of the MODIS data set is shown in figure 4.3. The x-axis shows the MODIS fire points running from 1 to 4,968. The y-axis shows the distance to the 4-th nearest neighbor for each fire point, in meters. The distances ascend from left to right. In this graph, the estimated threshold point occurred at about fire point number 4,906 where the 4-th nearest neighbor distance was 13,564 m. The initial estimates of the DBSCAN parameters from the sorted k-dist graph were: Eps = 13,564 m (13.564 km) and $MinPts$ = 5.

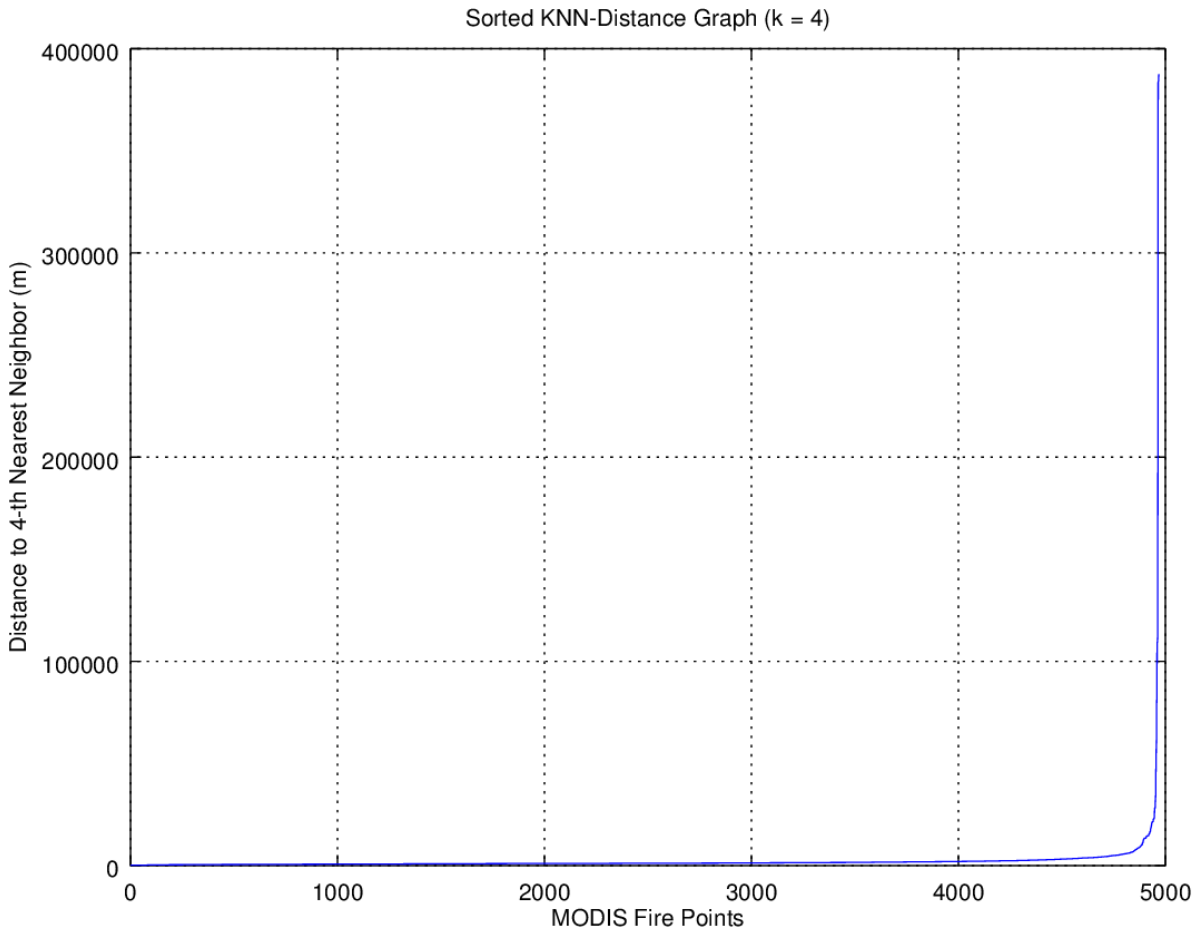


Figure 4.3: The sorted k-dist graph for $k = 4$

The plot shown in figure 4.3 was zoomed in four times (x4) and panned to the top-left to produce figure 4.4 which shows the estimated threshold point of the k-dist graph. All fire points with a higher k-dist value (right of the threshold point) in the graph were considered to be noise, while all other points (left of the threshold point) were assigned to a cluster. At this threshold point, 98.75% (4,906 / 4,968 fire points) of the data set was assigned to clusters while 1.25% (62 / 4,968 fire points) was identified as noise.

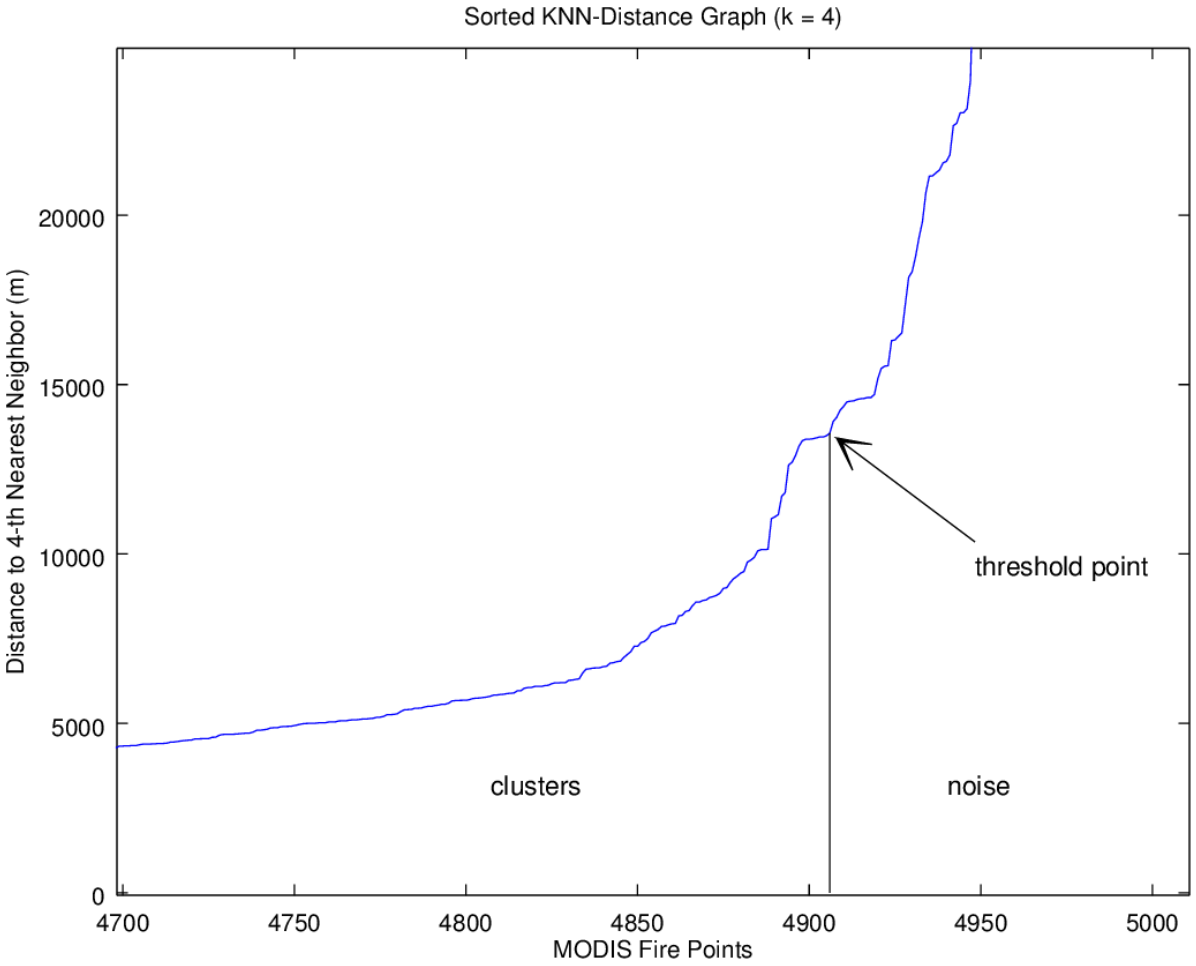


Figure 4.4: The threshold point for the sorted k-dist graph

4.3 Initial Clustering Result for the Estimated DBSCAN Parameters

The ELKI DBSCAN execution with the R*-Tree index structure had the lowest runtime complexity. Its mean runtime on the MODIS active fire data set of 4,968 fire points was 3.0 seconds. The M-Tree had a mean runtime of 26.5 seconds while the KD-Tree's was 28.3 seconds. Executing DBSCAN with no index structure achieved a mean runtime of 28.9 seconds. Table 4.2 summarizes these results.

Execution without an accelerating index structure was slower by a factor of 9.63 (28.9 / 3.0 seconds). These results clearly show that the R*-Tree index structure provides the best performance when ELKI's DBSCAN is executed with a geographical distance function. When developing an application that visualizes such clustering results, it is possible to perform a user-directed cluster analysis "on the fly" while retaining an acceptable response time. This study performed the clustering "off-line" before the application was developed because it was not known beforehand how long the clustering would take.

Table 4.2: Performance comparison of ELKI index structures

Index Structure	**Run 1 (sec)**	**Run 2 (sec)**	**Run 3 (sec)**	**Mean Runtime (sec)**
R*-Tree	3.1	3.0	3.0	3.0
M-Tree	26.3	26.5	26.7	26.5
KD-Tree	28.4	28.3	28.2	28.3
No index structure	28.5	29.9	28.3	28.9

Figure 4.5 shows the scatterplot diagram of the fire hot spot clusters, produced by ELKI for the initial DBSCAN parameters. The axes represent degrees of longitude (x-axis) and latitude (y-axis). A total of 29 hot spot clusters were identified. In the diagram, the noise points not assigned to any cluster are indicated by the blue '+' symbols.

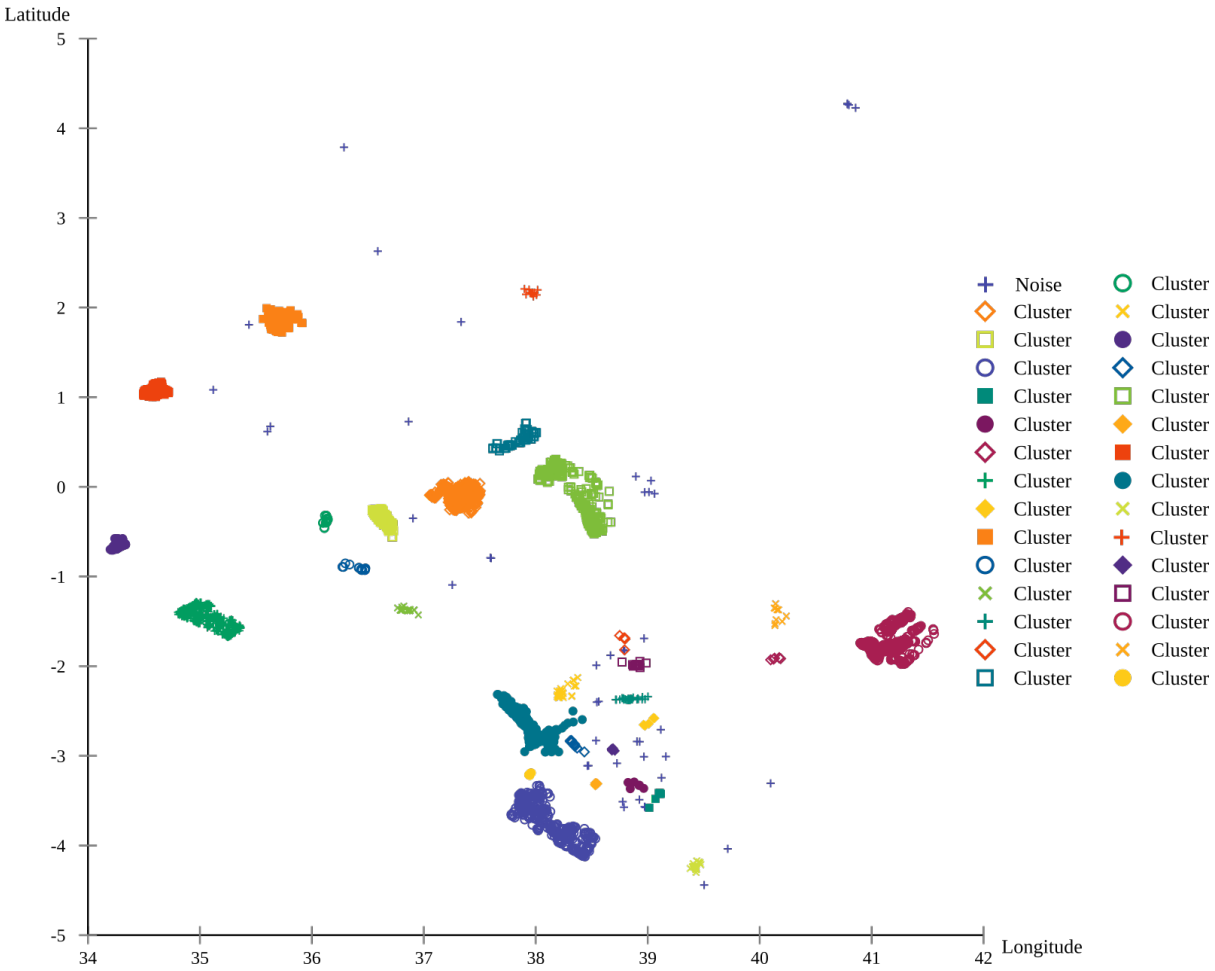


Figure 4.5: Scatterplot diagram for initial DBSCAN parameters

The sorted k-dist graph heuristic proposed by Ester et al. (1996) and also used by Usman, Sitanggang and Syaufina (2015) to estimate the DBSCAN parameters was not suitable for the MODIS active fire data set used in this study. The value of *Eps* (13.564 km) estimated from the sorted k-dist graph was too large since it clustered 98.75% of the data set. Smaller, significant fire hot spot clusters merged resulting in larger, less significant clusters that included a lot of noise. Most of these clusters covered entire WPAs which rendered them ineffective in meeting the primary research objective of this study. This was because the fire points occurring in remote WPAs with fewer than 4 (k) neighboring fire points had very large 4-th nearest neighbor distances. The highest such distance was 387.47 km. Examples of such WPAs were Malka Mari NP (4 fire points), Sibiloi NP (1 fire point), South Island NP (1 fire point), Losai NR (1 fire point), and Laikipia NR (1 fire point). As a result, the sorted k-dist graph produced an unsuitable threshold point of a relatively high k-distance value.

4.4 Results for Trial Run DBSCAN Parameters

Due to the *exploratory* nature of the data clustering task in this study, it was necessary to experiment with different values of the DBSCAN parameters in order to find the most suitable pair for the MODIS fire data set. The 25 pairs of parameter values for the DBSCAN trial runs were selected to include a wide range of geographical scopes (0.8 km^2 to 2.5 km^2) and minimum number of fire points within these scopes (5 to 12).

The clustering result of each pair of DBSCAN parameter values used in the trial runs was evaluated by examining its ELKI scatterplot diagram and the number of clusters produced. Figures 4.6 to 4.9 show the scatterplot diagrams of the fire hot spot clusters produced by ELKI for the various trial run DBSCAN parameters.

Table 4.3 shows the results for the trial runs. The parameters $MinPts = 7$ and $Eps = 700$ m in trial run no. 8 of the table were selected as the most suitable for the MODIS fire data in Kenya's WPAs. The most suitable parameter values were selected based on their ability to produce significant clusters in the presence of noise. In addition, the clusters were small enough in size to be identified within Kenya's WPAs.

The results of the trial runs showed that a *MinPts* value of 7 was large enough to yield significant clusters in the presence of noise within the MODIS fire data set. The smaller value of 5 produced many more clusters which included more noise points. The larger values of 8, 10, and 12 produced fewer clusters because a larger part of the data set was considered to be noise.

The *Eps* value of 700 m with an *Eps*-neighborhood value of 1.5 km^2 provided a reasonable geographical scope within which at least 7 fire points occurred, for the smallest possible fire hot spot cluster. The smaller values of 500 m and 600 m yielded fewer clusters due to the smaller geographical scopes (0.8 km^2 and 1.1 km^2 respectively) imposed on the data set. On the other hand, the larger values of 800 m and 900 m yielded more clusters which included more noise points due to the larger, more permissive geographical scopes (2.0 km^2 and 2.5 km^2 respectively).

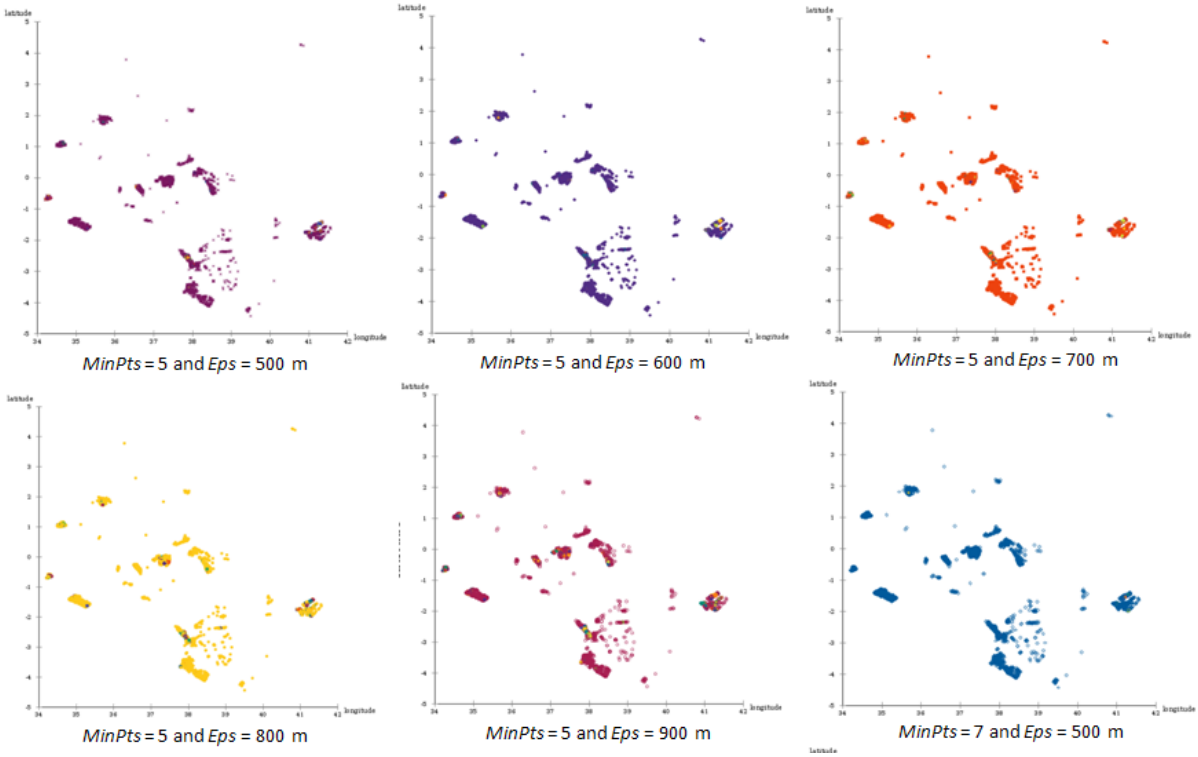


Figure 4.6: Scatterplot diagrams for trial run parameter values

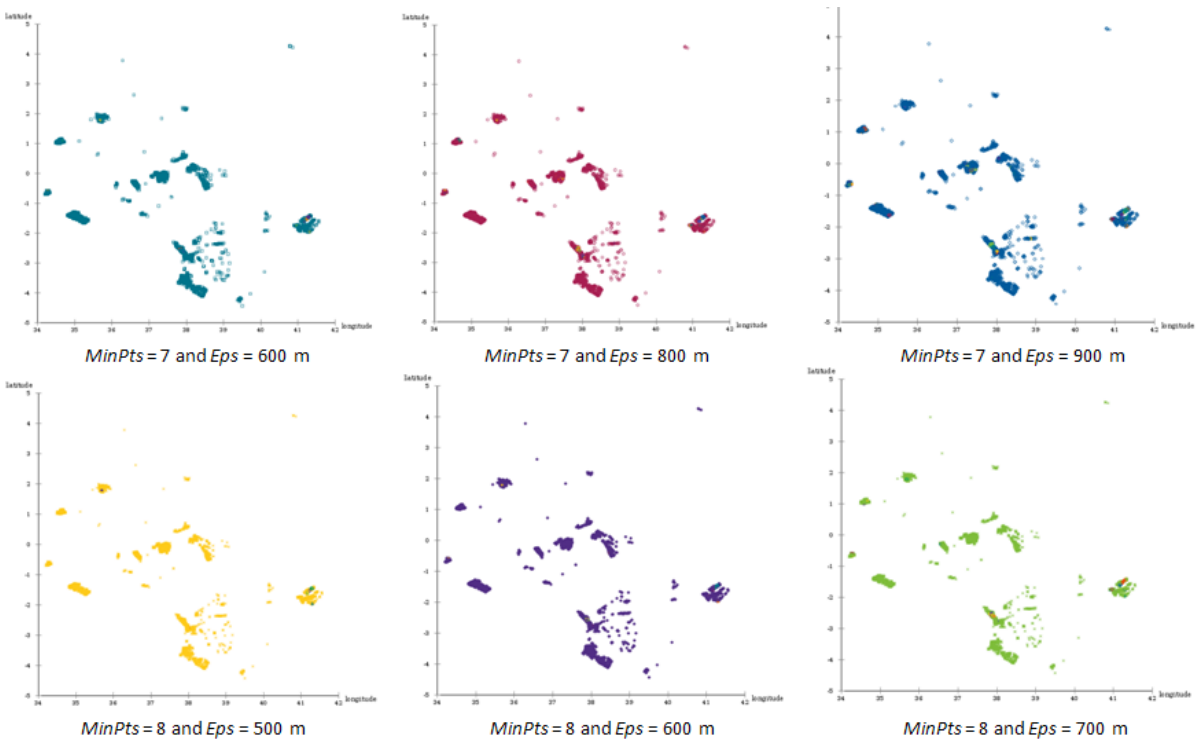


Figure 4.7: Scatterplot diagrams for trial run parameter values

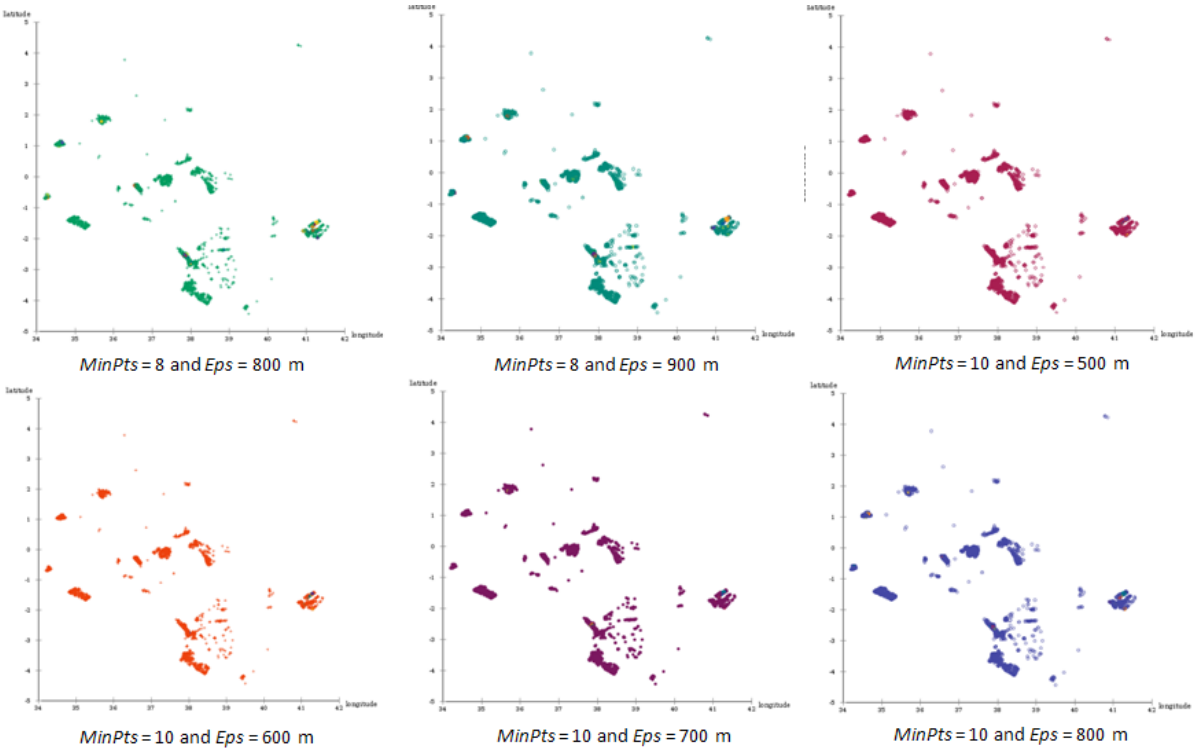


Figure 4.8: Scatterplot diagrams for trial run parameter values

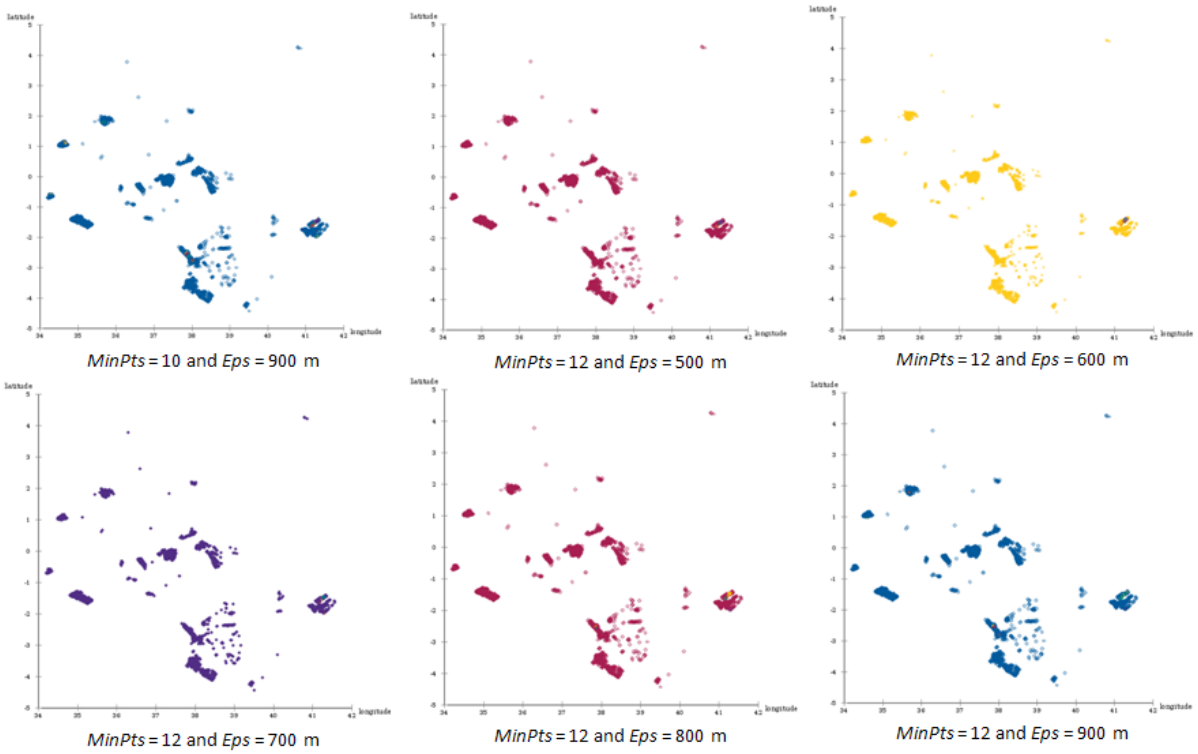


Figure 4.9: Scatterplot diagrams for trial run parameter values

Table 4.3: DBSCAN trial run results for different parameter values

Run No.	***MinPts***	***Eps*** **(m)**	***Eps*-neighborhood** **(km^2)**	**No. of Hot Spot Clusters**
Initial	5	13,564	578	29
1.	5	500	0.8	65
2.	5	600	1.1	91
3.	5	700	1.5	127
4.	5	800	2.0	139
5.	5	900	2.5	143
6.	7	500	0.8	15
7.	7	600	1.1	33
8.	**7**	**700**	**1.5**	**43**
9.	7	800	2.0	59
10.	7	900	2.5	74
11.	8	500	0.8	10
12.	8	600	1.1	15
13.	8	700	1.5	28
14.	8	800	2.0	39
15.	8	900	2.5	52
16.	10	500	0.8	8
17.	10	600	1.1	8
18.	10	700	1.5	13
19.	10	800	2.0	23
20.	10	900	2.5	30
21.	12	500	0.8	6
22.	12	600	1.1	4
23.	12	700	1.5	7
24.	12	800	2.0	10
25.	12	900	2.5	15

4.5 Final Clustering Result for the Most Suitable DBSCAN Parameters

Figure 4.10 shows the scatterplot diagram of the fire hot spot clusters produced by ELKI for the final DBSCAN parameters that were selected as the most suitable for the MODIS fire data in Kenya's WPAs. The final parameters yielded 43 fire hot spot clusters. DBSCAN assigned 15.40% (765 / 4,968 fire points) of the data set to clusters while 84.60% (4,203 / 4,968 fire points) was identified as noise. In the diagram, the noise points not assigned to any cluster are indicated by the green '×' symbols.

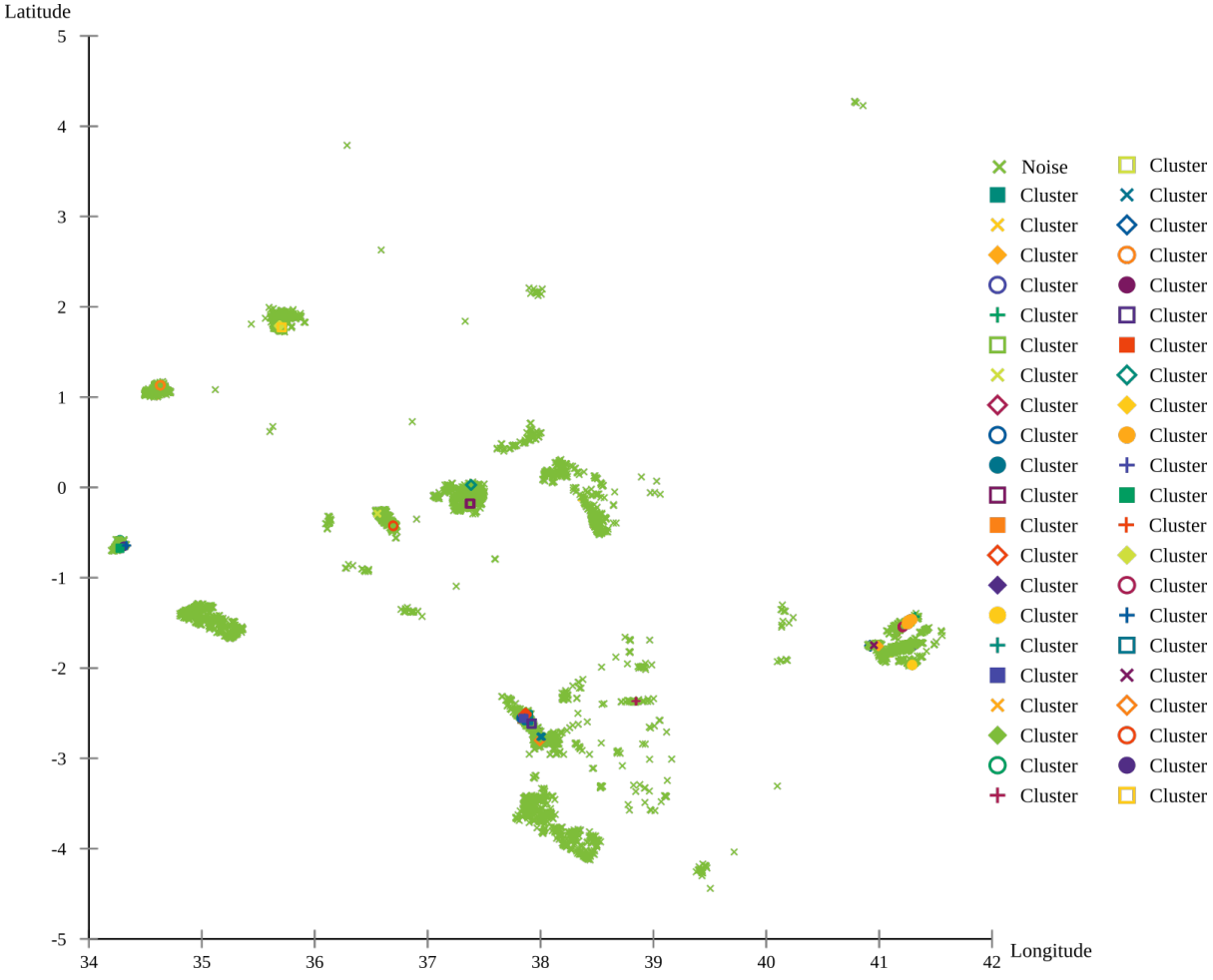


Figure 4.10: Scatterplot diagram for DBSCAN with $MinPts = 7$ and $Eps = 700$ m

4.6 Frequency Distribution and Sizes of Identified Fire Hot Spot Clusters

Table 4.4 shows the frequency distribution of the fire hot spot clusters identified in Kenya's WPAs. The hot spot clusters were identified in 14 WPAs (31%) out of the 45 WPAs which had recorded fire activity. The results indicated that the four WPAs most vulnerable to wildfires in Kenya were: Chyulu Hills NP, Dodori NR, Boni NR, and Ruma NP. All of these WPAs had a high number of MODIS fire points occurring in them, in addition to having several large fire hot spot clusters. Together, they accounted for 60.46% of the fire hot spot clusters.

Table 4.4: Frequency distribution table of the fire hot spot clusters

No.	**WPA Name**	**Absolute Frequency**	**Relative Frequency**
1.	Chyulu Hills NP[1]	12	27.91%
2.	Dodori NR[2]	6	13.95%
3.	Boni NR	4	9.30%
4.	Ruma NP	4	9.30%
5.	Tsavo West NP	3	6.98%
6.	Mount Kenya NP	3	6.98%
7.	Aberdare NP	2	4.65%
8.	Mount Elgon NP	2	4.65%
9.	North Kitui NR	2	4.65%
10.	South Turkana NR	2	4.65%
11.	Tsavo East NP	1	2.33%
12.	Mount Kenya NR	1	2.33%
13.	Chepkitale NR	1	2.33%
14.	Kiunga Marine NR	1	2.33%
	Total	**43**[3]	**102.34%**[4]

[1] NP - National Park
[2] NR - National Reserve
[3] The actual total from the table is 44 but one hot spot cluster is shared between Mount Kenya NP and Mount Kenya NR and should not be counted twice.
[4] The Relative Frequency total is slightly higher than the expected 100% due to rounding errors introduced by the SQL Server query used to compute the frequencies.

Figure 4.11 shows the frequency distribution bar graph for the absolute frequencies presented in table 4.4.

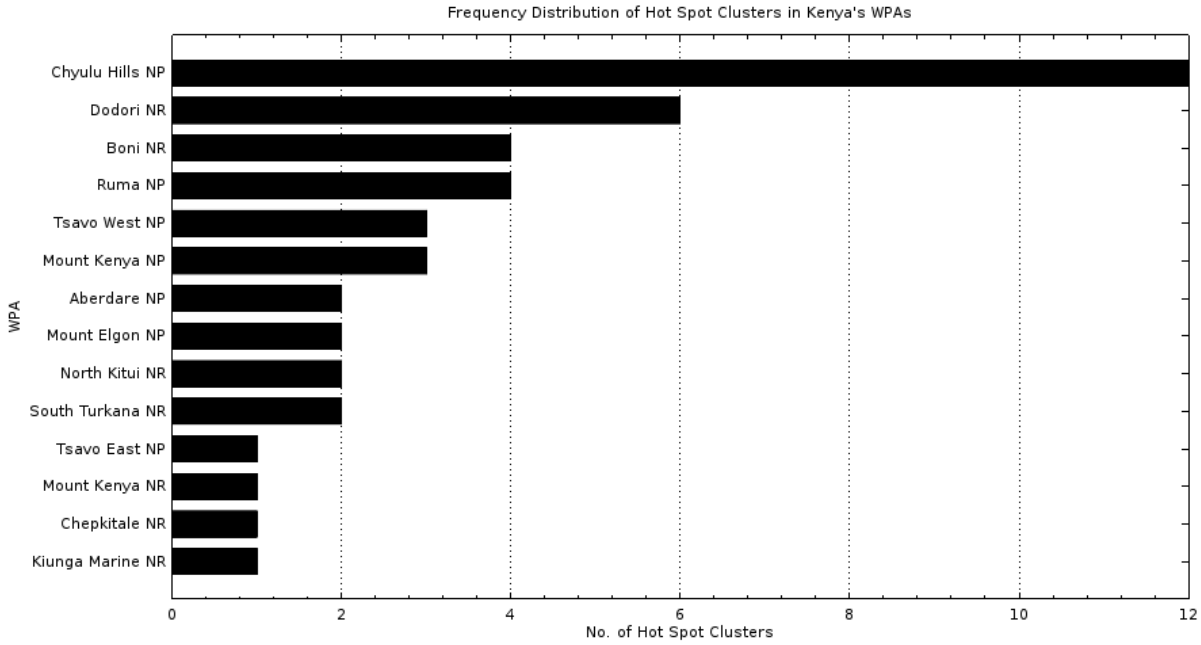


Figure 4.11: Frequency distribution bar graph of fire hot spots clusters

Table 4.5 shows the number of fire points in each of the 43 fire hot spot clusters. The average number of fire points per hot spot cluster was 18 (765 clustered fire points in 43 clusters). The identified hot spot clusters were large in size since the average number of fire points per cluster was far higher than the minimum of 7 defined by the DBSCAN *MinPts* parameter.

Table 4.5: Number of fire points in the fire hot spot clusters

No.	**Cluster ID**	**WPA Name**	**No. of Fire Points**
1.	14	Boni NR	237
2.	26	Boni NR	38
3.	6	Chyulu Hills NP	38
4.	10	Dodori NR	37
5.	41	Chyulu Hills NP	31
6.	25	Boni NR	21
7.	13	Chyulu Hills NP	21
8.	24	Kiunga Marine NR	18

No.	Cluster ID	WPA Name	No. of Fire Points
9.	11	Dodori NR	15
10.	32	Chyulu Hills NP	14
11.	1	South Turkana NR	14
12.	5	Tsavo West NP	13
13.	7	Boni NR	13
14.	18	Mount Elgon NP	13
15.	38	Mount Kenya NP	11
16.	29	Ruma NP	11
17.	16	Chyulu Hills NP	11
18.	35	Tsavo West NP	11
19.	17	Chyulu Hills NP	10
20.	34	Chepkitale NR	10
21.	21	North Kitui NR	10
22.	22	Dodori NR	10
23.	19	Chyulu Hills NP	9
24.	4	Chyulu Hills NP	8
25.	28	Chyulu Hills NP	8
26.	0	Chyulu Hills NP	8
27.	15	Aberdare NP	8
28.	31	Dodori NR	8
29.	33	Dodori NR	8
30.	42	Dodori NR	8
31.	37	Mount Elgon NP	8
32.	20	Ruma NP	8
33.	39	Tsavo West NP	7
34.	12	South Turkana NR	7
35.	40	Tsavo East NP	7
36.	27	Mount Kenya NP	7

No.	Cluster ID	WPA Name	No. of Fire Points
37.	36	North Kitui NR	7
38.	30	Ruma NP	7
39.	9	Ruma NP	7
40.	2	Aberdare NP	7
41.	3	Chyulu Hills NP	7
42.	23	Chyulu Hills NP	7
43.	8	Mount Kenya NP & Mount Kenya NR	7
	Total		**765**

The clustering results support prior KWS knowledge (Kamau, 2013; KWS, 2012d) that Chyulu Hills NP is highly vulnerable to wildfires. The data indicated fire presence in about 80% of the park. It also had the highest number of hot spot clusters (12), indicating a high distribution density of fires.

The results also confirmed that KBDCA is highly prone to wildfires. The largest fire hot spot cluster consisting of 237 fire points occurred in Boni NR. In total, KBDCA had 11 fire hot spot clusters, almost as high as the number in Chyulu Hills NP.

About 70% of Ruma NP had fire presence with a high distribution density of fire points. 4 fire hot spot clusters were identified in the park. The 166 fire points occuring in Ruma NP resulted in an average of 1.38 fire points per km^2. It was one of only two WPAs (the other being Chepkitale NR) with an average of more than 1 fire point per km^2.

The mountain ecosystems had a high number of fire points but with a relatively low distribution density. Out of the 569 fire points occurring in Mt. Kenya NP and Mt. Kenya NR, only 3 hot spot clusters were identified by DBSCAN. Most of the fire activity occurred in Mt. Kenya NP and the Northern part of Mt. Kenya NR. Aberdare NP had 264 fire points with only 2 hot spot clusters identified in it.

In TCA, the results showed that Tsavo West NP had much higher fire activity than Tsavo East NP, contrary to KWS knowledge. This may be because Tsavo East NP has more sparse vegetation cover that provides less fuel for wildfires. Both parks had only 4 hot spot regions, indicating a low distribution density of fires. The single hot spot in Tsavo East NP occurred on the banks of the Tiva river where there is acacia and palm forest vegetation. The web application's Google Maps interface clearly showed the exact locations of the wildfires that originate along the Nairobi-Mombasa highway and railway line in Tsavo West NP and Tsavo East NP. There were 26 such fire points occurring along the stretch of highway between Mtito Andei and Maungu towns.

4.7 Evaluation of the Web Application

The use of the Google Maps API in the web application developed in this study provided a very interactive means of visualizing the MODIS fire points and hot spot clusters. In particular, the use of a convex hull polygon with a high-contrast fill color and border clearly delineated the extent of the fire hot spot regions. This visualization method has also been used by the GeoClustering application (Wang, Wang & Liang, 2011). Google Maps also provided satellite imagery as a background on which fire points and clusters were overlaid. The vegetation context is helpful in enabling users to understand the probable cause and/or impact of the wildfires.

The web application's feature for viewing data for a specific WPA at a time was very useful for users who were mostly interested in the fire activity within their WPA. This usefulness was coupled with the ability to view the fire points and fire hot spots regions at different spatial resolutions.

While the Google Maps API is rich and highly interactive, it does require considerable network bandwidth to load the map tiles responsively. The web page load time for the web application is therefore determined by the user's Internet connection speed. It also has the disadvantage of not being functional when a user is not connected to the Internet. This may be an impediment for some of the KWS field stations where Internet connectivity is poor.

Chapter 5

Conclusion

5.1 Achievements

The primary research objective of this study was to identify the regions that are fire hot spots in Kenya's WPAs by performing a density-based cluster analysis on the MODIS active fire data set, for a 12 year period covering the years 2003 to 2014. The secondary research objective was to develop a web application that provides an interactive visualization of the fire hot spots identified by the cluster analysis. Both of the research objectives set out for this study were met satisfactorily.

The DBSCAN implementation in the ELKI software framework was effective in clustering the MODIS active fire data set. The final DBSCAN parameters selected as being the most suitable identified 43 fire hot spot regions within Kenya's WPAs. These are the areas that incur a high degree of biodiversity loss and wildlife habitat degradation resulting from wildfire damage.

The findings of this study indicate that density-based cluster analysis is a suitable clustering method for identifying hot spots in geospatial data sets. The cluster analysis performed in this study was a stand-alone exploratory task whose goal was to explore and reveal the hidden patterns in the MODIS active fire data set. The results strongly support observations in the literature which state that density-based clustering is well-suited for discovering clusters of arbitrary shape. This property in addition to DBSCAN's built-in notion of noise and the absence of the need to specify the number of clusters beforehand worked favorably in this study.

The arbitrary shapes of the identified fire hot spot clusters and the large percentage of low-density noise points detected in the MODIS fire data set indicate that density-based cluster analysis was the most suitable clustering method for this study. A clustering algorithm such as the widely used K-Means would not have been suitable for this study since there was no appropriate heuristic available for determining the number of fire hot spot clusters beforehand. In addition, it would have been limited to producing spherical clusters.

This study also showed that the sorted k-dist graph heuristic used to estimate parameters for the DBSCAN algorithm does not necessarily produce suitable parameter values. Its performance is influenced by the characteristics of a data set. Due to the exploratory nature of the clustering task, there was a need to run the DBSCAN algorithm with several different sets of parameter values and evaluate the result of each run, to identify the pair that yielded a suitable clustering result for a given data set.

The main weakness of the DBSCAN algorithm (which is addressed by OPTICS) is that it performs poorly when there are large differences in local densities for different regions

of the data space. This results from its use of one pair of global parameters over the entire data set. The selection of the most suitable pair of DBSCAN parameter values after experimenting with a large set of possibilities gave a meaningful clustering result for this study despite this weakness of the algorithm.

The data analysis performed during the study revealed that there was a relatively high level of fire activity in the WPAs between 2003 and 2014. However, 27% of the WPAs were safe from wildfires since no fire activity was observed in them over the 12 years. The data analysis also showed that the four most vulnerable WPAs to fire activity in Kenya were: Chyulu Hills NP, Dodori NR, Boni NR, and Ruma NP. The main causes of wildfires in these WPAs are the use of inappropriate honey gathering methods and burning to improve the quality of pasture for grazing.

With regards to the web application development, the use of the Google Maps API provided a rich and interactive means of visualizing the MODIS fire points and hot spot clusters. The use of the web platform to disseminate the information on the spatial distribution pattern of fire incidents in Kenya's WPAs enabled wide access for KWS WPA managers and research scientists.

5.2 Limitations

The MODIS active fire data set represents a rich repository from which information on fire incidents can be extracted. However, high quality fire observations from both the Terra and Aqua satellites are available from July 2002. Earlier time periods are not represented. In addition, in some cases, the MODIS instrument has limitations on the fire observations it can make. For instance, a fire detection may be missed due to cloud cover or forest canopy. MODIS may also fail to detect very small fires below about 50 m^2 under poor observing conditions since its spatial resolution is relatively low at 1 km^2 per pixel. Further, the accuracy of the fire point coordinates in the data set is impacted by the MODIS spatial resolution and the precision of 3-decimal places.

Since the MODIS fire data set is captured by remote sensing satellites in Earth orbit, the fire observations made by the MODIS instrument do not provide additional information such as the cause of the detected fires. A user can therefore not tell which of the observed fire points resulted from prescribed burning vis-a-vis wildfires in the WPAs. Such cases would require a KWS WPA manager or research scientist to manually label the fire activity based on its spatial and temporal context of occurrence.

A limitation faced in the use of the DBSCAN algorithm for this study pertains to the absence of suitable external clustering validation measures. The MODIS active fire data set was unlabeled since the fire hot spot regions in the WPAs were not known beforehand. Most of the existing internal clustering validation measures such as the Davies-Bouldin index are designed for centroid-based clustering algorithms.

5.3 Recommendations

Research

Density-based cluster analysis should be given priority over other clustering methods for data mining tasks involving the identification of hot spots in geospatial data sets. When using the DBSCAN algorithm for a density-based clustering task, the use of the sorted k-dist graph heuristic for estimating parameters should be accompanied by experimentation with a wide range of parameters before deciding on the most suitable values.

Practice

The 43 identified fire hot spot regions indicate the areas within the WPAs where fire monitoring efforts by KWS need to be focused during the fire season. Ground and aerial patrols of these regions should be intensified in an effort to reduce future incidents of wildfires. In addition, KWS WPA managers should highly consider the fire hot spot regions when deciding on the locations for constructing new fire watchtowers and firebreaks in the WPAs.

Chyulu Hills NP, Dodori NR, Boni NR, and Ruma NP should be given the highest priority by KWS with regards to the allocation of fire management resources. Within these WPAs, the use of inappropriate honey gathering methods and burning to improve the quality of pasture for grazing should be monitored closely since they are the principal causes of wildfires in the vulnerable WPAs. On the other hand, the 17 WPAs that did not record any fire activity over the 12 years under study require minimal fire management effort. These are: Amboseli NP, Arabuko Sokoke NP, Buffalo Springs NR, Central Island NP, Kakamega NR, Kisite Marine NP, Lake Bogoria NR, Lake Simbi National Sanctuary (NS), Malindi Marine NP, Maralal NS, Mombasa Marine NP, Mpunguti Marine NR, Ndere Island NP, Rimoi NR, Samburu NR, Watamu Marine NP, and Watamu Marine NR.

KWS research scientists need to conduct studies to determine the extent of biodiversity loss and habitat degradation that has occurred in the ecological zones surrounding the fire hot spot regions. The results of these studies will assist KWS WPA managers in understanding the impact that wildfires have had on Kenya's wildlife. They will also serve to justify the need for increased fire management resources in the most affected regions.

5.4 Future Work

The temporal aspect of the MODIS active fire data set was not considered in this study. This work can be extended by analyzing the change of the spatial distribution of the hot spot clusters over time. A year-by-year comparison of the hot spots would be beneficial in providing a trend of the fire activity over time.

For the web application, a number of useful but missing features were identified during evaluation with the KWS GIS section. The need to incorporate the fire points occurring outside the WPAs was considered to be useful since it would enable the user to compare fire activity inside and outside a WPA. This is required when analyzing wildlife habitat factors such as migration corridors or ranges of occurrence which extend outside the WPAs.

The temporal context for the MODIS active fire data set is provided by the Acquisition Date and Acquisition Time attribute fields. The provision of these fields for each fire incident on the Google Maps interface would help a user place the incident in both time and space for complete context. These fields would also make it possible to resolve the cause of a particular fire in the event that a KWS WPA manager or research scientist might recall a particular fire incident based on when it occurred.

An additional useful enhancement to the web application would be to provide an intuitive interface through which the user can dynamically specify different parameter values for the DBSCAN algorithm. Since the ELKI DBSCAN R*-Tree index structure provides a low enough runtime complexity for the MODIS data set used in this study, the clustering can be performed as a query while retaining an acceptable system response time.

References

Achtert, E., Kriegel, H-P., Schubert, E. & Zimek, A. 2013. 'Interactive data mining with 3D-parallel-coordinate-trees'. In Ross, K. A., Srivastava, D. & Papadias, D. (eds), *Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data*, ACM, New York City, pp. 1009–1012.

Aggarwal, C. C. 2014. 'An introduction to cluster analysis'. In Aggarwal, C. C. & Reddy, C. K. (eds), *Data Clustering: Algorithms and Applications*, CRC Press, pp. 1–27.

Alelyani, S., Tang, J. & Liu, H. 2014. 'Feature selection for clustering: a review'. In Aggarwal, C. C. & Reddy, C. K. (eds), *Data Clustering: Algorithms and Applications*, CRC Press, pp. 29–60.

Arabuko Sokoke Forest Management Team. 2002. *Arabuko Sokoke Strategic Forest Management Plan: 2002-2027*. KWS, Nairobi.

Cheng, W., Wang, W. & Batista, S. 2014. 'Grid-based clustering'. In Aggarwal, C. C. & Reddy, C. K. (eds), *Data Clustering: Algorithms and Applications*, CRC Press, pp. 127–148.

Davies, D., Ilavajhala, S., Wong, M. & Justice, C. 2009. 'Fire information for resource management system: archiving and distributing MODIS active fire data'. *IEEE Transactions on Geoscience and Remote Sensing*, vol. 47, no. 1, pp. 72–79.

Davies, D., Vosloo, H., Vannan, S. & Frost, P. 2008. 'Near real-time fire alert system in South Africa: from desktop to mobile service'. *In Proceedings of the 7th ACM conference on designing interactive systems*, ACM, Cape Town, pp. 315–322.

Deng, H. & Han, J. 2014. 'Probabilistic models for clustering'. In Aggarwal, C. C. & Reddy, C. K. (eds), *Data Clustering: Algorithms and Applications*, CRC Press, pp. 61–86.

Divya, G., Rejimol, R. R. & Selvan, K. 2014. 'Suitability of clustering algorithms for crime hotspot analysis'. *International Journal of Computer Science Engineering and Technology*, vol. 4, no. 7, pp. 231–234.

ELKI Development Team. 2014a. *Package de.lmu.ifi.dbs.elki.distance.distancefunction.geo: geographic (earth) distance functions*. Department of Computer Science Database Systems Group, Ludwig-Maximilians-Universität (LMU) Munich. Available from: `http://elki.dbs.ifi.lmu.de/releases/release0.6.5~20141030/doc/de/lmu/ifi/dbs/elki/distance/distancefunction/geo/package-summary.html` [Accessed 13 August 2015].

ELKI Development Team. 2014b. *Package de.lmu.ifi.dbs.elki.math.geodesy: class WGS84SpheroidEarthModel*. Department of Computer Science Database Systems Group, Ludwig-Maximilians-Universität (LMU) Munich. Available from: `http://elki.dbs.ifi.lmu.de/releases/release0.6.5~20141030/doc/de/lmu/`

ifi/dbs/elki/math/geodesy/WGS84SpheroidEarthModel.html [Accessed 13 August 2015].

ELKI Development Team. 2014c. *Using indexes for accelerated algorithms*. Department of Computer Science Database Systems Group, Ludwig-Maximilians-Universität (LMU) Munich. Available from: http://elki.dbs.ifi.lmu.de/wiki/HowTo/Index [Accessed 14 August 2015].

ESA. 2015. *Meteosat second generation (MSG) spacecraft*. Available from: https://directory.eoportal.org/web/eoportal/satellite-missions/m/meteosat-second-generation [Accessed 21 November 2015].

Ester, M. 2014. 'Density-based clustering'. In Aggarwal, C. C. & Reddy, C. K. (eds), *Data Clustering: Algorithms and Applications*, CRC Press, pp. 111–126.

Ester, M., Kriegel, H-P., Sander, J. & Xu, X. 1996. 'A density-based algorithm for discovering clusters in large spatial databases with noise'. In Simoudis, E., Han, J. & Fayyad, U. M. (eds), *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96)*, AAAI, Portland, pp. 226–231.

EUMETSAT. 2015. *Meteosat 0 degree visualised products*. Available from: http://oiswww.eumetsat.org/IPPS/html/MSG/PRODUCTS/ [Accessed 21 November 2015].

Fournier-Viger, P. 2015. *Example 51: clustering values with the DBScan algorithm*. SPMF: A Java Open-Source Pattern Mining Library. Available from: http://www.philippe-fournier-viger.com/spmf/index.php?link=documentation.php#dbscan [Accessed 17 August 2015].

Geoscience Australia. 2015. *Sentinel hotspots*. Australian Government - Geoscience Australia. Available from: http://sentinel.ga.gov.au/ [Accessed 20 June 2015].

Giglio, L. 2013. *MODIS collection 5 active fire product user's guide, version 2.5*. NASA FIRMS. Available from: https://earthdata.nasa.gov/files/MODIS_Fire_Users_Guide_2.5.pdf [Accessed 1 July 2015].

Giglio, L. n.d.a. *Advanced very high resolution radiometer (AVHRR)*. University of Maryland. Available from: http://modis-fire.umd.edu/pages/rationale.php?target=AVHRR [Accessed 28 November 2015].

Giglio, L. n.d.b. *Geostationary operational environmental satellites (GOES)*. University of Maryland. Available from: http://modis-fire.umd.edu/pages/rationale.php?target=GOES [Accessed 28 November 2015].

Giglio, L. n.d.c. *Landsat*. University of Maryland. Available from: http://modis-fire.umd.edu/pages/rationale.php?target=LANDSAT [Accessed 28 November 2015].

Giglio, L. n.d.d. *MODIS active fire and burned area products: rationale*. University of Maryland. Available from: http://modis-fire.umd.edu/pages/rationale.php [Accessed 28 November 2015].

Giglio, L. n.d.e. *Total ozone mapping spectrometer (TOMS)*. University of Maryland. Available from: `http://modis-fire.umd.edu/pages/rationale.php?target=TOMS` [Accessed 28 November 2015].

Giglio, L., Descloitres, J., Justice, C. O. & Kaufman, Y. 2003. 'An enhanced contextual fire detection algorithm for MODIS'. *Remote Sensing of Environment*, vol. 87, pp. 273–282.

Hahsler, M., Arya, S. & Mount, D. 2015. *Package 'dbscan'*. The Comprehensive R Archive Network (CRAN). Available from: `https://cran.r-project.org/web/packages/dbscan/dbscan.pdf` [Accessed 14 August 2015].

International Organization for Standardization (ISO). 2011. *ISO/IEC 25010:2011, Systems and software engineering - systems and software quality requirements and evaluation (SQuaRE) - system and software quality models*. ISO, Geneva.

Kamau, P. N. 2013. *Anthropogenic fires, forest resources, and local livelihoods at Chyulu Hills, Kenya*. Master's Thesis. Miami University, Oxford, Ohio.

Kibriya, A. M. 2014. *Interface DistanceFunction*. Machine Learning Group at the University of Waikato. Available from: `http://weka.sourceforge.net/doc.dev/weka/core/DistanceFunction.html` [Accessed 14 August 2015].

KWS. 2012a. *Hell's Gate-Mt. Longonot ecosystem management plan: 2010-2015*. KWS, Nairobi. Available from: `http://www.kws.go.ke/content/management-plans` [Accessed 24 November 2015].

KWS. 2012b. *Meru conservation area management plan: 2007-2017*. KWS, Nairobi. Available from: `http://www.kws.go.ke/content/management-plans` [Accessed 24 November 2015].

KWS. 2012c. *Ruma national park management plan: 2012-2017*. KWS, Nairobi. Available from: `http://www.kws.go.ke/content/management-plans` [Accessed 24 November 2015].

KWS. 2012d. *Tsavo conservation area management plan: 2008-2018*. KWS, Nairobi. Available from: `http://www.kws.go.ke/content/management-plans` [Accessed 24 November 2015].

KWS. 2013a. *Aberdare ecosystem management plan: 2010-2020*. KWS, Nairobi.

KWS. 2013b. *Kiunga-Boni-Dodori (KBDCA) conservation area management plan: 2013-2023*. KWS, Nairobi. Available from: `http://www.kws.go.ke/content/management-plans` [Accessed 24 November 2015].

KWS. 2013c. *KWS strategic plan: 2012-2017*. KWS, Nairobi. Available from: `http://www.kws.go.ke/content/strategic-plans` [Accessed 24 November 2015].

KWS. 2013d. *Mt. Kenya ecosystem management plan: 2010-2020*. KWS, Nairobi. Avail-

able from: http://www.kws.go.ke/content/management-plans [Accessed 24 November 2015].

KWS & KFS. 2012. *Kakamega forest ecosystem management plan: 2012-2022*. KWS, Nairobi. Available from: http://www.kws.go.ke/content/management-plans [Accessed 24 November 2015].

Li, T. & Ding, C. 2014. 'Nonnegative matrix factorizations for clustering: a survey'. In Aggarwal, C. C. & Reddy, C. K. (eds), *Data Clustering: Algorithms and Applications*, CRC Press, pp. 149–175.

Liu, J. & Han, J. 2014. 'Spectral clustering'. In Aggarwal, C. C. & Reddy, C. K. (eds), *Data Clustering: Algorithms and Applications*, CRC Press, pp. 177–199.

NASA Earthdata. 2015. *Frequently asked questions*. Available from: https://earthdata.nasa.gov/faq [Accessed 25 June 2015].

NASA FIRMS. 2014. *FIRMS web fire mapper*. Available from: https://firms.modaps.eosdis.nasa.gov/firemap/ [Accessed 20 June 2015].

NASA FIRMS. 2015. *FIRMS MODIS fire archive download*. Available from: https://firms.modaps.eosdis.nasa.gov/download/ [Accessed 11 July 2015].

NASA FIRMS. n.d. *About FIRMS*. Available from: https://earthdata.nasa.gov/earth-observation-data/near-real-time/firms/about-firms [Accessed 25 June 2015].

NOAA. 2012. *AVHRR*. Available from: https://www.class.ngdc.noaa.gov/data_available/avhrr/index.htm [Accessed 20 November 2015].

NOAA SSD. 2014. *Wildfire automated biomass burning algorithm (WF-ABBA)*. NOAA. Available from: http://www.ssd.noaa.gov/PS/FIRE/Layers/ABBA/abba.html [Accessed 20 November 2015].

NOAA SSD. 2015. *AVHRR fire detects from the fire identification, mapping and monitoring algorithm (FIMMA)*. NOAA. Available from: http://www.ssd.noaa.gov/PS/FIRE/Layers/FIMMA/fimma.html [Accessed 20 November 2015].

Oehlschlaegel, J. 2015. *Package 'fpc'*. The Comprehensive R Archive Network (CRAN). Available from: https://cran.r-project.org/web/packages/fpc/fpc.pdf [Accessed 14 August 2015].

Oliveira, M. G. & Souza Baptista, C. 2013. 'An approach to visualization and clustering-based analysis on spatiotemporal data'. *Journal of Information and Data Management*, vol. 4, no. 2, pp. 134–145.

Palumbo, I. 2013. 'The importance of fire ecology in protected areas management'. In Paron, P., Olago, D. O., & Omuto, C. T. (eds), *Kenya: A Natural Outlook – Geo-Environmental Resources and Hazards*, vol. 16, of *Developments in Earth Surface Processes*, Elsevier, Oxford, UK, pp. 181–191. Available from: http://

www.sciencedirect.com/science/article/pii/B9780444595591000141 [Accessed 27 November 2015].

Palumbo, I., Verbeeck, B., Clerici, M. & Grégoire, J-M. 2013. 'A web client for fire monitoring in support to protected area management in Africa'. In Lasaponara, R., Masini, N. & Biscione, M. (eds), *Proceedings of the 33rd EARSeL Symposium*, EARSeL, Matera, pp. 49–58.

Reddy, C. K. & Vinzamuri, B. 2014. 'A survey of partitional and hierarchical clustering algorithms'. In Aggarwal, C. C. & Reddy, C. K. (eds), *Data Clustering: Algorithms and Applications*, CRC Press, pp. 87–110.

Russell, S. & Norvig, P. 2010. *Artificial intelligence: a modern approach*. 3rd edn. Prentice Hall, Upper Saddle River, New Jersey.

scikit-learn developers. 2014. *sklearn.cluster.DBSCAN*. scikit-learn. Available from: http://scikit-learn.org/stable/modules/generated/sklearn.cluster.DBSCAN.html [Accessed 17 August 2015].

Schubert, M., Melnikova-Albrecht, Z. & Holzmann, R. 2014. *Class DBSCAN*. Machine Learning Group at the University of Waikato. Available from: http://weka.sourceforge.net/doc.packages/optics_dbScan/weka/clusterers/DBScan.html [Accessed 14 August 2015].

Tan, P., Steinbach, M. & Kumar, V. 2005. *Introduction to data mining*. Pearson Addison-Wesley, Boston, Massachusetts.

USGS. 2015a. *Landsat 8*. Available from: http://landsat.usgs.gov/landsat8.php [Accessed 20 November 2015].

USGS. 2015b. *Landsat enhanced thematic mapper plus (ETM+)*. Available from: https://lta.cr.usgs.gov/LETMP [Accessed 20 November 2015].

Usman, M., Sitanggang, I. S. & Syaufina, L. 2015. 'Hotspot distribution analyses based on peat characteristics using density-based spatial clustering'. *Procedia Environmental Sciences*, vol. 24, pp. 132–140.

Vadrevu, K. P., Csiszar, I., Ellicott, E., Giglio, L., Badarinath, K. V. S., Vermote, E. & Justice, C. 2013. 'Hotspot analysis of vegetation fires and intensity in the Indian region'. *Selected Topics in Applied Earth Observations and Remote Sensing, IEEE Journal of*, vol. 6, no. 1, pp. 224–238.

Wang, J., Wang, X., & Liang, S. H. L. 2011. 'GeoClustering: a web service for geospatial clustering'. In Li S., Dragićević, S. & Veenendaal, B. (eds), *Advances in Web-based GIS, Mapping Services and Applications*, CRC Press, pp. 37–54.

Wilson, J. 2007. *Ozone monitoring instrument (OMI)*. NASA. Available from: http://www.nasa.gov/mission_pages/aura/spacecraft/omi.html [Accessed 20 November 2015].

Appendix A

Code and Data File Listings

Listing A.1: firms2186714310171011_MCD14ML.csv

```
latitude,longitude,brightness,scan,track,acq_date,acq_time,satellite,
confidence,version,bright_t31,frp
5.716,35.218,328.5,1.1,1,2005-12-10, 0803,T,81,5.1,306.8,17.1
3.873,34.097,320.4,1.2,1.1,2005-12-10, 0804,T,49,5.1,304.3,13.1
3.719,34.115,319.7,1.2,1.1,2005-12-10, 0804,T,66,5.1,303.3,10.8
3.715,34.118,328.7,1.2,1.1,2005-12-10, 0804,T,82,5.1,305.5,24.4
...
```

Listing A.2: MCD14ML.txt

```
longitude latitude
35.218 5.716
34.097 3.873
34.115 3.719
34.118 3.715
...
```

Listing A.3: MCD14ML_WPA.csv

```
longitude,latitude,WPA_NAME
38.365,-4.055,Tsavo West NP
38.377,-4.057,Tsavo West NP
38.376,-4.067,Tsavo West NP
38.388,-4.069,Tsavo West NP
...
```

Listing A.4: fdist.sql

```
/* create a temporary database
 * and use it for the following SQL statements */
CREATE DATABASE mydb;
USE mydb;

/* create table to hold fire points */
CREATE TABLE `mydb`.`fire` (
    `long` DOUBLE NOT NULL,
    `lat` DOUBLE NOT NULL,
    `wpa` VARCHAR(30) NOT NULL)
ENGINE = MyISAM;
```

```
12
13 /* load the fire points from the input CSV file into the table */
14 LOAD DATA INFILE '/tmp/MCD14ML_WPA.csv'
15 INTO TABLE mydb.fire
16 COLUMNS TERMINATED BY ','
17 OPTIONALLY ENCLOSED BY ''
18 ESCAPED BY ''
19 LINES TERMINATED BY '\n'
20 IGNORE 1 LINES;
21
22 /* save total number of fire points (4,968) in variable `n'
23  * for use below in calculating the relative frequencies */
24 SELECT @n := COUNT(lat)
25 FROM mydb.fire;
26
27 /* write the number of fire points in each WPA in descending order
28  * into an output CSV file */
29 SELECT wpa, COUNT(*) AS absfreq,
30     CONCAT(ROUND(((COUNT(*) / @n) * 100), 2), '%') AS relfreq
31 FROM mydb.fire
32 GROUP BY wpa
33 ORDER BY absfreq DESC
34 INTO OUTFILE '/tmp/fdist.csv'
35     FIELDS TERMINATED BY ','
36     ENCLOSED BY ''
37     ESCAPED BY ''
38     LINES TERMINATED BY '\n';
39
40 /* delete the database and its contents */
41 DROP DATABASE mydb;
```

Listing A.5: fdist.csv

```
1 Tsavo West NP,693,13.95%
2 Chyulu Hills NP,581,11.69%
3 Boni NR,443,8.92%
4 Masai Mara NR,415,8.35%
5 Dodori NR,414,8.33%
  ...
```

Listing A.6: fdist.m

```
1 % create frequency distribution bar graph
2
3 % read data from CSV file into cell array x and column vector y
```

```
[x, y] = textread('fdist.csv', '%s %d %*', 'delimiter', ',',
    'whitespace', '');

% reverse both x and y to have higher values at top of bar graph
x = flip(x);
y = flip(y);

% plot horizontal bar graph
barh(y, 0.7, 'facecolor', 'black', 'edgecolor', 'black');
set(gca, 'yticklabel', x, 'ticklength', [0.007, 0.007], 'xgrid', 'on',
    'xminortick', 'on', 'gridlinestyle', ':', 'fontsize', 8);

% fix the current y axis limits manually
axis([-Inf, Inf, 0, length(y)+1], 'manual');

% set axis labels and graph title
xlabel('No. of Fire Points', 'fontsize', 9);
ylabel('WPA', 'fontsize', 9);
title("Frequency Distribution of MODIS Fire Points in Kenya's WPAs",
    'fontsize', 9);

% save the bar graph to a png file
print -dpng fdist.png
```

Listing A.7: scatterplot.m

```
% create scatterplot of preprocessed MODIS fire points
% renamed from scatter.m, which is the name of an Octave function file

% read data from CSV file into column vectors x and y
[x, y] = textread('MCD14ML_WPA.csv', '%f %f %*', 'delimiter', ',',
    'headerlines', 1, 'whitespace', '');

% plot scatterplot
scatter(x, y, 5, 'black', '+');

% fix current y axis limits manually; make axes equal
axis([33, 43, -5, 5], 'manual', 'equal');

% set axis labels and graph title
xlabel('Longitude');
ylabel('Latitude');
title("Scatterplot Diagram of MODIS Fire Points in Kenya's WPAs");

% save graph to png file
```

```
print -dpng scatterplot.png
```

Listing A.8: knn-distances.txt

```
111.28164593701999
156.8733266768559
156.8733266768559
156.87332667687468
156.87332667687468
...
210464.43266414793
383443.4757380694
383948.18111164623
384144.3296365825
387472.3639902396
```

Listing A.9: noise.txt

```
# Cluster: Noise
# Cluster name: Noise
# Cluster noise flag: true
# Cluster size: 4203
ID=4968 41.164 -1.505 Boni NR
ID=4967 41.162 -1.516 Boni NR
ID=4966 35.702 1.734 South Turkana NR
ID=4965 35.735 1.831 South Turkana NR
ID=4964 41.488 -1.74 Kiunga Marine NR
...
ID=5 38.386 -4.079 Tsavo West NP
ID=4 38.388 -4.069 Tsavo West NP
ID=3 38.376 -4.067 Tsavo West NP
ID=2 38.377 -4.057 Tsavo West NP
ID=1 38.365 -4.055 Tsavo West NP
```

Listing A.10: cluster_0.txt

```
# Cluster: Cluster 0
# Cluster name: Cluster
# Cluster noise flag: false
# Cluster size: 8
ID=2067 37.84 -2.568 Chyulu Hills NP
ID=3290 37.841 -2.568 Chyulu Hills NP
ID=3321 37.837 -2.566 Chyulu Hills NP
ID=2573 37.842 -2.572 Chyulu Hills NP
```

```
ID=1076 37.835 -2.569 Chyulu Hills NP
ID=2048 37.835 -2.57 Chyulu Hills NP
ID=2581 37.841 -2.562 Chyulu Hills NP
ID=2124 37.834 -2.57 Chyulu Hills NP
```

Listing A.11: Query for the frequency distribution of fire hot spot clusters in the WPAs

```
SELECT
    name AS wpa,
    COUNT(DISTINCT clusterid) AS absfreq,
    ROUND(
        (
            CAST(
                COUNT(DISTINCT clusterid) AS FLOAT
            ) / 43 * 100
        ), 2
    ) AS relfreq
FROM clu_wpa INNER JOIN clu_fire
ON clu_wpa.wpaid = clu_fire.wpaid
WHERE clusterid != 'Noise'
GROUP BY name
ORDER BY absfreq DESC
```

Listing A.12: Query for the number of fire points in each fire hot spot cluster

```
SELECT
    DISTINCT clusterid,
    name AS wpa,
    COUNT(clusterid) AS size
FROM clu_wpa INNER JOIN clu_fire
ON clu_wpa.wpaid = clu_fire.wpaid
WHERE clusterid != 'Noise'
GROUP BY clusterid, name
ORDER BY size DESC
```

Listing A.13: Query for the average number of fire points per km^2 in each WPA

```
SELECT
    name AS wpa,
    COUNT(clusterid) / size AS density
FROM clu_wpa INNER JOIN clu_fire
ON clu_wpa.wpaid = clu_fire.wpaid
GROUP BY name, size
ORDER BY density DESC
```

Listing A.14: wpa.sql

```
USE [kwsids]
GO

SET ANSI_NULLS ON
GO

SET QUOTED_IDENTIFIER ON
GO

SET ANSI_PADDING ON
GO

CREATE TABLE [dbo].[clu_wpa] (
    [wpaid] [tinyint] IDENTITY(1,1) NOT NULL,
    [name] [varchar](30) NOT NULL,
    [size] [decimal](9, 4) NOT NULL,
    CONSTRAINT [PK_clu_wpa] PRIMARY KEY CLUSTERED ([wpaid] ASC)
    WITH (PAD_INDEX = OFF, STATISTICS_NORECOMPUTE = OFF,
        IGNORE_DUP_KEY = OFF, ALLOW_ROW_LOCKS = ON,
        ALLOW_PAGE_LOCKS = ON
    ) ON [PRIMARY]
) ON [PRIMARY]
GO

SET ANSI_PADDING OFF
GO
```

Listing A.15: fire.sql

```
USE [kwsids]
GO

SET ANSI_NULLS ON
GO

SET QUOTED_IDENTIFIER ON
GO

SET ANSI_PADDING ON
GO

CREATE TABLE [dbo].[clu_fire] (
    [fireid] [smallint] IDENTITY(1,1) NOT NULL,
```

```
    [lat] [decimal](9, 4) NOT NULL,
    [long] [decimal](9, 4) NOT NULL,
    [wpaid] [tinyint] NOT NULL,
    [clusterid] [varchar](10) NOT NULL,
    CONSTRAINT [PK_clu_fire] PRIMARY KEY CLUSTERED ([fireid] ASC)
    WITH (PAD_INDEX = OFF, STATISTICS_NORECOMPUTE = OFF,
        IGNORE_DUP_KEY = OFF, ALLOW_ROW_LOCKS = ON,
        ALLOW_PAGE_LOCKS = ON
    ) ON [PRIMARY]
) ON [PRIMARY]
GO

SET ANSI_PADDING OFF
GO

/* foreign key defining relationship between wpa and fire tables */
ALTER TABLE [dbo].[clu_fire]
WITH CHECK
ADD CONSTRAINT [FK_clu_wpa_clu_fire]
    FOREIGN KEY([wpaid])
    REFERENCES [dbo].[clu_wpa] ([wpaid])
    ON UPDATE CASCADE
GO

ALTER TABLE [dbo].[clu_fire]
CHECK CONSTRAINT [FK_clu_wpa_clu_fire]
GO
```

Listing A.16: fire.awk

```
# This program produces the data for the Fire table. The CSV output
# file has 4 columns for lat, long, wpa_name, and cluster_id.
{
    if (FNR == 1)
        cid = $NF;
    else if (FNR > 4) {
        printf("%s,%s,", $3, $2);
        for (i = 4; i <= NF; i++) {
            c = (i == 4) ? "" : " ";
            printf("%s%s", c, $i);
        }
        printf(",%s", cid);
        printf("\n");
    }
}
```

Listing A.17: wpa.awk

```
# This program further processes the data for the Fire table.

FILENAME == "wpa.csv" {
    a[$1] = FNR
}

# replace wpa names with respective ids
FILENAME == "fire.csv" {
    print $1, $2, a[$3], $4
}
```

Listing A.18: wpa.csv

```
Aberdare NP,765.7
Amboseli NP,392
Arabuko Sokoke NP,6
Arawale NR,533
Bisanadi NR,606
...
```

Listing A.19: fire.csv

```
-2.568,37.84,10,0
-2.568,37.841,10,0
-2.566,37.837,10,0
-2.572,37.842,10,0
-2.569,37.835,10,0
...
```